Relevance Supervision Can Teach Neural Retrievers to Prefer LLM-Generated Text

Anonymous ACL submission

Abstract

Recent studies show that neural retrievers often exhibit source bias, ranking LLM-generated passages above semantically similar human-written ones, which threatens the reliability of modern information access systems. While this bias is often viewed as a model-level property, we show that relevance supervision itself can induce it. Empirically, source bias is clearest in relevance-supervised retrievers, and the same encoders become more LLM-preferring after relevance fine-tuning. We trace this effect to surface-level artifacts in supervised retrieval data: annotated positives and negatives differ systematically in properties such as fluency and lexical specificity, and these differences parallel those between LLM-generated and human-written passages. As a result, retrievers can learn artifact-based shortcuts for relevance. We further show theoretically that, under contrastive learning, artifacts that reliably separate positives from negatives can be encoded into learned relevance scores. Motivated by this mechanism, we propose two mitigation strategies: balancing artifact distributions during negative sampling and removing the artifact-aligned embedding component at inference time. Both reduce source bias with little loss in retrieval effectiveness. Overall, our findings support a data-centric view of source bias: retrievers may prefer LLM-generated text not merely because of inherent model properties, but because relevance supervision teaches them to rely on surface-level artifacts.

1 Introduction

The rapid rise of large language models (LLMs) has reshaped the information landscape, creating corpora in which human-written and LLM-generated texts increasingly coexist. In this setting, recent work shows that neural retrievers can rank LLM-generated passages above semantically similar human-written ones, a phenomenon known as *source bias* (Dai et al., 2024b,c). This preference matters because retrievers serve as front ends for search, recommendation, and retrieval-augmented generation: if they systematically favor generated text, they may surface fluent but less reliable passages, reduce the visibility of human-authored content, and amplify generated text through feedback loops (Chen et al., 2024; Zhou et al., 2025).

Prior work has mainly explained source bias through model-side or text-level factors, such as architectural similarities between PLM-based retrievers and LLMs (Dai et al., 2024b,c) or retrievers' preference for low-perplexity passages (Wang et al., 2025). Accordingly, existing mitigation methods intervene either on the retriever side (Wang et al., 2025) or on the generation side (Dai et al., 2025). Yet these explanations leave open a data-centric possibility: relevance supervision itself may induce source bias. In supervised retrieval data, relevant and non-relevant passages may differ beyond semantic relevance. They may also differ in surface-level *artifacts*, such as fluency or lexical specificity, that retrievers can exploit as shortcuts.

In this paper, we examine whether such artifact-based shortcuts make supervised retrievers favor LLM-generated text. Specifically, we ask whether artifacts that separate positives from negatives during training can transfer to source preferences at test time. This leads to three research questions:

RQ1: Where does source bias emerge? We first examine source preference across three retriever families: relevance-supervised retrievers, general-purpose embedding models, and unsupervised retrievers. We find that preference for LLM-generated passages is strongest and most consistent in relevance-supervised retrievers, and that fine-tuning unsupervised retrievers on MS MARCO shifts them toward LLM-generated passages.

RQ2: How can relevance supervision shape source bias? Across retrieval datasets, positive and negative documents differ systematically in surface-level properties such as fluency and lexi-

cal specificity, and these patterns align with those between LLM-generated and human-written text. We provide a theoretical analysis showing how contrastive objectives can encode such artifact signals when they help separate positives from negatives.

RQ3: How can source bias be mitigated? Motivated by this mechanism, we test two strategies: balancing artifact distributions during negative sampling at training time and suppressing the artifact-aligned component of retriever embeddings at inference time. Both reduce source bias while largely preserving retrieval effectiveness.

Overall, our findings support a data-centric view of source bias: rather than being an inherent property of neural retrievers, source bias can emerge when relevance supervision encourages models to rely on artifact-based shortcuts that also distinguish LLM-generated text from human-written text. This perspective suggests complementary mitigation directions at both training and inference time, including artifact-aware data construction and embedding-level debiasing.

2 Related Work

Source Bias in Information Retrieval. Recent work has shown that neural retrievers can prefer LLM-generated passages over semantically similar human-written ones, a phenomenon known as *source bias* (Dai et al., 2024c,b). Cocktail (Dai et al., 2024a) extends BEIR (Thakur et al., 2021) with LLM-generated counterparts, enabling systematic evaluation across retrieval datasets. Related source-preference effects have also been observed in multimodal retrieval, recommendation, and retrieval-augmented generation (Xu et al., 2024; Zhou et al., 2025; Chen et al., 2024).

Mechanisms and Mitigation. Prior work links source bias to architectural similarity between PLM retrievers and LLMs (Dai et al., 2024c), preference for low-perplexity documents (Wang et al., 2025), or distribution mismatch (Dai et al., 2024b). Mitigation methods thus target either retrievers or LLM outputs (Wang et al., 2025; Dai et al., 2025). In contrast, we study a data-centric mechanism: relevance supervision may make surface artifacts predictive of relevance, leading retrievers to learn shortcuts that favor LLM-generated text.

3 RQ1: Where Does Source Bias Emerge?

Prior work reports that neural retrievers often rank LLM-generated passages above semantically similar human-written ones (Dai et al., 2024b,c). However, prior studies mainly evaluate retrievers trained with large-scale relevance supervision, such as MS MARCO (Nguyen et al., 2016). This raises the question of whether source bias appears broadly across neural retrievers, or is especially associated with relevance-supervised retrieval.

We address this question with two analyses. First, we compare three retrieval model families: relevance-supervised retrievers, general-purpose embedding models, and unsupervised retrievers. Second, we fine-tune several unsupervised retrievers on MS MARCO to test whether relevance supervision makes the same base encoders favor LLM-generated passages.

3.1 Experimental Setup

We evaluate 13 neural retrievers from three families. The relevance-supervised group includes ANCE (Xiong et al., 2020), TAS-B (Hofstätter et al., 2021), coCondenser (Gao and Callan, 2022), RetroMAE (Xiao et al., 2022), and DRAGON (Lin et al., 2023). The general-purpose embedding group includes BGE (Xiao et al., 2024), BCE (NetEase Youdao, 2023), GTE (Li et al., 2023), E5 (Wang et al., 2024), and M3E (Wang et al., 2023). The unsupervised group includes Contriever (Izacard et al., 2022), E5-unsup (Wang et al., 2024), and unsupervised SimCSE (Gao et al., 2021). We treat these groups as common training regimes rather than a strict supervised/unsupervised partition, since some general-purpose embedding models may include retrieval-style or weakly supervised data in their training mixtures.

We evaluate on 14 Cocktail datasets (Dai et al., 2024a), which augment BEIR collections (Thakur et al., 2021) with LLM-generated counterparts designed to preserve the semantics of human-written passages. The original study reports high semantic consistency on a manually checked subset. Dataset statistics, model checkpoints, and resource links are provided in Appendix C and Appendix B.

Following prior work (Huang et al., 2025), we measure source preference by the top-ranked source composition, using ΔNDSR@5:

$$\mathrm{NDSR}_c@k = \frac{\sum_{i=1}^{k} \mathbb{1}(\mathrm{source}(d_i) = c)\ w_i}{\sum_{i=1}^{k} w_i},$$

$$\Delta\mathrm{NDSR}@k = \mathrm{NDSR}_{\mathrm{Human}}@k - \mathrm{NDSR}_{\mathrm{LLM}}@k,$$

where $w_i = 1/\log_2(1+i)$. $\Delta\mathrm{NDSR}@k$ ranges from -1 to 1: positive values indicate preference for human-written passages, while negative values

Table 1: ΔNDSR@5 across 14 datasets and 13 retrievers spanning three model families. Negative values indicate LLM preference and positive values indicate human preference. $*$ denotes $p < 0.05$.

Dataset (↓)	Relevance-Supervised Retrievers					General-Purpose Embedding Models					Unsupervised Retrievers		
	ANCE	TAS-B	coCondenser	RetroMAE	DRAGON	BGE	BCE	GTE	E5	M3E	Contriever	E5-Unsup	SimCSE
MS MARCO	-0.040*	-0.119*	-0.018*	-0.080*	-0.081*	-0.021*	0.084*	-0.074*	-0.036*	0.053*	0.280*	0.094*	0.384*
DL19	-0.073	-0.224*	-0.072	-0.180*	-0.233*	-0.017	0.119	-0.178*	0.015	0.139	0.271*	0.086	0.428*
DL20	-0.029	-0.070	-0.078	-0.081	-0.116*	0.057	0.048	-0.049	0.012	0.203*	0.275*	0.190*	0.389*
NQ	-0.040*	-0.074*	-0.067*	-0.055*	-0.096*	-0.078*	0.324*	-0.003	0.153*	0.040*	0.186*	0.228*	0.140*
NFCorpus	-0.087*	-0.082*	-0.067*	-0.098*	-0.079*	0.030	-0.064*	-0.142*	0.034	-0.143*	-0.083*	-0.348*	0.127*
TREC-COVID	-0.162*	-0.328*	-0.340*	-0.193*	-0.133*	0.014	-0.025	-0.236*	-0.118	-0.085	-0.135*	-0.224*	0.162*
HotpotQA	-0.015*	-0.011*	-0.008*	-0.013*	0.014*	0.061*	0.184*	0.010*	0.078*	0.063*	-0.273*	-0.091*	0.097*
FiQA-2018	-0.179*	-0.169*	-0.257*	-0.244*	-0.160*	-0.150*	0.414*	-0.050*	-0.116*	0.102*	-0.068*	-0.052*	0.210*
Touché-2020	-0.101	-0.165*	-0.128*	-0.099	-0.052*	-0.042	0.218*	-0.017	-0.185*	0.242*	-0.133*	-0.073	0.027
DBpedia	-0.095*	-0.039*	-0.053*	-0.077*	-0.054*	0.017	0.069*	-0.035*	0.003	0.019	-0.130*	-0.062*	0.064*
SCIDOCS	-0.040*	-0.054*	-0.058*	-0.073*	-0.048*	-0.061*	0.517*	-0.046*	0.010	0.275*	0.028*	0.059*	0.268*
FEVER	-0.199*	-0.024*	-0.032*	-0.006*	-0.040*	0.040*	0.306*	-0.027*	0.031*	0.031*	0.028*	-0.008*	0.031*
Climate-FEVER	-0.314*	-0.082*	-0.153*	-0.105*	-0.091*	-0.038*	0.642*	-0.080*	0.215*	0.123*	-0.003	0.017	0.070*
SciFact	-0.024	-0.058*	-0.049*	-0.048*	-0.041*	0.011	0.015	-0.079*	0.004	-0.206*	0.017	-0.101*	-0.059*

Table 2: ΔNDSR@5 for the unsupervised retrievers in Table 1 after MS MARCO fine-tuning. The "-FT" suffix denotes fine-tuning on MS MARCO. Negative values indicate LLM preference and positive values indicate human preference. $*$ denotes $p < 0.05$.

Dataset (↓)	Relevance-Supervised Retrievers		
	Contriever-FT	E5-FT	SimCSE-FT
MS MARCO	0.012*	-0.044*	-0.053*
DL19	-0.035	-0.198*	-0.133
DL20	0.121*	0.022	-0.178*
NQ	-0.038*	-0.051*	-0.060*
NFCorpus	-0.139*	-0.189*	-0.060*
TREC-COVID	-0.282*	-0.271*	-0.205*
HotpotQA	-0.004	-0.019*	-0.013*
FiQA-2018	-0.215*	-0.212*	-0.189*
Touché-2020	-0.087*	-0.196*	-0.169*
DBpedia	-0.010	-0.036*	-0.053*
SCIDOCS	-0.050*	-0.072*	-0.041*
FEVER	-0.018*	-0.064*	0.000
Climate-FEVER	-0.099*	-0.091*	-0.049*
SciFact	-0.086*	-0.077*	-0.044*

indicate preference for LLM-generated passages.

3.2 Results

Source preference across retriever families. Table 1 reports ΔNDSR@5 across 14 datasets and 13 retrievers. Relevance-supervised retrievers consistently favor LLM-generated passages, with negative values for nearly all datasets and models. By contrast, general-purpose embedding models and unsupervised retrievers show mixed preferences that vary across datasets. These results suggest that source bias is strongest and most consistent among retrievers trained with relevance supervision. Retrieval effectiveness is reported in Appendix D.

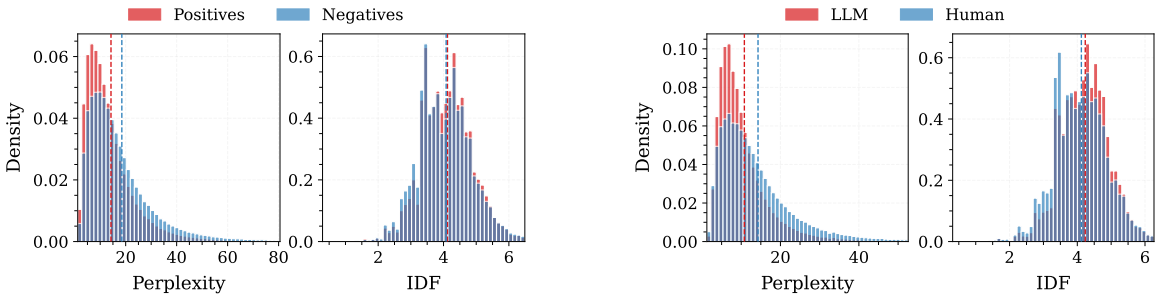
Effect of relevance fine-tuning. To more directly examine the role of supervised retrieval training, we fine-tune Contriever, E5-unsup, and SimCSE on MS MARCO. This comparison keeps the base encoder family fixed while introducing relevance-labeled query–passage training under a standard retrieval fine-tuning setup. As shown in Table 2, fine-tuning shifts these models toward LLM-generated passages, with negative ΔNDSR@5 values on most datasets. These results indicate that relevance fine-tuning can shift models toward LLM-oriented source preference, suggesting that the effect is not simply a consequence of encoder architecture.

Robustness to passage length. We assess whether the source preference is robust to passage length. LLM-generated passages in Cocktail are often shorter than their human-written counterparts, and neural retrievers can be sensitive to length-related artifacts (Thakur et al., 2024; Fayyaz et al., 2025). We therefore construct length-controlled variants by lengthening LLM-generated passages while preserving semantics. As detailed in Appendix J, relevance-supervised retrievers continue to prefer LLM-generated passages even when longer, although the preference magnitude decreases. This suggests that length affects the magnitude of source preference, but not its direction.

Summary. The results show that LLM preference is strongest and most consistent under relevance-supervised retrieval, and that relevance fine-tuning shifts initially unsupervised retrievers toward LLM-generated passages. Although source bias may arise from multiple factors, these findings point to supervised retrieval training as an important contributor. This motivates our next question: why can relevance supervision lead retrievers to prefer LLM-generated text?

4 RQ2: How Can Relevance Supervision Shape Source Bias?

Having observed that source bias is most systematic in relevance-supervised retrieval pipelines, we



(a) Positives vs. Negatives.

(b) LLM-generated vs. Human-written.

Figure 1: PPL and IDF distributions for (a) annotated positives versus training negatives and (b) LLM-generated versus human-written passages. Dashed lines indicate means. In both comparisons, positives and LLM-generated passages exhibit lower PPL and higher IDF, indicating parallel artifact-related shifts.

now ask how relevance supervision may encourage this behavior. We hypothesize that positive and negative passages in supervised retrieval data can differ in *surface-level artifacts* that correlate with relevance labels but are not identical to query-specific semantic relevance. These artifacts include textual properties such as fluency and lexical specificity. When such artifacts also overlap with properties of LLM-generated text, retrievers may learn them as shortcuts and assign higher scores to LLM-generated passages at inference time.

We examine this mechanism in three steps: first using illustrative linguistic diagnostics, then drawing on embedding-space evidence, and finally adopting a theoretical perspective.

4.1 Illustrative Linguistic Diagnostics

We first examine two simple linguistic diagnostics that may reflect surface-level differences: perplexity (PPL) and inverse document frequency (IDF). PPL measures how predictable a passage is under a language model and can reflect fluency or textual regularity, while IDF reflects lexical specificity. Together, they provide an interpretable, low-dimensional view of possible artifact differences associated with relevance labels and source type.

For a passage $d = (w_1, \ldots, w_{|d|})$, we compute

$$\mathrm{PPL}(d) = \exp\left(-\frac{1}{|d|}\sum_{i=1}^{|d|} \log p_\theta(w_i \mid w_{<i})\right),$$

using Llama-3-8B-Instruct (Dubey et al., 2024). For IDF, we use $\mathrm{IDF}(t) = \log \frac{N}{1+\mathrm{df}(t)}$ and take the median token IDF as the passage-level score, with statistics estimated on the full MS MARCO collection using the Apache Lucene tokenizer (Hatcher and Gospodnetic, 2004).

Positives vs. Negatives. We treat MS MARCO passages annotated as relevant to at least one training query as positives. We compare these positives with the remaining collection, which serves as the candidate pool for sampled or mined training negatives. This choice reflects the fact that the exact negatives used by released retrievers are often unavailable and may vary across training pipelines. We refer to this pool as the "negative pool" for brevity, while noting that it may contain unannotated positives and is not a clean set of judged non-relevant passages. Thus, the comparison should be interpreted as a corpus-level diagnostic of surface-level differences between positives and likely negative candidates, rather than as a reconstruction of any particular retriever's training negatives.

As shown in Figure 1a, positives exhibit lower PPL and slightly higher IDF than the negative pool. The PPL difference is more pronounced, whereas the IDF difference is smaller but directionally consistent. Together, these patterns indicate a distributional tendency for positives to be more predictable under the language model and marginally more lexically specific. Appendix F provides detailed statistics and additional evidence that similar PPL imbalances appear across other IR datasets, suggesting that this pattern is not unique to MS MARCO.

LLM-generated vs. Human-written Passages. We next compare LLM-generated passages with human-written passages across the 14 BEIR-derived datasets in Cocktail. Figure 1b shows representative results on MS MARCO, where LLM-generated passages exhibit lower PPL and higher IDF than human-written passages. Full results are provided in Appendix F. Thus, under these diagnostic measures, the positive–negative shift and the LLM–human shift are aligned in direction.

Summary. Taken together, these diagnostics reveal parallel surface-level patterns associated with relevance supervision and source type. Since PPL and IDF capture only selected aspects of textual variation, we next examine whether retriever embeddings encode a corresponding direction aligning the positive–negative supervision contrast with the LLM–human source contrast.

4.2 Embedding-Space Shifts

In this section, we examine whether the embedding shift associated with retrieval supervision (positives vs. negatives) aligns with the shift associated with source type (LLM-generated vs. human-written passages). Such alignment would suggest that source-related differences overlap with representational directions learned from relevance supervision. We proceed in three steps: (1) estimate a positive–negative supervision direction; (2) estimate the LLM–human source direction and assess its consistency; and (3) evaluate whether the two directions are aligned.

Notation. Let q denote a query and d a passage. For supervised retrieval, d^+ and d^- denote an annotated positive passage and a sampled negative passage, respectively. For source-type analysis, d^{LLM} and d^{Human} denote an LLM-generated passage and its human-written counterpart. The query and document encoders $h_q(\cdot)$ and $h_d(\cdot)$ map inputs to embeddings in $\mathbb{R}^m$, and the retrieval score is

$$s_\theta(q, d) = \langle h_q(q), h_d(d) \rangle.$$

We use δ to denote an embedding displacement vector. For example, the LLM–human displacement is

$$\delta^{\mathrm{LH}} = h_d(d^{\mathrm{LLM}}) - h_d(d^{\mathrm{Human}}).$$

The symbol $\bar{\delta}$ denotes the average displacement over a set of paired passages, and $\mathbb{E}[\cdot]$ denotes expectation over the indicated distribution.

Estimating the Positive–Negative Embedding Direction. To estimate the embedding direction associated with retrieval supervision, we construct positive–negative pairs from MS MARCO. Rather than comparing annotated positives with arbitrary unannotated passages, we follow common practice (Karpukhin et al., 2020): for each query, we retrieve the top-10 BM25 candidates and randomly sample one as the negative paired with the annotated positive. This produces a retrieval-like contrast in which positives and negatives have substantial lexical and topical overlap. We then estimate the average positive–negative displacement as

$$\bar{\delta}_{\mathrm{PN}} = \mathbb{E}\big[h_d(d^+) - h_d(d^-)\big].$$

We refer to this vector as the supervision-induced positive–negative direction.

Reference Baselines in High-Dimensional Space. To interpret the magnitude of directional consistency, we compare cosine similarities with those of random vectors in the same dimensionality. In 768 dimensions, random vectors are nearly orthogonal, so their cosine similarities concentrate around zero. Over 99.7% of random pairs fall within $\pm 3\sigma$ of the mean (Appendix G). We therefore use this range as a descriptive reference: cosine similarities outside it indicate unusually strong alignment compared with random directions.

Does the LLM–Human Contrast Exhibit Directional Consistency? The LLM–human comparison uses semantically aligned counterparts, allowing us to compute pairwise displacements directly:

$$\delta_i^{\mathrm{LH}} = h_d(d_i^{\mathrm{LLM}}) - h_d(d_i^{\mathrm{Human}}).$$

We evaluate whether these displacements form a consistent direction along three dimensions.
(1) Within datasets. We compute the average pairwise cosine similarity

$$\mathbb{E}_{i \neq j}\left[\cos(\delta_i^{\mathrm{LH}}, \delta_j^{\mathrm{LH}})\right].$$

Values exceeding the random-vector reference range indicate stronger-than-random directional consistency within a dataset, as shown in Figure 2a.
(2) Across datasets. For each dataset D, we compute the dataset-level mean displacement

$$\bar{\delta}_{\mathrm{LH},D} = \mathbb{E}_{d_i \in D}[\delta_i^{\mathrm{LH}}],$$

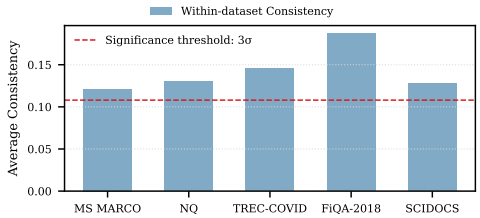
and evaluate cross-dataset alignment by

$$\cos(\bar{\delta}_{\mathrm{LH},D_1}, \bar{\delta}_{\mathrm{LH},D_2}).$$

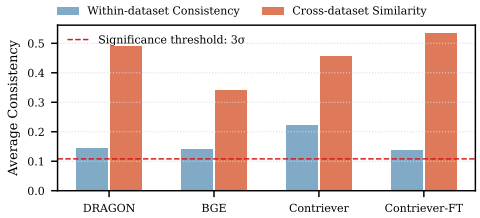
As shown in Figure 2b, the mean source displacements are consistently aligned across datasets.
(3) Across retrievers. We repeat the analysis with multiple retrievers. Figure 2c shows that the LLM–human displacement exhibits similar directional patterns across the retrievers examined.

Together, these results show that the LLM–human contrast corresponds to a reproducible embedding displacement across datasets and retrievers. The sign-flip control further shows that this pattern reflects a shared direction rather than only non-directional properties such as vector magnitude.

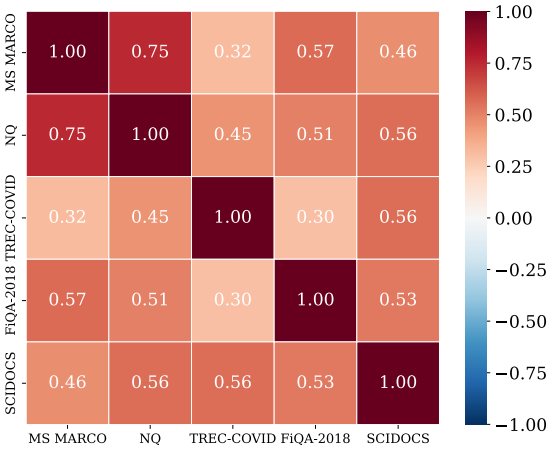
Do the Positive–Negative and LLM–Human Directions Align? We next examine whether the source-related displacement aligns with the



(a) Within-dataset consistency.

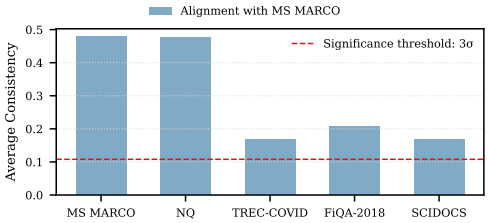


(c) Cross-retriever consistency.

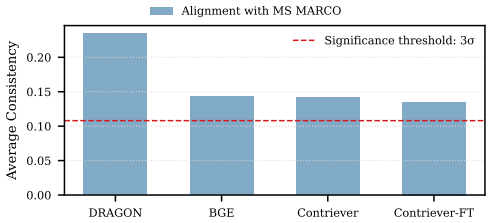


(b) Cross-dataset consistency.

Figure 2: The LLM–Human distinction forms a stable embedding-space direction. The plots demonstrate this consistency along three dimensions: (a) within datasets, (b) across datasets, and (c) across retrievers. Panels (a) and (b) use the DRAGON retriever. Results for all 14 datasets are in Appendix H.



(a) Cross-dataset consistency.



(b) Cross-retriever consistency.

Figure 3: The LLM–Human displacement aligns with the positive–negative supervision direction. Panels (a) and (b) show cross-dataset and cross-retriever consistency, respectively. Across both settings, cosine similarities exceed the 3σ threshold, confirming a stable and coherent embedding-space direction.

positive–negative supervision contrast. For each dataset D, we compute the cosine similarity between the mean LLM–human displacement, $\overline{\delta}_{\mathrm{LH},D}$, and the positive–negative direction estimated from MS MARCO, $\overline{\delta}_{\mathrm{PN}}$. As shown in Figure 3a, the two directions are consistently aligned across datasets and exceed the random-vector baseline. Figure 3b shows a similar pattern across retrievers, indicating that the alignment is not confined to a single model.

Empirical Validation Check. Because dense retriever embeddings can be anisotropic, the random-vector baseline serves mainly as an idealized geometric reference. As an empirical complement, we also report a paired sign-flip check in Appendix H. This check randomly flips the sign of each LLM–human displacement and recomputes the within-dataset consistency, cross-dataset consistency, and PN–LH alignment metrics, serving as a sign-randomized empirical baseline under the observed embedding geometry. The results remain consistent with our main findings.

This alignment indicates that the LLM–human source contrast and the positive–negative supervision contrast share a representational component. Rather than treating this component as purely stylistic, we interpret it as a source-related signal that can overlap with relevance supervision in the learned embedding space. We next provide a theoretical perspective on how such signals can enter the retrieval score under contrastive training.

4.3 Theoretical Perspective

We present a compact theoretical account of how contrastive retrievers learn source-related components that align with relevance supervision. Formal statements and proofs are provided in Appendix E.

Let a document d be described by semantic features M_d and artifact features A_d, such as fluency and lexical specificity. These artifacts need not be reliable indicators of query-specific relevance, but they can still influence training if they help predict the positive–negative contrast beyond semantic

relevance alone. In the density-ratio view of contrastive learning, the learned score is related to the log ratio between positive-pair and negative-pair densities (Appendix E). Thus, any feature predictive of this contrast, including artifact features, can contribute to the score, leading to:

$$s^*(q, d) \approx \mathrm{Score}_{\mathrm{sem}}(q, M_d) + \mathrm{Score}_{\mathrm{art}}(q, A_d).$$

The additive form can be viewed as a first-order Taylor approximation of the score around the observed data: the artifact term captures the leading local contribution of artifact features, while higher-order interactions are left in the residual. Thus, when artifact features help distinguish positives from negatives in the training distribution, they can become **learnable shortcuts.**

For dot-product retrievers, this shortcut can appear as an embedding-space component. Locally linearizing the document encoder with respect to artifact features gives

$$\begin{aligned} s_\theta(q, d) &= \langle h_q(q), h_d(d) \rangle \\ &\approx \langle h_q(q), h_d^{\mathrm{sem}}(d) \rangle + \langle h_q(q), h_d^{\mathrm{art}}(d) \rangle. \end{aligned}$$

This view is consistent with our empirical finding that the LLM–human displacement is directionally stable and aligned with the positive–negative supervision direction. It also provides a basis for mitigation strategies that reduce artifact imbalance in training or suppress artifact-aligned embedding components at inference time.

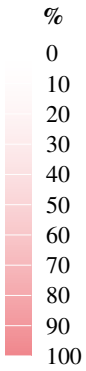
5 RQ3: How Can Source Bias Be Mitigated?

The analyses above suggest that source bias can arise when relevance supervision makes surface-level artifacts predictive of the positive–negative contrast. This perspective points to two intervention points. First, at training time, we can reduce the artifact mismatch between positives and negatives so that such cues become less useful as shortcuts. Second, at inference time, we can suppress embedding components aligned with the LLM–human displacement identified in Section 4.2.

Importantly, these interventions not only mitigate source bias, but also test the proposed mechanism: if reducing artifact mismatch or suppressing artifact-aligned directions weakens source preference, this provides empirical support for our explanation. Our goal is therefore not to advance state-of-the-art debiasing, but to use simple interventions to probe the mechanism underlying source bias. We examine both strategies below.

Table 3: ΔNDSR@5 under three negative sampling strategies with increasing expected artifact mismatch (in-batch only → standard → hard-neg only). Darker shading in the Average row denotes stronger bias.

	In-batch only	Standard	Hard-neg only
MS MARCO	0.014	-0.051	-0.057
DL19	0.025	-0.155	-0.182
DL20	0.041	-0.120	-0.152
NQ	0.020	-0.081	-0.085
NFCorpus	-0.050	-0.068	-0.093
TREC-COVID	-0.182	-0.252	-0.285
HotpotQA	0.003	0.017	-0.021
FiQA-2018	-0.055	-0.227	-0.238
Touché-2020	-0.077	-0.202	-0.193
DBpedia	-0.021	-0.041	-0.043
SCIDOCS	0.010	-0.051	-0.035
FEVER	0.014	-0.005	-0.013
Climate-FEVER	-0.032	-0.071	-0.080
SciFact	-0.032	-0.051	-0.053
Average	-0.024	-0.099	-0.109



Training-time Intervention: Reducing Artifact Mismatch. We first examine whether source bias can be reduced by changing the negative sampling strategy during retriever training. As a simple intervention, we use *in-batch only* sampling, where negatives are annotated positives associated with other queries in the same batch. Since both positives and negatives are drawn from the annotated-positive pool, this setting serves as a low-mismatch condition that reduces systematic artifact differences between the two sides of the training contrast.

We compare it with two reference settings: a *standard* setup that combines in-batch negatives with one mined hard negative per query, and a *hard-neg only* setup where negatives are drawn from the broader unannotated corpus. These settings form a controlled spectrum from lower to higher expected artifact mismatch. For a controlled comparison, we fine-tune BERT-based retrievers on MS MARCO using the BEIR pipeline (Devlin et al., 2019; Thakur et al., 2021), changing only the negative sampling strategy.

Table 3 shows a clear trend: in-batch-only sampling yields the weakest LLM preference, standard sampling leads to stronger preference, and hard-neg-only sampling is strongest on average. This pattern is consistent with our hypothesis that artifact mismatch in the training contrast can amplify source bias. Since removing mined hard negatives slightly reduces retrieval effectiveness (Appendix I), we view in-batch-only sampling primarily as a mechanism probe rather than an optimal training recipe. More practical variants can retain hard negatives while controlling their artifact distribution, such as sampling hard negatives from the annotated-positive pool (Appendix K).

Table 4: ΔNDSR@5 before and after inference-time debiasing across 5 datasets and 5 relevance-supervised retrievers, using leave-one-dataset-out calibration to estimate the debiasing direction. Positive values indicate human preference, while negative values indicate LLM preference. In the Average row, the first line reports mean ΔNDSR@5, and the second line reports the remaining |ΔNDSR@5| relative to the original value. Darker shading indicates stronger source bias. Full results on all 14 datasets appear in Appendix I.

Dataset (↓)	ANCE		coCondenser		DRAGON		RetroMAE		TAS-B	
	Original	Debias	Original	Debias	Original	Debias	Original	Debias	Original	Debias
MS MARCO	-0.042	-0.004	-0.020	-0.002	-0.083	-0.035	-0.083	-0.053	-0.121	-0.064
TREC-COVID	-0.162	-0.035	-0.340	-0.180	-0.134	-0.106	-0.194	-0.082	-0.328	-0.224
NQ	-0.042	0.003	-0.072	-0.030	-0.099	-0.050	-0.060	-0.029	-0.078	-0.055
FiQA-2018	-0.179	0.006	-0.219	-0.142	-0.161	-0.123	-0.205	-0.127	-0.170	-0.127
SCIDOCS	-0.040	-0.017	-0.058	-0.017	-0.048	-0.035	-0.073	-0.002	-0.054	-0.021
Average	-0.093 (100%)	-0.009 (10%)	-0.142 (100%)	-0.074 (52%)	-0.105 (100%)	-0.070 (67%)	-0.123 (100%)	-0.059 (48%)	-0.150 (100%)	-0.098 (65%)

%: 0, 20, 40, 60, 80, 100

Inference-time Intervention: Suppressing Artifact-aligned Directions. We next examine whether source bias can be reduced without retraining. As shown in Section 4.2, LLM-generated passages exhibit a consistent shift in embedding space relative to human-written ones. Let

$$n = \frac{\overline{\delta}_{\mathrm{LH}}}{\|\overline{\delta}_{\mathrm{LH}}\|}$$

denote the normalized mean LLM–human displacement. Given a passage embedding $v = h_d(d)$, we suppress its component along this direction:

$$v' = v - \langle v, n \rangle n.$$

We estimate n through *leave-one-dataset-out calibration*, treating MS MARCO, DL19, and DL20 as one group because they share the same collection. For each target dataset, we sample 100 human–LLM passage pairs from each of the remaining dataset groups, yielding 1100 pairs in total, and average their displacement vectors. This protocol leverages the cross-dataset consistency of the LLM–human displacement and yields a transferable estimate of the source-aligned direction.

We apply this projection to five relevance-supervised retrievers, where source bias is most pronounced. As shown in Table 4, the intervention reduces source preference in most cases. Retrieval effectiveness is largely preserved, with NDCG@5 remaining nearly unchanged across retrievers. Full results are reported in Appendix I. The method requires no retraining and only applies a simple linear projection to passage embeddings, making it a lightweight post-hoc correction.

Summary. The two interventions target the proposed mechanism at complementary stages of the retrieval pipeline. At training time, reducing the artifact mismatch between positives and negatives weakens LLM preference, while increasing this mismatch strengthens it. At inference time, removing a source-aligned embedding component provides a lightweight correction that reduces source bias without retraining. These results provide intervention-based evidence for our mechanism: source bias can be amplified by artifact imbalances in relevance supervision and mitigated by controlling or suppressing the corresponding signals.

6 Conclusion

We provide a data-centric account of source bias in neural retrieval, showing that relevance supervision can inadvertently encourage retrievers to prefer LLM-generated passages. Our analyses suggest that this behavior arises in part from surface-level artifact imbalances in supervised retrieval data: positives and negatives differ in properties such as fluency and lexical specificity, and these differences mirror those between LLM-generated and human-written passages. Embedding-space evidence further shows that the positive–negative supervision direction aligns with the stable LLM–human shift, suggesting that artifact differences learned from the training contrast can favor LLM-generated passages at inference time.

We test this mechanism through two interventions. Reducing artifact mismatch during negative sampling and removing a source-aligned embedding component at inference both reduce source preference while largely preserving retrieval effectiveness. Together, these findings show that source bias is not merely a model-side phenomenon, but is also shaped by relevance supervision, highlighting artifact-aware data construction and lightweight embedding-level correction as promising directions for robust retrieval in mixed human–LLM corpora.

7 Limitations

Our artifact analysis is necessarily incomplete. PPL and IDF are only two interpretable examples, and source-related artifacts may be diverse, entangled, and difficult to enumerate exhaustively. Moreover, perfectly controlling for these artifacts in retrieval training is extremely challenging, as changing negatives may also change hardness, topical overlap, and false-negative rates. Therefore, although our model comparisons, fine-tuning experiments, and negative-sampling interventions support the role of relevance supervision, they cannot fully exclude other co-varying factors. Our results suggest that relevance supervision is a plausible contributor to source bias, rather than its sole cause.

8 Ethical Considerations

All experiments use public datasets and models and do not involve personal information. Although source bias may affect content exposure, our work focuses on understanding its mechanism, and the direct ethical risks of our experiments are limited.

References

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and 1 others. 2020. Overview of touché 2020: argument retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 384–395. Springer.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval*, pages 716–722, Cham. Springer International Publishing.

Xiaoyang Chen, Ben He, Hongyu Lin, Xianpei Han, Tianshu Wang, Boxi Cao, Le Sun, and Yingfei Sun. 2024. Spiral of silence: How is large language model killing information retrieval?—a case study on open domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14930–14951, Bangkok, Thailand. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track. *Preprint*, arXiv:2102.07662.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the trec 2019 deep learning track. *Preprint*, arXiv:2003.07820.

Sunhao Dai, Weihao Liu, Yuqi Zhou, Liang Pang, Rongju Ruan, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. 2024a. Cocktail: A comprehensive information retrieval benchmark with llm-generated documents integration. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7052–7074, Bangkok, Thailand. Association for Computational Linguistics.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024b. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 6437–6447, New York, NY, USA. Association for Computing Machinery.

Sunhao Dai, Yuqi Zhou, Liang Pang, Zhuoyang Li, Zhaocheng Du, Gang Wang, and Jun Xu. 2025. Mitigating source bias with llm alignment. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pages 370–380, New York, NY, USA. Association for Computing Machinery.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024c. Neural retrievers are biased towards llm-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 526–537, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mohsen Fayyaz, Ali Modarressi, Hinrich Schuetze, and Nanyun Peng. 2025. Collapse of dense retrievers: Short, early, and literal biases outranking factual evidence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9136–9152, Vienna, Austria. Association for Computational Linguistics.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1265–1268, New York, NY, USA. Association for Computing Machinery.

Erik Hatcher and Otis Gospodnetic. 2004. *Lucene in action (in action series)*. Manning Publications Co.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 113–122, New York, NY, USA. Association for Computing Machinery.

Wei Huang, Keping Bi, Yinqiong Cai, Wei Chen, Jiafeng Guo, and Xueqi Cheng. 2025. How do llm-generated texts impact term-based retrieval models? *Preprint*, arXiv:2508.17715.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Vladimir Karpukhin, Barlas Oguz, 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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Inc. NetEase Youdao. 2023. Bcembedding: Bilingual and crosslingual embedding for rag. `https://github.com/netease-youdao/BCEmbedding`.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*. NIPS Workshop.

Nandan Thakur, Luiz Bonifacio, Maik Fröbe, Alexander Bondarenko, Ehsan Kamalloo, Martin Potthast, Matthias Hagen, and Jimmy Lin. 2024. Systematic evaluation of neural retrieval models on the touché 2020 argument retrieval subset of beir. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1420–1430.

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 *Thirty-Fifth Conference on Neural Information Processing Systems*.

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.

Roman Vershynin. 2018. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1):1:1–1:12.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Haoyu Wang, Sunhao Dai, Haiyuan Zhao, Liang Pang, Xiao Zhang, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. 2025. Perplexity trap: Plm-based retrievers overrate low perplexity documents. *arXiv preprint arXiv:2503.08684*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Yuxin Wang, Qingxuan Sun, and Sicheng He. 2023. M3e: Moka massive mixed embedding model.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 641–649, New York, NY, USA. Association for Computing Machinery.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024. Invisible relevance bias: Text-image retrieval models prefer ai-generated images. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 208–217, New York, NY, USA. Association for Computing Machinery.

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.

Yuqi Zhou, Sunhao Dai, Liang Pang, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. 2025. Exploring the escalation of source bias in user, data, and recommender system feedback loop. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1676–1686.

A The Use of AI Writing Assistants

In this study, we used AI writing assistants solely to improve the clarity and readability of the text. The models/tools were not used for research ideation, literature review, methodological development, experimental design, data analysis, result generation, or interpretation. All scientific content, analyses, conclusions, and responsibility for the work are entirely those of the authors.

B Reproducibility Resources

To facilitate reproducibility, we provide the datasets, model checkpoints, and main implementation details used in this work. All datasets and models are obtained from publicly available HuggingFace releases or their official websites. Our usage follows the respective licenses and research-only terms of the original sources. Unless otherwise stated, we use the official dataset splits and evaluation files provided by Cocktail and follow the standard BEIR-style dense retrieval evaluation protocol.

B.1 Datasets

Table 5 lists the Cocktail datasets used in our experiments. Each Cocktail dataset augments a BEIR collection with LLM-generated counterparts for human-written passages. For source-preference evaluation, we retrieve from the mixed human–LLM corpus and compute ΔNDSR@5 over the ranked results. For retrieval-effectiveness evaluation, we report NDCG@5 using the relevance judgments provided with the benchmark.

B.2 Model Checkpoints

Table 6 lists the dense retriever checkpoints used in our experiments. For models with separate query

Table 5: Datasets used in this paper (Cocktail versions) and their HuggingFace links.

Dataset	HuggingFace Link
MS MARCO (Nguyen et al., 2016)	https://huggingface.co/datasets/IR-Cocktail/msmarco
TREC-DL'19 (Craswell et al., 2020)	https://huggingface.co/datasets/IR-Cocktail/dl19
TREC-DL'20 (Craswell et al., 2021)	https://huggingface.co/datasets/IR-Cocktail/dl20
Natural Questions (Kwiatkowski et al., 2019)	https://huggingface.co/datasets/IR-Cocktail/nq
NFCorpus (Boteva et al., 2016)	https://huggingface.co/datasets/IR-Cocktail/nfcorpus
TREC-COVID (Voorhees et al., 2021)	https://huggingface.co/datasets/IR-Cocktail/trec-covid
HotpotQA (Yang et al., 2018)	https://huggingface.co/datasets/IR-Cocktail/hotpotqa
FiQA-2018 (Maia et al., 2018)	https://huggingface.co/datasets/IR-Cocktail/fiqa
Touché-2020 (Bondarenko et al., 2020)	https://huggingface.co/datasets/IR-Cocktail/webis-touche2020
DBpedia-Entity (Hasibi et al., 2017)	https://huggingface.co/datasets/IR-Cocktail/dbpedia-entity
SCIDOCS (Cohan et al., 2020)	https://huggingface.co/datasets/IR-Cocktail/scidocs
FEVER (Thorne et al., 2018)	https://huggingface.co/datasets/IR-Cocktail/fever
Climate-FEVER (Diggelmann et al., 2021)	https://huggingface.co/datasets/IR-Cocktail/climate-fever
SciFact (Wadden et al., 2020)	https://huggingface.co/datasets/IR-Cocktail/scifact

and passage encoders, such as DRAGON, we use the corresponding released encoders. For single-encoder models, the same encoder is used for both queries and passages. We follow the official usage of each checkpoint, including its tokenizer, pooling strategy, embedding normalization, and input prefixes when applicable. For example, E5-style models are evaluated with their official query and passage prefixes.

B.3 Implementation Details

For all dense retrieval experiments, we encode queries and passages independently and follow the usage recommended by each released checkpoint for similarity computation, pooling, embedding normalization, and input prefixes when applicable. The maximum query and passage lengths are set to 32 and 512 tokens, respectively.

For the fine-tuning experiments in RQ1, we fine-tune Contriever, E5-unsup, and SimCSE on MS MARCO using the same contrastive retrieval objective and training pipeline. The loss is InfoNCE with dot-product similarity. Following the BEIR training script, we train for 10 epochs using AdamW with a learning rate of 2×10^{-5}, a batch size of 75, and 1,000 warmup steps. The remaining optimizer settings follow the SentenceTransformers defaults. All runs use random seed 42. The same maximum query and passage lengths as in evaluation are used during fine-tuning.

For the negative-sampling intervention in RQ3, we keep the model architecture, loss form, optimizer, learning rate, batch size, and training steps fixed, and vary only the negative sampling strategy. The *in-batch only* setting uses other queries' annotated positives in the same mini-batch as negatives. The *standard* setting combines in-batch negatives with one mined hard negative per query. The *hard-neg only* setting uses one mined hard negative per query without in-batch negatives; this is implemented with a custom loss that computes the contrastive objective only over each query's positive and sampled hard negative. All mined hard negatives are sampled from the official MS MARCO hard-negative file.

For inference-time debiasing, we follow the leave-one-dataset-out calibration protocol in Section 5, sampling 100 human–LLM passage pairs from each non-target dataset group to estimate the LLM–human direction. The projection is applied only to passage embeddings, without post-projection renormalization.

Unless otherwise noted, all experiments are conducted on a single NVIDIA RTX A6000 GPU with 48 GB of memory.

C Dataset Statistics

Table 7 summarizes the statistics of the 14 datasets used in this paper. This table is adapted from the Cocktail benchmark (Dai et al., 2024a), with minor modifications.

D Retrieval Effectiveness of Evaluated Models

For completeness, we report the retrieval effectiveness of all evaluated models on the Cocktail benchmark. Table 8 presents NDCG@5 across 14 datasets for the 13 retrievers spanning the three model families. Table 9 further reports results after fine-tuning unsupervised retrievers on MS MARCO. These results complement the source preference analyses in Section 3.

Table 6: Dense retriever checkpoints used in this paper and their HuggingFace links.

Model	HuggingFace Link
Relevance-Supervised Retrievers	
ANCE (Xiong et al., 2020)	https://huggingface.co/sentence-transformers/msmarco-roberta-base-ance-firstp
TAS-B (Hofstätter et al., 2021)	https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b
coCondenser (Gao and Callan, 2022)	https://huggingface.co/sentence-transformers/msmarco-bert-co-condensor
RetroMAE (Xiao et al., 2022)	https://huggingface.co/nthakur/RetroMAE_BEIR
DRAGON (query encoder) (Lin et al., 2023)	https://huggingface.co/nthakur/dragon-plus-query-encoder
DRAGON (corpus encoder) (Lin et al., 2023)	https://huggingface.co/nthakur/dragon-plus-context-encoder
General-Purpose Embedding Models	
BGE-base (Xiao et al., 2024)	https://huggingface.co/BAAI/bge-base-en-v1.5
BCE (NetEase Youdao, 2023)	https://huggingface.co/maidalun1020/bce-embedding-base_v1
GTE (Li et al., 2023)	https://huggingface.co/thenlper/gte-base
E5 (Wang et al., 2024)	https://huggingface.co/intfloat/e5-base-v2
M3E (Wang et al., 2023)	https://huggingface.co/moka-ai/m3e-base
Unsupervised Retrievers	
Contriever (Izacard et al., 2022)	https://huggingface.co/nishimoto/contriever-sentencetransformer
E5-Unsupervised (Wang et al., 2024)	https://huggingface.co/intfloat/e5-base-unsupervised
SimCSE (Gao et al., 2021)	https://huggingface.co/princeton-nlp/unsup-simcse-bert-base-uncased

Table 7: Statistics of the 14 datasets in the Cocktail benchmark used in this paper. Avg. D/Q denotes the average number of relevant documents per query. This table is adapted from Dai et al. (2024a).

Dataset	Domain	Task	Relevancy	#Pairs	#Queries	#Corpus	Avg. D/Q	Avg. Length (Q / Human / LLM)
MS MARCO	Misc.	Passage Retrieval	Binary	532,663	6,979	542,203	1.1	6.0 / 58.1 / 55.1
DL19	Misc.	Passage Retrieval	Binary	-	43	542,203	95.4	5.4 / 58.1 / 55.1
DL20	Misc.	Passage Retrieval	Binary	-	54	542,203	66.8	6.0 / 58.1 / 55.1
TREC-COVID	Biomedical	Biomedical IR	3-level	-	50	128,585	430.1	10.6 / 197.6 / 165.9
NFCorpus	Biomedical	Biomedical IR	3-level	110,575	323	3,633	38.2	3.3 / 221.0 / 206.7
NQ	Wikipedia	QA	Binary	-	3,446	104,194	1.2	9.2 / 86.9 / 81.0
HotpotQA	Wikipedia	QA	Binary	169,963	7,405	111,107	2.0	17.7 / 67.9 / 66.6
FiQA-2018	Finance	QA	Binary	14,045	648	57,450	2.6	10.8 / 133.2 / 107.8
Touché-2020	Misc.	Argument Retrieval	3-level	-	49	101,922	18.4	6.6 / 165.4 / 134.4
DBpedia	Wikipedia	Entity Retrieval	3-level	-	400	145,037	37.3	5.4 / 53.1 / 54.0
SCIDOCS	Scientific	Citation Prediction	Binary	-	1,000	25,259	4.7	9.4 / 169.7 / 161.8
FEVER	Wikipedia	Fact Checking	Binary	140,079	6,666	114,529	1.2	8.1 / 113.4 / 91.1
Climate-FEVER	Wikipedia	Fact Checking	Binary	-	1,535	101,339	3.0	20.2 / 99.4 / 81.3
SciFact	Scientific	Fact Checking	Binary	919	300	5,183	1.1	12.4 / 201.8 / 192.7

E Formal Statements and Proofs

We formalize a mechanism by which artifact differences between positives and negatives can enter the contrastive retrieval objective. The analysis proceeds in three steps: (1) deriving the population-optimal scorer under InfoNCE, (2) decomposing the resulting density ratio into semantic and artifact-related contributions, and (3) relating this decomposition to a local embedding-space view for dot-product retrievers.

Notation and Setting. Let q denote a query and d a document. Each document d is associated with semantic features M_d and artifact features A_d, such as perplexity, lexical specificity, or stylistic attributes. We treat (M_d, A_d) as random variables induced by a measurable feature map applied to documents.

A dense retriever consists of query and document encoders $h_q(q), h_d(d) \in \mathbb{R}^m$, with a typical dot-product score

$$s_\theta(q, d) = \langle h_q(q), h_d(d) \rangle.$$

Let $p_{\text{pos}}(q, d)$ denote the distribution of positive query–document pairs. We first consider the idealized case in which negatives are sampled independently from a query-independent marginal distribution $p_{\text{neg}}(d) = p(d)$. This setting matches the standard density-ratio analysis of InfoNCE and is closest to random or in-batch negative sampling. For query-dependent negative samplers, such as mined hard negatives, the same derivation yields a density ratio with $p_{\text{neg}}(d \mid q)$ in the denominator.

Table 8: NDCG@5 results across 14 datasets for 13 dense retrievers. Higher is better.

Dataset (↓)	Relevance-Supervised Retrievers					General-Purpose Embedding Models					Unsupervised Retrievers		
	ANCE	TAS-B	coCondenser	RetroMAE	DRAGON	BGE	BCE	GTE	E5	M3E	Contriever	E5-Unsup	SimCSE
MS MARCO	0.647	0.680	0.683	0.688	0.735	0.688	0.590	0.688	0.702	0.473	0.504	0.575	0.245
DL19	0.686	0.760	0.734	0.743	0.771	0.755	0.708	0.750	0.747	0.507	0.515	0.624	0.346
DL20	0.701	0.724	0.708	0.751	0.758	0.729	0.651	0.718	0.743	0.489	0.492	0.597	0.289
NQ	0.640	0.708	0.711	0.746	0.790	0.778	0.625	0.789	0.790	0.494	0.623	0.737	0.353
NFCorpus	0.266	0.340	0.345	0.336	0.389	0.403	0.275	0.394	0.368	0.257	0.339	0.371	0.109
TREC-COVID	0.671	0.670	0.677	0.735	0.678	0.783	0.574	0.763	0.714	0.390	0.391	0.605	0.296
HotpotQA	0.553	0.705	0.663	0.747	0.799	0.792	0.533	0.761	0.801	0.575	0.650	0.668	0.369
FiQA-2018	0.275	0.408	0.467	0.498	0.529	0.384	0.285	0.380	0.373	0.366	0.225	0.373	0.093
Touché-2020	0.479	0.427	0.349	0.441	0.390	0.402	0.333	0.423	0.411	0.242	0.308	0.333	0.252
DBpedia	0.408	0.493	0.493	0.528	0.533	0.514	0.360	0.514	0.541	0.370	0.427	0.488	0.259
SCIDOCS	0.095	0.111	0.102	0.116	0.123	0.177	0.118	0.190	0.141	0.069	0.114	0.174	0.041
FEVER	0.820	0.835	0.842	0.870	0.876	0.928	0.682	0.924	0.905	0.865	0.878	0.925	0.510
Climate-FEVER	0.270	0.306	0.255	0.311	0.318	0.368	0.274	0.373	0.303	0.161	0.223	0.264	0.195
SciFact	0.465	0.602	0.549	0.611	0.631	0.715	0.533	0.732	0.688	0.448	0.614	0.719	0.239

Table 9: NDCG@5 results of unsupervised retrievers after MS MARCO fine-tuning, corresponding to the same base models in Table 8. The "-FT" suffix denotes fine-tuning on MS MARCO.

Dataset	Relevance-Supervised Retrievers		
	Contriever-FT	E5-FT	SimCSE-FT
MS MARCO	0.676	0.711	0.630
DL19	0.696	0.763	0.727
DL20	0.673	0.720	0.703
NQ	0.732	0.764	0.670
NFCorpus	0.339	0.378	0.279
TREC-COVID	0.446	0.731	0.590
HotpotQA	0.712	0.735	0.577
FiQA-2018	0.255	0.336	0.220
Touché-2020	0.347	0.428	0.389
DBpedia	0.495	0.532	0.444
SCIDOCS	0.117	0.138	0.083
FEVER	0.857	0.895	0.837
Climate-FEVER	0.289	0.312	0.261
SciFact	0.593	0.679	0.470

We use the term *artifact imbalance* to refer to empirical differences in artifact distributions between positives and negatives. A marginal diagnostic is

$$\Delta_A = \mathbb{E}[A_{d^+}] - \mathbb{E}[A_{d^-}],$$

but the theoretical condition below concerns artifact differences after conditioning on query and semantic features.

Step 1: Optimal scorer under InfoNCE. InfoNCE encourages the model to identify the positive passage among a set of candidates. Under population risk minimization with an unrestricted scoring function, its optimal score has a density-ratio form.

Lemma 1. *For InfoNCE with negatives sampled independently from $p(d)$, the population-optimal unrestricted scorer is equivalent, up to an additive function of q, to*

$$s^*(q,d) = \log \frac{p_{\text{pos}}(d \mid q)}{p(d)} + C(q).$$

Equivalently, when the query marginal under positive-pair sampling is $p(q)$,

$$s^*(q,d) = \log \frac{p_{\text{pos}}(q,d)}{p(q)p(d)} + C(q).$$

Insight 1: Under independent negative sampling and an unrestricted scorer, the population optimum of InfoNCE has a log-density-ratio form.

Step 2: Decomposition into semantic and artifact contributions. We next decompose the density ratio in the feature space (M_d, A_d). The decomposition separates a component depending only on the query and semantic features from a residual component that captures the additional contribution of artifacts conditional on those variables.

Proposition 1 (Conditional artifact contribution)**.** *Let (M_d, A_d) be semantic and artifact features induced by a measurable feature map. Consider the population-optimal scorer under a negative sampler $p_{\text{neg}}(d \mid q)$:*

$$s^*(q,d) = \log \frac{p_{\text{pos}}(d \mid q)}{p_{\text{neg}}(d \mid q)} + C(q).$$

In the feature space (M_d, A_d), the corresponding density ratio can be written as

$$s^*(q,d) = \phi(q, M_d) + \psi(q, M_d, A_d) + C(q),$$

where $\phi(q, M_d)$ collects terms depending only on (q, M_d), and $\psi(q, M_d, A_d)$ captures the conditional artifact contribution.

If $p_{\text{pos}}(A_d \mid q, M_d) \neq p_{\text{neg}}(A_d \mid q, M_d)$ on a set of positive measure, then $\psi(q, M_d, A_d)$ is non-constant as a function of A_d on that set. Consequently, the optimal scorer cannot in general be expressed as a function of (q, M_d) alone.

Insight 2: When artifacts distinguish positives from negatives after conditioning on query and semantic content, artifact-dependent terms can enter the optimal contrastive scorer.

Step 3: A local embedding-space view. We finally relate the preceding decomposition to dot-product retrievers. The goal is to show how artifact-related variation can appear as a local linear component in the document embedding.

Proposition 2 (Local embedding decomposition)**.** *Let the document encoder h_d be differentiable in a local feature chart (M_d, A_d). Around a reference artifact value a_0, the dot-product score admits the first-order expansion*

$$\begin{aligned} s_\theta(q,d) = &\underbrace{\langle h_q(q), h_{\mathrm{sem}}(M_d)\rangle}_{\textit{semantic reference}} \\ &+ \underbrace{\langle h_q(q), h_{\mathrm{art}}(A_d; M_d)\rangle}_{\textit{artifact perturbation}} \\ &+ \varepsilon(q, M_d, A_d), \end{aligned}$$

where

$$\begin{aligned} h_{\mathrm{sem}}(M_d) &= g(M_d, a_0), \\ h_{\mathrm{art}}(A_d; M_d) &= J_a(M_d, a_0)(A_d - a_0), \end{aligned}$$

$g(M_d, A_d)$ denotes the document embedding as a function of the feature variables, $J_a(M_d, a_0)$ is the Jacobian of g with respect to artifact features at a_0, and

$$\varepsilon(q, M_d, A_d) = o(\|A_d - a_0\|) \quad \textit{as} \quad \|A_d - a_0\| \to 0.$$

If the artifact Jacobian varies weakly with M_d, the artifact perturbation can be approximated by a shared linear direction or low-dimensional subspace.

Insight 3: A local first-order approximation suggests that artifact-related variation can appear as approximately linear components in embedding space.

Formal proofs of Lemma 1, Proposition 1, and Proposition 2 are provided below. Together, these results characterize a mechanism through which relevance supervision can encode artifact-dependent signals. When such signals align with differences between LLM-generated and human-written passages, they can contribute to source preference. The embedding-space view further motivates our empirical search for stable source-aligned directions.

E.1 Proof of Lemma 1

Proof. We derive the population-optimal scorer for InfoNCE under independent negative sampling. The proof proceeds in three steps: (i) formalize the sampling and objective, (ii) show that risk minimization matches the true posterior over the positive index, and (iii) compute this posterior.

Step 1: Sampling scheme and objective. Draw a query $q \sim p(q)$ and sample an index $I \sim \mathrm{Unif}\{0, \ldots, K\}$, where I denotes the position of the positive passage. Conditioned on (q, I), sample the positive passage $d_I \sim p_{\mathrm{pos}}(d \mid q)$ and sample negatives $d_j \sim p(d)$ for all $j \neq I$. This yields a candidate batch $\boldsymbol{d} = (d_0, \ldots, d_K)$.

Given scores $s(q, d_j)$, the model predicts

$$\pi_\theta(i \mid q, \boldsymbol{d}) = \frac{\exp(s(q, d_i))}{\sum_{j=0}^{K} \exp(s(q, d_j))}.$$

A temperature parameter can be included by rescaling the score; it does not affect the density-ratio form of the optimum.

The InfoNCE loss is the expected negative log-likelihood:

$$\mathcal{L}(\theta) = \mathbb{E}_{q,\boldsymbol{d}}\left[\mathbb{E}_{I|q,\boldsymbol{d}}\big[-\log \pi_\theta(I \mid q, \boldsymbol{d})\big]\right].$$

Let

$$P_i = \mathbb{P}(I = i \mid q, \boldsymbol{d}), \qquad \pi_i = \pi_\theta(i \mid q, \boldsymbol{d}).$$

For fixed $(q, \boldsymbol{d})$, the conditional risk is

$$R(\boldsymbol{s}; q, \boldsymbol{d}) = -\sum_{i=0}^{K} P_i \log \pi_i.$$

Step 2: Bayes optimality. The conditional risk can be written as

$$\begin{aligned} R(\boldsymbol{s}; q, \boldsymbol{d}) &= -\sum_i P_i \log P_i + \sum_i P_i \log \frac{P_i}{\pi_i} \\ &= H(P) + \mathrm{KL}(P \| \pi). \end{aligned}$$

Since $H(P)$ is independent of the model and $\mathrm{KL}(P\|\pi) \geq 0$, the risk is minimized iff $\pi = P$.

Because π is a softmax over scores, any optimal scoring function satisfies

$$s(q, d_i) = \log P_i + C(q, \boldsymbol{d}),$$

where $C(q, \boldsymbol{d})$ is shared across all candidates in the batch and therefore does not affect the softmax.

Step 3: Computing the posterior. By Bayes' rule and the sampling scheme,

$$\begin{aligned} P_i &= \mathbb{P}(I = i \mid q, \boldsymbol{d}) \\ &\propto \mathbb{P}(I = i)\, p_{\mathrm{pos}}(d_i \mid q) \prod_{j \neq i} p(d_j) \\ &= \frac{1}{K+1} p_{\mathrm{pos}}(d_i \mid q) \prod_{j \neq i} p(d_j). \end{aligned}$$

Normalizing over i gives

$$P_i = \frac{\dfrac{p_{\text{pos}}(d_i \mid q)}{p(d_i)}}{\displaystyle\sum_{j=0}^{K} \frac{p_{\text{pos}}(d_j \mid q)}{p(d_j)}}.$$

Therefore,

$$\log P_i = \log \frac{p_{\text{pos}}(d_i \mid q)}{p(d_i)} - \log \sum_{j=0}^{K} \frac{p_{\text{pos}}(d_j \mid q)}{p(d_j)}.$$

The second term is shared across all candidates in the batch and can be absorbed into the additive constant. Thus,

$$s^*(q, d) = \log \frac{p_{\text{pos}}(d \mid q)}{p(d)} + C(q).$$

If the positive-pair query marginal is $p(q)$, then

$$p_{\text{pos}}(d \mid q) = \frac{p_{\text{pos}}(q, d)}{p(q)},$$

and hence

$$s^*(q, d) = \log \frac{p_{\text{pos}}(q, d)}{p(q)p(d)} + C(q).$$

This completes the proof.

Remark. If negatives are drawn from a query-dependent distribution $p_{\text{neg}}(d \mid q)$, the same derivation gives

$$s^*(q, d) = \log \frac{p_{\text{pos}}(d \mid q)}{p_{\text{neg}}(d \mid q)} + C(q).$$

Thus, the optimal score is a density ratio between the positive-pair distribution and the actual negative sampler. □

E.2 Proof of Proposition 1

Proof. We prove the result for a general negative sampler $p_{\text{neg}}(d \mid q)$. Let (Q, M, A) denote the random variables induced by sampling (q, d) and applying the feature map $d \mapsto (M_d, A_d)$. We write the positive and negative feature-space distributions as

$$p_{\text{pos}}(m, a \mid q), \qquad p_{\text{neg}}(m, a \mid q).$$

By Lemma 1 and its query-dependent extension, the population-optimal score can be written, up to an additive function of q, as

$$s^*(q, d) = \log \frac{p_{\text{pos}}(m_d, a_d \mid q)}{p_{\text{neg}}(m_d, a_d \mid q)} + C(q),$$

where (m_d, a_d) are the feature realizations of d.

Using the chain rule,

$$\log \frac{p_{\text{pos}}(m_d, a_d \mid q)}{p_{\text{neg}}(m_d, a_d \mid q)} = \log \frac{p_{\text{pos}}(m_d \mid q)}{p_{\text{neg}}(m_d \mid q)} + \log \frac{p_{\text{pos}}(a_d \mid q, m_d)}{p_{\text{neg}}(a_d \mid q, m_d)}.$$

Define

$$\phi(q, m_d) = \log \frac{p_{\text{pos}}(m_d \mid q)}{p_{\text{neg}}(m_d \mid q)},$$

$$\psi(q, m_d, a_d) = \log \frac{p_{\text{pos}}(a_d \mid q, m_d)}{p_{\text{neg}}(a_d \mid q, m_d)}.$$

Then

$$s^*(q, d) = \phi(q, M_d) + \psi(q, M_d, A_d) + C(q).$$

If

$$p_{\text{pos}}(A_d \mid q, M_d) \neq p_{\text{neg}}(A_d \mid q, M_d)$$

on a set of positive measure, then the likelihood ratio

$$\frac{p_{\text{pos}}(A_d \mid q, M_d)}{p_{\text{neg}}(A_d \mid q, M_d)}$$

is not almost surely constant on that set. Hence $\psi(q, M_d, A_d)$ is non-constant as a function of A_d there. Consequently, $s^*(q, d)$ cannot in general be represented as a function of (q, M_d) alone. This establishes that the optimal scorer contains a conditional artifact contribution under the stated condition. □

E.3 Proof of Proposition 2

Proof. Let $g(m, a)$ denote the document embedding as a function of semantic and artifact features: $g(m, a) = h_d(d)$ for documents with $(M_d, A_d) = (m, a)$. Assume g is differentiable with respect to a in a neighborhood of a reference artifact value a_0. A first-order Taylor expansion around a_0 gives

$$g(m, a) = g(m, a_0) + J_a(m, a_0)(a - a_0) + r(m, a),$$

where

$$J_a(m, a_0) = \left[\frac{\partial g(m, a)}{\partial a}\right]_{a=a_0}$$

and

$$\|r(m, a)\| = o(\|a - a_0\|) \quad \text{as} \quad \|a - a_0\| \to 0.$$

Define

$$h_{\text{sem}}(m) = g(m, a_0),$$
$$h_{\text{art}}(a; m) = J_a(m, a_0)(a - a_0).$$

Then

$$h_d(d) = h_{\text{sem}}(M_d) + h_{\text{art}}(A_d; M_d) + r(M_d, A_d).$$

For a dot-product retriever,

$$\begin{aligned} s_\theta(q, d) &= \langle h_q(q), h_d(d) \rangle \\ &= \langle h_q(q), h_{\text{sem}}(M_d) \rangle \\ &\quad + \langle h_q(q), h_{\text{art}}(A_d; M_d) \rangle \\ &\quad + \langle h_q(q), r(M_d, A_d) \rangle. \end{aligned}$$

Let

$$\varepsilon(q, M_d, A_d) = \langle h_q(q), r(M_d, A_d) \rangle.$$

Since $\|r(M_d, A_d)\| = o(\|A_d - a_0\|)$, we have

$$\varepsilon(q, M_d, A_d) = o(\|A_d - a_0\|)$$

for fixed $h_q(q)$. Thus,

$$s_\theta(q, d) = \underbrace{\langle h_q(q), h_{\text{sem}}(M_d)\rangle}_{\text{semantic reference}} + \underbrace{\langle h_q(q), h_{\text{art}}(A_d; M_d)\rangle}_{\text{artifact perturbation}} + \varepsilon(q, M_d, A_d).$$

If $J_a(M_d, a_0)$ varies weakly with M_d, or if its variation is dominated by a shared low-dimensional component, then the artifact perturbation can be approximated by a shared direction or subspace in the embedding space. This gives the idealized linear view used to motivate the empirical embedding-direction analysis. □

Remark. The embedding-space argument is local and first-order. It does not require semantic and artifact information to be globally separable in the learned representation. Rather, it shows that, around a reference artifact value, changes in artifact features can induce approximately linear perturbations in the document embedding. This motivates examining whether LLM–human and positive–negative contrasts exhibit stable directions in the learned embedding space.

F Additional Linguistic Analyses

In this appendix, we provide supplementary analyses promised in Section 4.1. Specifically, we report (i) additional effect-size analyses for the comparisons in the main text, and (ii) results on the other 13 datasets beyond MS MARCO.

F.1 Effect-Size Analyses

We quantify the magnitude of linguistic differences using standard effect-size measures (Hedges' g for mean differences) and report associated significance levels. These statistics complement the significance tests in the main paper by showing not only whether differences are significant but also their practical magnitude. Table 10 summarizes results on MS MARCO for two contrasts: (i) positives vs. the unannotated pool, and (ii) LLM-generated vs. human-written passages.

We observe that both comparisons yield highly significant differences despite modest effect sizes. For perplexity (PPL), positives are more fluent than the unannotated pool ($g = -0.214$), and LLM passages are even more fluent than human passages ($g = -0.274$). For IDF, the effects are smaller ($g = 0.047$ and 0.145 respectively) but consistently positive, indicating that both positives and LLM rewrites exhibit slightly greater lexical specificity. Taken together, these results show that supervision and source type both introduce systematic, statistically robust shifts in linguistic features, even if the magnitudes are moderate.

Table 10: Effect sizes (Hedges' g) and significance for linguistic feature comparisons on MS MARCO. Positive values indicate higher scores for the first group. p-values smaller than numerical precision are reported as $p < 10^{-15}$.

Comparison	PPL (g)	IDF (g)	p-value
Positives vs. Unannotated	-0.214	$+0.047$	$< 10^{-15}$
LLM vs. Human	-0.274	$+0.145$	$< 10^{-15}$

F.2 Positives vs. Negatives on Additional Datasets

To assess whether the imbalance between annotated positives and negatives generalizes beyond MS MARCO, we extend the perplexity analysis to other datasets in Cocktail (Figure4). For datasets that share the same corpus (e.g., MS MARCO and DL19/20), we report results only once. For NFCorpus and HotpotQA, all passages are annotated with relevance labels, so no negative pool exists and only positives are shown. Across the remaining datasets, positives consistently exhibit lower perplexity than negatives, mirroring the trend in MS MARCO. This indicates that stylistic disparities between positives and negatives are not dataset-specific idiosyncrasies but a systematic property of retrieval supervision. As discussed in the main text, positives are often drawn from edited, high-quality sources intended to serve as good answers, whereas negatives derive from more heterogeneous and less polished text.

F.3 LLM vs. Human across Additional Datasets

To ensure that the findings generalize beyond MS MARCO, we replicate the analysis on the other datasets in Cocktail. Figure 5 reports perplexity distributions, and Figure 6 reports IDF distributions, comparing LLM-generated versus human-written passages.

Consistent with the MS MARCO case, LLM-generated passages consistently exhibit lower perplexity and slightly higher IDF than their human-written counterparts. The PPL differences are stable and consistent across all datasets, while the IDF

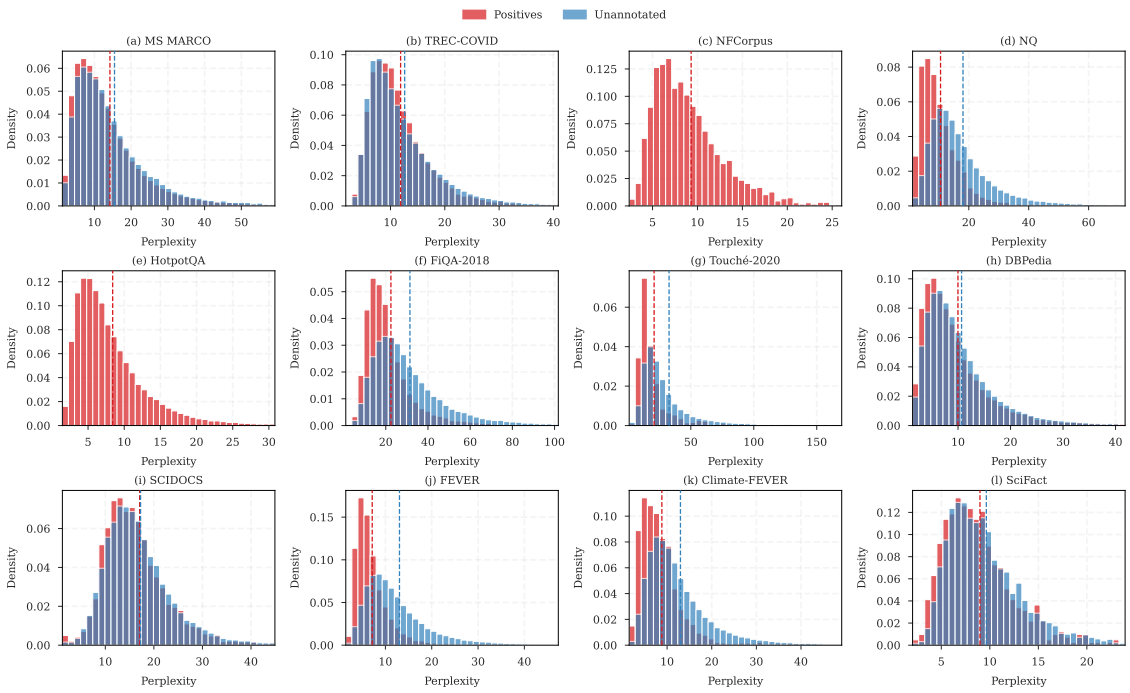


Figure 4: Perplexity distributions of positives versus negatives across retrieval datasets in Cocktail. For MS MARCO and DL19/20, results are reported once due to corpus overlap. For NFCorpus and HotpotQA, all passages are annotated as relevant, so only positive distributions are shown.

differences are more modest in magnitude but follow the same direction throughout. These results confirm that source-based stylistic artifacts are systematic and broadly consistent across domains.

G Cosine Similarity Between Random High-Dimensional Vectors

We derive the null distribution of cosine similarities between independent random vectors, which serves as the statistical baseline for our embedding-space analyses. Let $x, y \in \mathbb{R}^m$ be isotropic random vectors. Normalizing to the unit sphere ($\hat{x} = x/\|x\|$, $\hat{y} = y/\|y\|$) yields $\hat{x}, \hat{y} \sim \mathrm{Unif}(\mathbb{S}^{m-1})$, and their cosine similarity is

$$Z = \langle \hat{x}, \hat{y} \rangle \in [-1, 1].$$

By rotational invariance, Z follows a Beta-type density (Vershynin, 2018):

$$f_Z(z) \;=\; \frac{\Gamma(\frac{m}{2})}{\sqrt{\pi}\,\Gamma(\frac{m-1}{2})}(1-z^2)^{\frac{m-3}{2}}, \quad z \in [-1, 1],$$

which is symmetric around zero. Equivalently, the tail probability can be expressed via the regularized incomplete Beta function:

$$\Pr(|Z| > t) \;=\; \mathrm{I}_{1-t^2}\!\left(\tfrac{m-1}{2}, \tfrac{1}{2}\right).$$

By symmetry, $\mathbb{E}[Z] = 0$. Since each coordinate of a uniform unit vector has variance $1/m$, the variance of Z is

$$\mathrm{Var}(Z) \;=\; \frac{1}{m}.$$

For large m, the density concentrates sharply at zero. Expanding $\log(1-z^2) \approx -z^2$ near the origin gives the Gaussian approximation

$$Z \;\approx\; \mathcal{N}\!\left(0, \tfrac{1}{m}\right).$$

In dimension $m = 768$, the standard deviation is $\sigma = 1/\sqrt{m} \approx 0.0361$, so that $3\sigma \approx 0.108$. Under the normal approximation,

$$\Pr(|Z| > 3\sigma) \approx 0.27\%,$$

which closely matches the exact Beta distribution. Thus, over 99.7% of random pairs fall within $\pm 3\sigma$, validating the use of this threshold as a significance criterion in high-dimensional embedding spaces. Figure 7 illustrates this concentration.

H Additional Embedding Analyses

This appendix provides two complementary sets of embedding-space analyses. First, we extend the representative results in the main text to the full set of unique corpora in the Cocktail benchmark, using the DRAGON retriever as a representative model. Second, because dense retriever embeddings are not necessarily isotropic, we report an empirical paired sign-flip validation that uses the observed paired displacements to construct a more realistic null distribution.

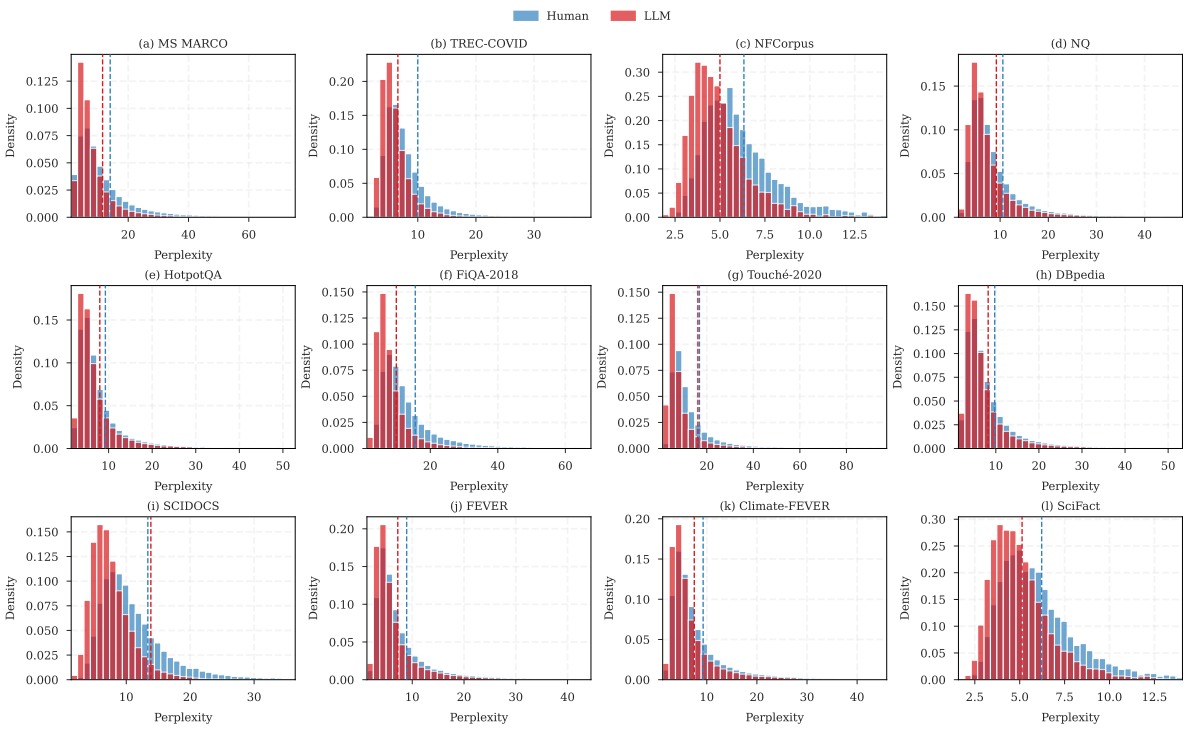


Figure 5: Perplexity (PPL) distributions of LLM-generated vs. human-written passages across additional datasets. Red = LLM, Blue = Human. LLM passages consistently exhibit lower perplexity, indicating higher fluency.

H.1 Additional Embedding Analyses on the Full Benchmark

Our experiments use 14 datasets from the Cocktail benchmark. Since three of them (MS MARCO, DL19, and DL20) share the same underlying corpus, we report visualization results at the corpus level, resulting in 12 unique corpora. These figures complement the representative results shown in the main text and report: (1) within-dataset displacement consistency (Figure 8), (2) cross-dataset similarity of mean displacement directions (Figure 9), and (3) alignment between LLM–human and supervision-induced directions (Figure 10).

Random-vector reference. The dashed 3σ reference lines in Figures 8 and 10 are based on the cosine similarities of random vectors in the same embedding dimensionality. We use this as a descriptive geometric reference: in high-dimensional space, independent isotropic random vectors are nearly orthogonal, so values outside the 3σ range indicate unusually strong alignment compared with idealized random directions. However, dense retriever embeddings are not isotropic and may contain common components, dataset-level structure, and anisotropy. Therefore, we complement this reference with the empirical paired sign-flip validation in Appendix H.2.

Overall, these full-benchmark visualizations extend the main-text findings to the full set of unique corpora. The majority of corpora follow the same trends as reported in the main text, while a small number exhibit weaker effects, which we treat as exceptions rather than contradictions.

H.2 Empirical Paired Sign-Flip Validation

The random-vector reference above provides an idealized geometric baseline, but it does not fully capture the geometry of real dense retriever embeddings. Retriever embeddings may be anisotropic and may contain common components, dataset-level structure, or other non-random regularities. We therefore complement the random-vector reference with a paired sign-flip permutation test.

For each human–LLM passage pair, we define the LLM–human displacement as

$$\delta_i^{\mathrm{LH}} = h_d(d_i^{\mathrm{LLM}}) - h_d(d_i^{\mathrm{Human}}).$$

In each sign-flip permutation, we randomly multiply each paired displacement by $+1$ or -1:

$$\delta_i^{\pi} = z_i \delta_i^{\mathrm{LH}}, \qquad z_i \sim \mathrm{Unif}\{-1, +1\}.$$

This procedure preserves the magnitude of each displacement, the semantic pairing between the human-written passage and its LLM-generated counterpart, the dataset composition, and the empirical geometry of the retriever embedding space.

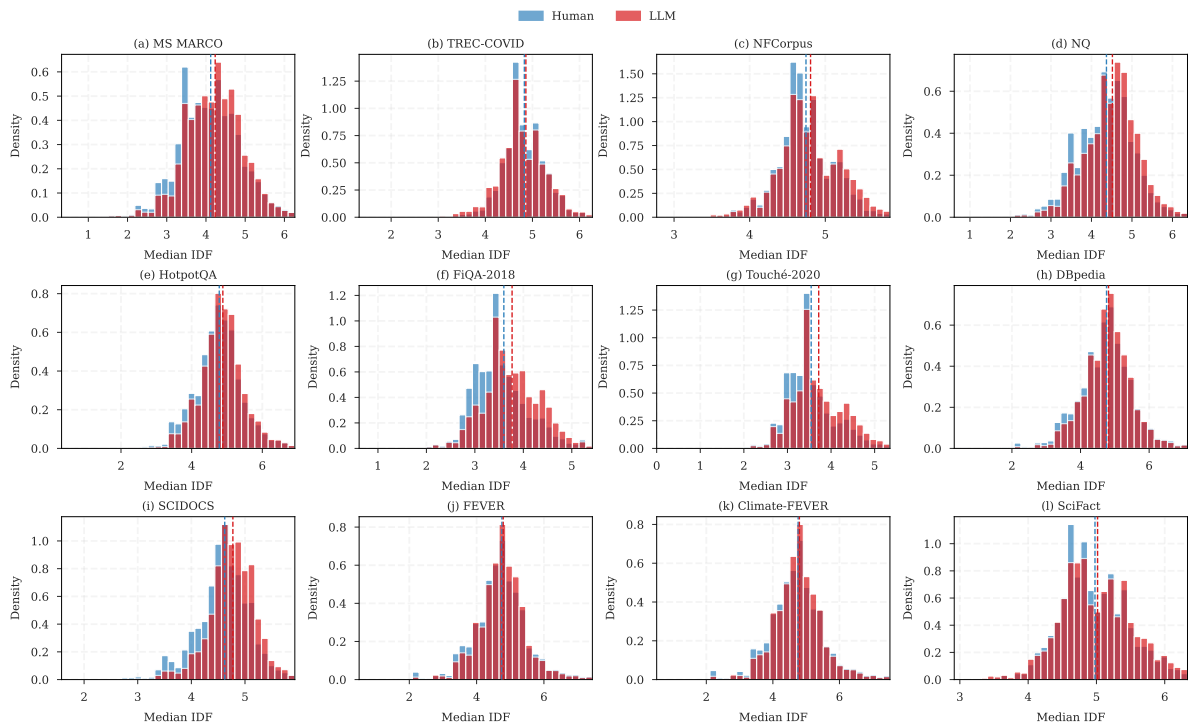


Figure 6: Median IDF distributions of LLM-generated vs. human-written passages across additional datasets. Red = LLM, Blue = Human. LLM passages generally exhibit higher IDF, though the gap varies across datasets.

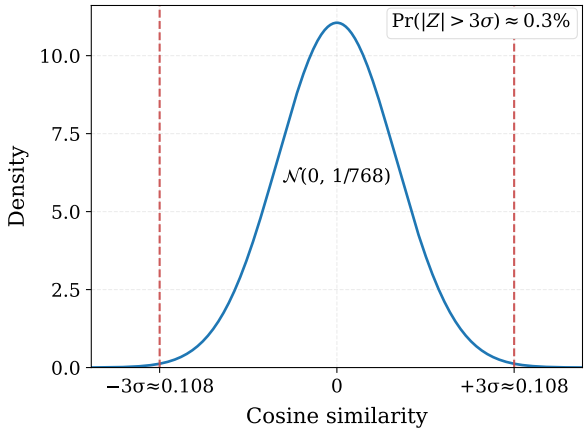


Figure 7: Null distribution of cosine similarity between random vectors in $m = 768$ dimensions, approximated by $\mathcal{N}(0, 1/m)$. Over 99.7% of values lie within $\pm 3\sigma \approx 0.108$, supporting its use as a significance criterion.

At the same time, it removes any systematic orientation from human-written to LLM-generated text. The resulting sign-randomized distribution therefore serves as an empirical null for testing whether the observed directionality is stronger than expected under the same embedding geometry.

For a statistic T, let T_{obs} denote the value computed from the original displacements, and let $T_1^\pi, \ldots, T_B^\pi$ denote the values computed from B sign-flip permutations. We use the one-sided empirical permutation p-value

$$p = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left[T_b^\pi \geq T_{\text{obs}}\right]}{B+1}.$$

The additive correction prevents zero p-values with a finite number of Monte Carlo permutations. We use $B = 1000$ permutations, so the smallest attainable corrected permutation p-value is $1/(B+1) = 0.000999$.

We apply this validation to the three embedding-space quantities analyzed in the main text and in Appendix H.1.

Within-dataset LLM–human consistency. For each dataset D, we first compute the dataset-level mean LLM–human displacement

$$\bar{\delta}_{\text{LH},D} = \frac{1}{N_D} \sum_{i \in D} \delta_i^{\text{LH}}.$$

We then measure within-dataset consistency as the average alignment between each pair-level LLM–human displacement and this dataset-level mean displacement:

$$T_{\text{obs}}^{\text{within}}(D) = \frac{1}{N_D} \sum_{i \in D} \cos\left(\delta_i^{\text{LH}}, \bar{\delta}_{\text{LH},D}\right).$$

This statistic matches the within-dataset consistency analyzed in Appendix H.1: it asks whether individual LLM–human displacements within the same dataset consistently point toward the dataset-level source direction.

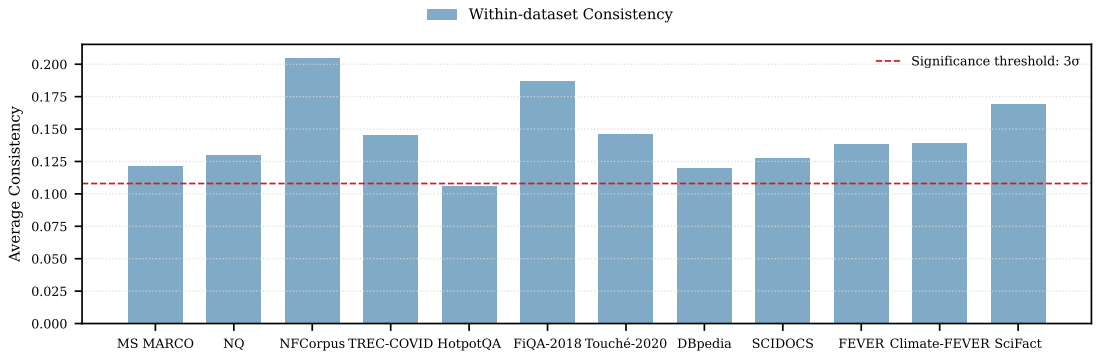


Figure 8: Within-dataset consistency of LLM–Human displacements. Bars show average pairwise cosine similarity among displacement vectors δ_i^{LH} within each dataset, relative to the 3σ random-vector reference. Most datasets exceed the reference range, with a few exceptions near or below it.

For each sign-flip permutation, we replace each displacement with $\delta_i^{\pi} = z_i \delta_i^{\mathrm{LH}}$ and recompute the permuted dataset-level mean direction

$$\overline{\delta}^{\pi}_{\mathrm{LH},D} = \frac{1}{N_D} \sum_{i \in D} z_i \delta_i^{\mathrm{LH}}.$$

The corresponding null statistic is

$$T_{\pi}^{\mathrm{within}}(D) = \frac{1}{N_D} \sum_{i \in D} \cos\left(z_i \delta_i^{\mathrm{LH}}, \overline{\delta}^{\pi}_{\mathrm{LH},D}\right).$$

This tests whether pair-level LLM–human displacements are more aligned with their dataset-level mean source direction than expected after randomizing the human-to-LLM orientation within each paired displacement.

Cross-dataset LLM–human consistency. For each dataset D, we compute the dataset-level mean displacement

$$\overline{\delta}_{\mathrm{LH},D} = \frac{1}{N_D} \sum_{i \in D} \delta_i^{\mathrm{LH}}.$$

The observed cross-dataset consistency is the average cosine similarity between dataset-level mean directions:

$$T_{\mathrm{obs}}^{\mathrm{cross}} = \frac{2}{K(K-1)} \sum_{a<b} \cos\left(\overline{\delta}_{\mathrm{LH},D_a}, \overline{\delta}_{\mathrm{LH},D_b}\right),$$

where K is the number of dataset-level collections. For each permutation, we independently sign-flip the paired LLM–human displacements within each dataset, recompute the dataset-level mean directions,

$$\overline{\delta}^{\pi}_{\mathrm{LH},D} = \frac{1}{N_D} \sum_{i \in D} z_i \delta_i^{\mathrm{LH}},$$

and then recompute the same cross-dataset statistic:

$$T_{\pi}^{\mathrm{cross}} = \frac{2}{K(K-1)} \sum_{a<b} \cos\left(\overline{\delta}^{\pi}_{\mathrm{LH},D_a}, \overline{\delta}^{\pi}_{\mathrm{LH},D_b}\right).$$

This tests whether the observed cross-dataset alignment remains stronger than expected after randomizing the human-to-LLM orientation within each dataset.

Positive–negative and LLM–human alignment. We also validate the alignment between the supervision-induced positive–negative direction and the LLM–human source direction. The positive–negative direction is estimated from MS MARCO as

$$\overline{\delta}_{\mathrm{PN}} = \mathbb{E}\left[h_d(d^+) - h_d(d^-)\right],$$

where d^+ is an annotated positive passage and d^- is a BM25-retrieved negative passage. We treat $\overline{\delta}_{\mathrm{PN}}$ as a fixed supervision-induced reference direction. For each dataset D, the observed PN–LH alignment is

$$T_{\mathrm{obs}}^{\mathrm{PN\text{-}LH}}(D) = \cos\left(\overline{\delta}_{\mathrm{PN}}, \overline{\delta}_{\mathrm{LH},D}\right).$$

The sign-flip null keeps $\overline{\delta}_{\mathrm{PN}}$ fixed and applies sign flips only to the LLM–human displacements:

$$T_{\pi}^{\mathrm{PN\text{-}LH}}(D) = \cos\left(\overline{\delta}_{\mathrm{PN}}, \overline{\delta}^{\pi}_{\mathrm{LH},D}\right).$$

This tests whether the observed LLM–human direction is more strongly aligned with the positive–negative supervision direction than expected if the human-to-LLM orientation were randomized within each pair.

Aggregating permutation results. For dataset-level quantities, namely within-dataset LH consistency and PN–LH alignment, we report an overall statistic by macro-averaging over datasets. Specifically, for a statistic $T(D)$ defined on dataset D, the observed overall value is

$$T_{\mathrm{obs}}^{\mathrm{overall}} = \frac{1}{K} \sum_{D=1}^{K} T_{\mathrm{obs}}(D),$$

where K is the number of dataset-level collections. The corresponding overall null distribution is constructed by applying the same macro-average within each sign-flip permutation:

$$T_{\pi_b}^{\mathrm{overall}} = \frac{1}{K} \sum_{D=1}^{K} T_{\pi_b}(D).$$

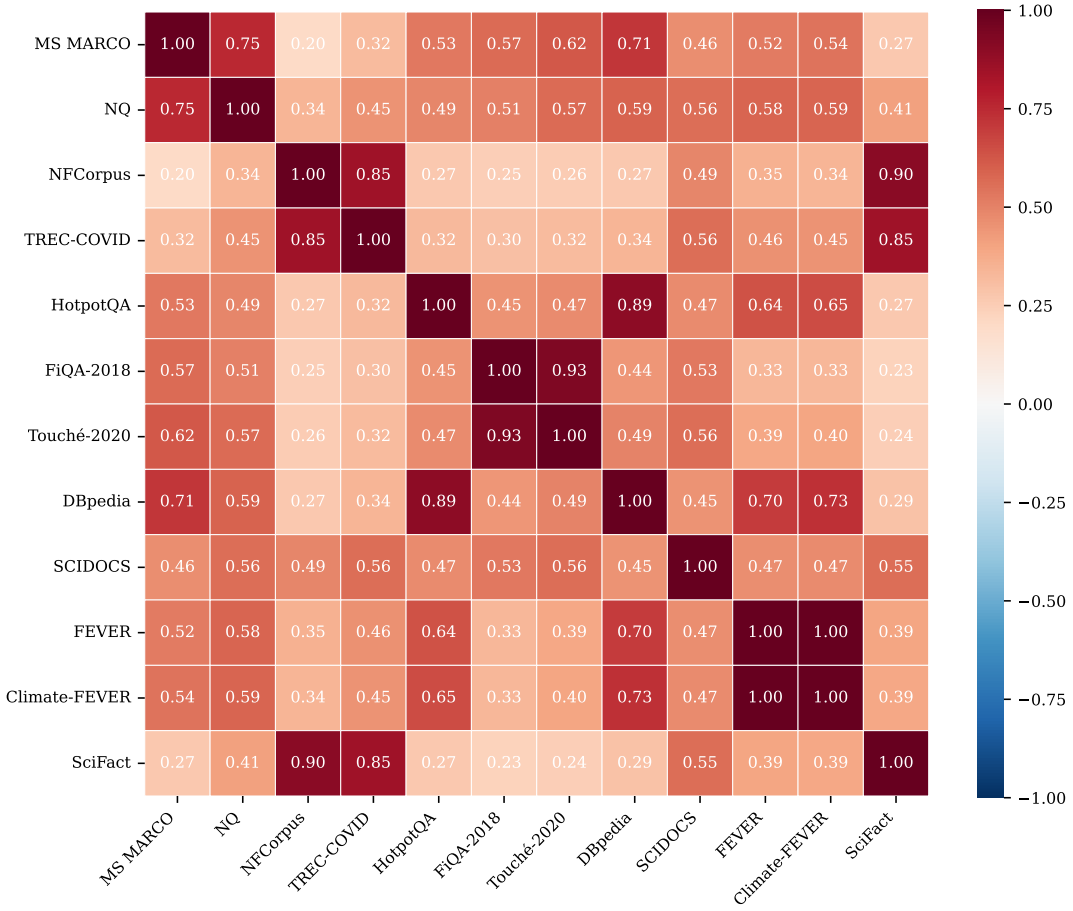


Figure 9: Cross-dataset similarity of mean LLM–Human displacement directions. Values denote cosine similarity between dataset-level means $\bar{\delta}_{\mathrm{LH},D}$. Darker cells indicate stronger alignment, revealing consistent source-related directions across corpora.

The null mean and standard deviation in Table 11 are computed from $\{T^{\mathrm{overall}}_{\pi_b}\}_{b=1}^{B}$, and the reported p-value compares $T^{\mathrm{overall}}_{\mathrm{obs}}$ against this overall null distribution. We do not average per-dataset p-values.

For cross-dataset LH consistency, the statistic is already defined globally as the average cosine similarity over all pairs of dataset-level mean directions. Therefore, its observed value and null distribution are computed directly from the cross-dataset statistic for each sign-flip permutation. All sign-flip validations are computed over the 14 dataset-level collections used in the benchmark.

Table 11 summarizes the paired sign-flip validation using DRAGON. All three observed statistics are substantially larger than their corresponding sign-flip null distributions. With $B = 1000$ permutations, the smallest attainable permutation p-value under the additive correction is $1/(B + 1) = 0.000999$, reported as 0.0010 in the table. Thus, the within-dataset LH consistency, cross-dataset LH consistency, and PN–LH alignment are all significant under the paired sign-flip test. These results support the conclusion that the observed embedding-space patterns are not artifacts of the idealized random-vector reference or of the marginal geometry of the embedding space.

Metric	Obs.	Null μ	Null σ	p
Within-dataset LH	0.1436	0.0044	0.0001	0.0010
Cross-dataset LH	0.5089	0.0001	0.0053	0.0010
PN–LH alignment	0.2692	0.0015	0.0240	0.0010

Table 11: Summary of paired sign-flip validation using DRAGON over the 14 dataset-level collections. Obs. denotes the observed statistic; Null μ and Null σ denote the mean and standard deviation of the corresponding overall sign-flip null distribution. For dataset-level metrics, the overall statistic is obtained by macro-averaging over datasets within each permutation. With $B = 1000$ permutations, the minimum attainable corrected permutation p-value is $1/(B + 1) = 0.000999$.

I Additional Results for RQ3

In this section, we provide the supplementary results for Section 5, including (a) retrieval effective-

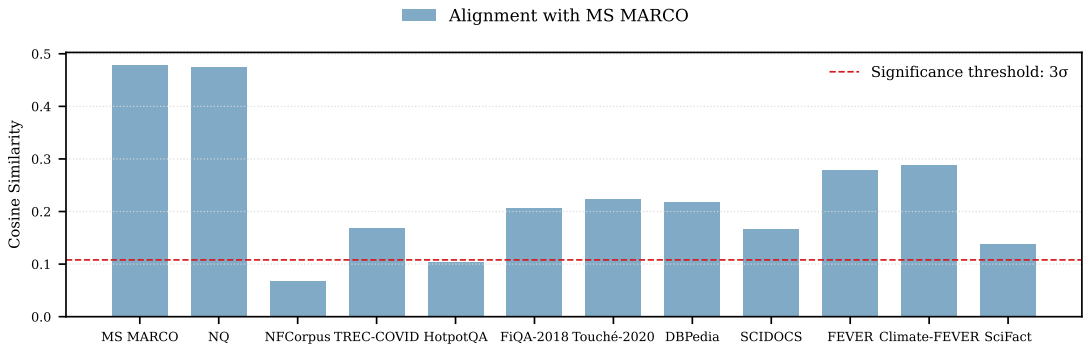


Figure 10: Cosine similarity between the LLM–Human displacement direction and the MS MARCO positive–negative contrast across datasets. The red dashed line marks the 3σ random-vector reference. Most datasets show strong alignment beyond the reference range, with a few cases near or below it.

ness for the training-time sampling experiments, which were omitted from the main text due to space constraints, and (b) additional inference-time debiasing results on more datasets. These results complement the main findings and further validate our conclusions.

I.1 Training-time Interventions: Retrieval Effectiveness

Table 12 reports retrieval effectiveness (NDCG@5) for the three negative sampling strategies (*in-batch only*, *standard*, and *hard-neg only*) across all datasets. Overall, settings that include mined hard negatives achieve higher retrieval performance, while using only in-batch negatives leads to lower effectiveness on most datasets. This trend is consistent with widely noted observations in the dense retrieval community that mined hard negatives are essential for strong retrieval effectiveness.

Table 12: NDCG@5 results on 14 datasets under different negative sampling strategies. The *standard* strategy combines in-batch and mined hard negatives, while the other two use only one type of negative.

Dataset	In-batch only	Standard	Hard-neg only
MS MARCO	0.629	0.629	0.623
DL19	0.640	0.706	0.728
DL20	0.642	0.701	0.719
TREC-COVID	0.571	0.611	0.568
NFCorpus	0.303	0.287	0.278
NQ	0.652	0.670	0.666
HotpotQA	0.570	0.579	0.579
FiQA-2018	0.209	0.218	0.216
Touché-2020	0.350	0.418	0.411
DBpedia	0.428	0.436	0.437
SCIDOCS	0.096	0.086	0.086
FEVER	0.850	0.842	0.829
Climate-FEVER	0.280	0.271	0.241
SciFact	0.435	0.452	0.443
Average	0.475	0.493	0.487

I.2 Inference-time Interventions

We provide the full inference-time results deferred to the appendix. Table 13 reports ΔNDSR@5 across all datasets, extending the five-dataset summary in the main text. Table 14 reports the corresponding NDCG@5 results. Overall, the projection method generally reduces source bias while largely preserving retrieval effectiveness across datasets.

J Revisiting Length Effects in Evaluating Source Bias

J.1 Length Bias in Neural Retrieval

Prior work has shown that neural retrievers exhibit non-semantic bias correlated with passage length. On BEIR's Touché-2020 argument retrieval task, Thakur et al. (2024) reports that neural retrievers tend to rank very short passages, many of which are non-argumentative, leading to large drops in effectiveness. Complementary evidence documents neural retrievers exhibit systematic brevity, early-position, and literal-match biases, and that these biases often result in shorter passages receiving disproportionately high scores, even outranking passages that contain the correct answer (Fayyaz et al., 2025).

This consideration is particularly relevant for evaluating *source bias*, since differences between human-written and LLM-generated passages may be partially reflected in their length. The Cocktail benchmark (Dai et al., 2024a), widely used to compare LLM-generated and human-written passages, constructs LLM rewrites that are typically shorter. This leads to the central question we examine: *is the observed source bias simply a manifestation of length bias?*

Table 13: ΔNDSR@5 before and after inference-time debiasing across 14 datasets and 5 relevance-supervised retrievers, using leave-one-dataset-out calibration to estimate the debiasing direction. Positive values indicate human preference, while negative values indicate LLM preference. In the Average row, the first line reports mean ΔNDSR@5, and the second line reports the remaining |ΔNDSR@5| relative to the original value. Darker shading indicates stronger source bias.

Dataset (↓)	ANCE		coCondenser		DRAGON		RetroMAE		TAS-B	
	Original	Debias	Original	Debias	Original	Debias	Original	Debias	Original	Debias
MS MARCO	-0.042	-0.004	-0.020	-0.002	-0.083	-0.035	-0.083	-0.053	-0.121	-0.064
DL19	-0.073	-0.029	-0.072	-0.041	-0.233	-0.161	-0.186	-0.131	-0.224	-0.176
DL20	-0.034	0.021	-0.079	-0.068	-0.121	-0.069	-0.088	-0.054	-0.072	-0.065
TREC-COVID	-0.162	-0.035	-0.340	-0.180	-0.134	-0.106	-0.194	-0.082	-0.328	-0.224
NFCorpus	-0.087	0.004	-0.068	-0.011	-0.079	-0.057	-0.081	-0.018	-0.082	-0.061
NQ	-0.042	0.003	-0.072	-0.030	-0.099	-0.050	-0.060	-0.029	-0.078	-0.055
HotpotQA	-0.020	0.016	-0.014	-0.005	-0.018	-0.016	-0.019	0.010	-0.018	0.003
FiQA-2018	-0.179	0.006	-0.219	-0.142	-0.161	-0.123	-0.205	-0.127	-0.170	-0.127
Webis-Touche2020	-0.168	0.046	-0.226	-0.130	-0.178	-0.114	-0.175	-0.047	-0.247	-0.099
DBpedia-Entity	-0.097	-0.047	-0.054	-0.032	-0.057	-0.052	-0.059	-0.022	-0.042	-0.025
SCIDOCS	-0.040	-0.017	-0.058	-0.017	-0.048	-0.035	-0.073	-0.002	-0.054	-0.021
FEVER	-0.200	-0.126	-0.037	-0.027	-0.043	-0.040	-0.010	0.013	-0.029	-0.020
Climate-FEVER	-0.314	-0.131	-0.153	-0.091	-0.091	-0.056	-0.105	-0.018	-0.083	-0.047
SciFact	-0.025	0.095	-0.049	0.010	-0.041	-0.021	-0.048	0.010	-0.058	-0.011
Average	-0.106	-0.014	-0.104	-0.055	-0.099	-0.067	-0.099	-0.039	-0.115	-0.071
	(100%)	(13%)	(100%)	(53%)	(100%)	(68%)	(100%)	(39%)	(100%)	(62%)

J.2 Extended Rewriting for Length Control

To answer this question, we require an evaluation setting in which passage *source* can be varied while passage *length* is manipulated independently of meaning. However, because the LLM-generated versions in Cocktail are generally shorter than the original human-written passages, they do not allow us to determine whether any observed preference for LLM-generated text reflects its source or simply its reduced length.

To control for length effects, we construct an additional rewritten version of each dataset that preserves the semantic content of the original passage while producing a systematically longer variant. Following the semantic-preserving rewriting protocol of previous work (Dai et al., 2024a), we use an LLM to regenerate each passage based solely on its original content, using a prompt that enforces meaning preservation while generating a slightly longer paraphrase. The full prompt is provided in Appendix J.5.

This yields three aligned versions for every dataset: the human-written version, the Cocktail version, and our length-extended version. These parallel versions enable controlled comparison in settings where rewritten passages are either shorter than or longer than their human counterparts. We report the resulting length statistics for all datasets in the next subsection.

J.3 Length Statistics

Table 15 reports the average passage lengths across the three versions. Consistent with prior observations, the Cocktail version reduces passage length by approximately 5–20% across datasets. By contrast, our extended version increases length by 3–45%. Taken together, these variants provide controlled conditions in which LLM-generated passages can be either shorter or longer than their human-written counterparts.

The length statistics in our analysis are computed using the standard Apache Lucene tokenizer without stemming or stopword removal. As tokenization procedures differ across implementations, our computed averages show small deviations from the values reported in the Cocktail benchmark (Dai et al., 2024a). For completeness, we include those reported values in Table 7.

J.4 Does Length Alone Account for Retriever Preferences?

As shown in Section 3, source bias is not a universal property of neural retrievers: it appears consistently only in models trained with relevance supervision, whereas unsupervised and general-purpose embedding models do not exhibit systematic source preference. A natural follow-up question is whether these supervised models seem to favor LLM-generated passages simply because the LLM passages in Cocktail are, on average, shorter than their human-written counterparts.

To assess whether length differences alone can

Table 14: NDCG@5 results (original vs. debias) on 14 datasets for 5 relevance-supervised retrievers.

Dataset (↓)	ANCE		coCondenser		DRAGON		RetroMAE		TAS-B	
	Original	Debias	Original	Debias	Original	Debias	Original	Debias	Original	Debias
MS MARCO	0.590	0.588	0.620	0.620	0.665	0.666	0.626	0.624	0.617	0.618
DL19	0.695	0.696	0.750	0.755	0.767	0.763	0.739	0.739	0.743	0.741
DL20	0.716	0.703	0.750	0.754	0.778	0.776	0.760	0.761	0.737	0.740
TREC-COVID	0.679	0.690	0.707	0.712	0.684	0.675	0.744	0.737	0.644	0.649
NFCorpus	0.301	0.300	0.382	0.386	0.397	0.398	0.373	0.376	0.375	0.376
NQ	0.628	0.623	0.687	0.688	0.737	0.736	0.704	0.705	0.689	0.689
HotpotQA	0.537	0.528	0.640	0.640	0.719	0.719	0.716	0.715	0.674	0.673
FiQA-2018	0.255	0.257	0.244	0.248	0.323	0.323	0.278	0.281	0.257	0.262
Touché-2020	0.487	0.496	0.326	0.344	0.501	0.505	0.444	0.457	0.429	0.427
DBpedia	0.435	0.437	0.525	0.523	0.540	0.539	0.526	0.528	0.518	0.516
SCIDOCS	0.114	0.111	0.124	0.125	0.148	0.148	0.136	0.135	0.133	0.133
FEVER	0.824	0.826	0.785	0.786	0.895	0.895	0.891	0.892	0.858	0.858
Climate-FEVER	0.240	0.244	0.237	0.240	0.290	0.291	0.279	0.282	0.286	0.287
SciFact	0.429	0.424	0.530	0.531	0.599	0.602	0.571	0.575	0.563	0.561
Average	0.495	0.494	0.522	0.525	0.575	0.574	0.556	0.558	0.537	0.537

account for the observed preferences, we repeat the source-preference evaluation on the length-controlled versions introduced above. In this setting, each LLM passage is a meaning-preserving yet systematically lengthened version of its human-written counterpart. This setup removes the original brevity advantage of LLM passages in Cocktail and, in most cases, even places them at a length disadvantage. Under these conditions, any remaining preference for LLM-generated text cannot be attributed to shorter passage length.

Table 16 shows the results for the full set of 13 retrievers. Among the relevance-supervised models, lengthening the LLM passages reduces the magnitude of the bias, often substantially, but does not eliminate it. These models continue to prefer LLM-generated passages on most datasets, despite the fact that the LLM versions are now longer than the human counterparts. A few model–dataset pairs even flip to a slight preference for human-written passages, but the dominant pattern remains a persistent bias toward LLM passages. This indicates that passage length modulates the strength of the effect but is insufficient to account for its origin.

Table 17 further examines the three unsupervised retrievers from Section 3 after MS MARCO fine-tuning. Models that originally exhibited little or no source preference develop a consistent bias toward LLM-generated passages once trained with relevance supervision—even when the LLM passages face the same length disadvantage. In contrast, general-purpose embedding models and the unsupervised retrievers in their base form tend to favor the human passages under the length-controlled condition, reflecting their greater sensitivity to length but also their lack of a stable preference in either direction.

Taken together, these findings suggest that although passage length influences the strength of source preference, it is insufficient to explain why the effect persists. Lengthening the LLM passages attenuates the bias but does not remove it, and relevance supervision still induces a preference for LLM-generated text when length advantages are reversed. This aligns with the overall conclusion that source bias primarily arises from stylistic and distributional artifacts in supervised retrieval data.

J.5 Prompt for Length-Controlled Rewriting

To produce the length-controlled passages used in our evaluation, we employ a prompt that enforces meaning preservation while generating slightly longer variants of the original text. The full prompt template is shown in Figure 11. While the prompt specifies a modest increase (e.g., around 10%), LLM outputs exhibit natural variability, and some rewritten passages become substantially longer. This variability does not affect our analysis, as the experiment requires only that the LLM passages be consistently longer than the originals, not that they match a specific percentage.

K Additional Analyses of Supervision and Negative Sampling

This section provides further evidence on how different *supervision pipelines* and *negative sampling*

Table 15: Average passage lengths across three versions of each dataset, along with relative changes (Δ) with respect to the original human-written passages. For datasets sharing the MS MARCO corpus (e.g., DL19 and DL20), length statistics are reported once.

Dataset	**Human (Orig.)**	**Cocktail**	**Ours**	**Δ Cocktail (%)**	**Δ Ours (%)**
MS MARCO	59.1	55.8	82.0	−5.6	+38.7
TREC-COVID	204.8	171.8	222.0	−16.1	+8.4
NFCorpus	227.6	184.8	249.3	−18.8	+9.5
NQ	89.6	82.2	110.8	−8.2	+23.6
HotpotQA	68.5	67.4	92.9	−1.6	+35.6
FiQA-2018	135.4	109.3	147.6	−19.3	+9.0
Touché-2020	167.2	135.5	172.7	−18.9	+3.3
DBpedia	54.0	54.7	78.5	+1.3	+45.3
SCIDOCS	174.0	166.0	182.7	−4.6	+5.0
FEVER	99.4	91.9	118.7	−7.5	+19.4
Climate-FEVER	87.2	82.0	107.9	−6.0	+23.8
SciFact	209.4	178.9	230.2	−14.5	+9.9

Table 16: ΔNDSR@5 results across 14 datasets for 13 neural retrievers spanning three model families. Negative values are shaded in red to indicate a preference for LLM-generated passages, while positive values are shaded in blue to indicate a preference for human-written passages. Asterisks (*) denote statistically significant deviations from zero (two-sided t-test, $p < 0.05$).

Dataset (↓)	Relevance-Supervised Retrievers					General-Purpose Embedding Models					Unsupervised Retrievers		
	ANCE	TAS-B	coCondenser	RetroMAE	DRAGON	BGE	BCE	GTE	E5	M3E	Contriever	E5-Unsup	SimCSE
MS MARCO	-0.026*	-0.042*	0.032*	-0.014*	-0.007	0.040*	0.383*	-0.038*	0.157*	0.446*	0.271*	-0.113*	0.658*
DL19	-0.092	-0.039	-0.065*	-0.056	-0.016	0.079	0.561*	-0.072	0.256*	0.554*	0.320*	-0.079	0.741*
DL20	-0.007	-0.033	-0.029	-0.020	-0.039	0.212*	0.458*	-0.004	0.227*	0.648*	0.367*	0.017	0.793*
NQ	-0.050*	-0.045*	-0.035*	-0.002	-0.047*	-0.085*	0.498*	0.024*	0.313*	0.320*	0.197*	0.152*	0.369*
NFCorpus	-0.071*	-0.031*	-0.025*	-0.055*	-0.018	-0.031*	0.090*	-0.015	0.356*	0.049*	-0.004	-0.165*	0.423*
TREC-COVID	-0.028	-0.115	-0.056	-0.233*	-0.079	-0.061	0.358*	-0.012	0.151*	0.390*	-0.113*	0.039	0.569*
HotpotQA	0.024*	0.067*	0.024*	0.034*	-0.002	0.151*	0.354*	0.051*	0.238*	0.237*	-0.484*	-0.044*	0.110*
FiQA-2018	-0.195*	-0.095*	-0.065*	-0.154*	-0.103*	-0.143*	0.607*	-0.016	-0.085*	0.414*	0.085*	-0.063*	0.513*
Touche-2020	0.016	-0.010	-0.050	-0.002	-0.022	-0.055	0.667*	0.022	-0.168*	0.636*	-0.101	0.003	0.268*
DBpedia	-0.033	-0.017	-0.018	-0.012	-0.023	0.112*	0.369*	0.046*	0.242*	0.444*	-0.314*	0.004	0.371*
SCIDOCS	-0.023*	-0.075*	-0.074*	-0.007	-0.024*	-0.148*	0.572*	-0.019*	0.188*	0.453*	0.159*	0.032*	0.386*
FEVER	-0.155*	-0.034*	-0.021*	-0.016*	-0.021*	0.037*	0.442*	0.061*	0.159*	0.247*	-0.075*	0.066*	0.105*
Climate-FEVER	-0.245*	-0.103*	-0.064*	-0.147*	-0.132*	-0.127*	0.742*	-0.022*	0.487*	0.253*	-0.027*	0.147*	0.068*
SciFact	-0.084*	-0.027	-0.041*	-0.010	-0.011	-0.016	-0.006	-0.021	0.133*	-0.059*	0.173*	-0.071*	0.181*

strategies influence source bias. Our goal is not to analyze negative sampling in isolation, but to understand how supervision design induces (or avoids) *artifact imbalances* between positives and negatives, a central component of the mechanism developed in Section 4.

This section begins by clarifying how different supervised retrievers construct their hard negatives, since these pipelines determine the stylistic differences between positive and negative passages. We then evaluate a retriever trained solely on NQ to examine whether the observed source bias generalizes beyond MS MARCO. Finally, we evaluate a more practical alternative to the in-batch-only setup by selecting hard negatives directly from the positive pool.

K.1 Supervision Pipelines and Hard-Negative Construction

Relevance-supervised retrievers (e.g., ANCE, TAS-B, DRAGON, RetroMAE, coCondenser) typically rely on large-scale human relevance annotations, coupled with mined hard negatives. Although their exact pipelines differ, they all follow a similar supervised contrastive setup: for each query q, a judged relevant passage p^+ is paired with one or more hard negatives p^- drawn from top-ranked retrieval candidates that are either unjudged or explicitly labeled non-relevant.

Concretely, ANCE, coCondenser and DRAGON iteratively refresh hard negatives using the current dense retriever, while TAS-B and RetroMAE mainly rely on BM25- or multi-retriever–mined

```
<|system|>
You are a helpful assistant.

<|user|>
Please follow the instructions below:
1. Maintain the original meaning of the input passage.
2. Make the paraphrased passage slightly longer than the original (e.g., 10% longer) while preserving the
     same information.
3. Output the paraphrased passage directly.

Following is the passage you need to paraphrase:
{text}
Your answer must be formatted as:
Rewritten Text:
<your rewritten text>
```

Figure 11: Prompt used to generate meaning-preserving, length-extended passages with Qwen2.5-7B-Instruct.

Table 17: ΔNDSR@5 results of unsupervised retrievers after MS MARCO fine-tuning, corresponding to the same base models in Table 16. The "-FT" suffix denotes fine-tuning on MS MARCO. Negative values are shaded in red to indicate a preference for LLM-generated passages, while positive values are shaded in blue to indicate a preference for human-written passages. Asterisks (*) denote statistically significant deviations from zero (two-sided t-test, $p < 0.05$).

Dataset (↓)	Relevance-Supervised Retrievers		
	Contriever-FT	E5-FT	SimCSE-FT
MS MARCO	0.110*	-0.008*	-0.020*
DL19	-0.015	-0.046	-0.082
DL20	0.218*	0.071	-0.021
NQ	-0.017*	-0.049*	-0.048*
NFCorpus	-0.007	-0.083*	-0.019
TREC-COVID	-0.238*	-0.242*	-0.036
HotpotQA	0.111*	-0.012*	0.035*
FiQA-2018	-0.016	-0.123*	-0.206*
Touche-2020	-0.041	-0.042	-0.089
DBpedia	0.060	-0.017	-0.044*
SCIDOCS	-0.026*	-0.022*	-0.056*
FEVER	0.078*	-0.041*	0.029*
Climate-FEVER	-0.091*	-0.145*	-0.120*
SciFact	-0.021	-0.058*	-0.052

candidates from MS MARCO. In all cases, positives and negatives are sampled from *different parts of the corpus*: positives concentrate on answer-like spans around annotated answers, whereas negatives come from a much broader and noisier pool.

A common pattern therefore emerges:

- **Positives** are drawn from *answer-like, high-quality passages*, often with higher fluency and information density.
- **Negatives** are drawn from *retrieval candidates* (e.g., BM25 top-k or model-mined candidates) that are more heterogeneous, noisier, and stylistically less polished.

As shown in Figure 1a and Appendix F, this supervision scheme induces a systematic stylistic gap between positive and negative pools. If we let A denote a stylistic attribute (e.g., fluency), and write

$$\Delta A = \mathbb{E}[A^+] - \mathbb{E}[A^-],$$

then in typical relevance-supervised training we observe $\Delta A > 0$ across multiple artifact dimensions. This helps explain why *all relevance-supervised retrievers* in Table 1 exhibit pronounced source bias: supervision itself encodes a non-semantic separation between positives and negatives that overlaps with properties of LLM-generated text.

In contrast, the other model families in Table 1 use supervision schemes that do *not* introduce such a consistent gap:

- **General-purpose embedding models** (e.g., E5, BGE, GTE, M3E) are trained with multi-task and weakly supervised objectives, often with symmetric sampling from the same underlying corpus. Positives and negatives are drawn from similar distributions (e.g., STS/NLI pairs, contrastive sentence pairs), which keeps ΔA small.
- **Unsupervised retrievers** (e.g., Contriever, unsupervised SimCSE) sample positives via augmentation (e.g., dropout views, adjacent sentences) and use in-batch samples as negatives. Both come from the same raw corpus, again implying $\Delta A \approx 0$.

These differences can therefore be unified under a single mechanism: *whether the supervised training process creates a systematic stylistic imbalance between positives and negatives*.

K.2 Source Bias Persists Under NQ-Only Supervision

A natural question is whether source bias is specific to MS MARCO and its particular negative-construction scheme. To test this, we evaluate a publicly available DPR retriever trained solely on Natural Questions (NQ)[1]. NQ differs from

[1]We use the publicly released DPR encoders trained on NQ: the question encoder (https://huggingface.

Table 18: Source bias of the NQ-only DPR retriever on Cocktail. We report ΔNDSR@5; negative values indicate a preference for LLM-generated passages and are shaded in red. Asterisks (*) denote statistically significant deviations from zero (two-sided t-test, $p < 0.05$).

Dataset	ΔNDSR@5	Dataset	ΔNDSR@5
MS MARCO	-0.264*	FiQA-2018	-0.631*
DL19	-0.258*	Touche-2020	-0.430*
DL20	-0.319*	DBpedia	-0.087*
NQ	-0.106*	SCIDOCS	-0.353*
NFCorpus	-0.170*	FEVER	-0.075*
TREC-COVID	-0.503*	Climate-FEVER	-0.119*
HotpotQA	-0.065*	SciFact	-0.300*

Table 19: Comparison of negative sampling strategies. We report average ΔNDSR@5 (bias; lower magnitudes indicate weaker source bias) and NDCG@5 (retrieval effectiveness) across the 14 Cocktail datasets. The "Positive-pool hard negs" condition retains hard negatives while reducing the stylistic gap between positive and negative pools.

Training configuration	ΔNDSR@5	NDCG@5
Standard	-0.099	0.493
In-batch only	-0.024	0.475
Positive-pool hard negs	-0.023	0.478

MS MARCO in several fundamental respects, including its Wikipedia-only domain, its distinct annotation and answer format, the retrieval candidates used for negative sampling, and the overall style and distribution of passages.

We evaluate this NQ-only retriever on Cocktail. Despite the differences in corpus, annotation, and negative sampling pipeline, it still exhibits substantial source bias. On average across the 14 datasets we observe $\Delta\text{NDSR@5} = -0.263$, and the per-dataset results are shown in Table 18. All datasets show a clear preference for LLM-generated passages.

Although the magnitude varies across datasets, the direction of the effect mirrors that of MS MARCO-trained retrievers. This indicates that source bias is not an artifact of MS MARCO, but a more general outcome of relevance supervision. Whenever a supervised dataset introduces a stylistic discrepancy between positive and negative passages (i.e., $\Delta A > 0$), the resulting retriever tends to inherit a corresponding preference that favors LLM-generated text.

co/sentence-transformers/facebook-dpr-question_encoder-single-nq-base) and the context encoder (https://huggingface.co/sentence-transformers/facebook-dpr-ctx_encoder-single-nq-base).

Table 20: Per-dataset ΔNDSR@5 and NDCG@5 results for the "positive-pool hard negs" configuration.

Dataset	ΔNDSR@5	NDCG@5
MS MARCO	0.027	0.621
DL19	-0.013	0.667
DL20	0.012	0.661
NQ	0.019	0.668
NFCorpus	-0.060	0.239
TREC-COVID	-0.158	0.574
HotpotQA	0.013	0.577
FiQA-2018	0.005	0.215
Touche-2020	-0.064	0.396
DBpedia	-0.033	0.430
SCIDOCS	-0.017	0.091
FEVER	0.037	0.844
Climate-FEVER	-0.069	0.268
SciFact	-0.023	0.436
Average	-0.023	0.478

K.3 From In-Batch Negatives to Hard Negatives from the Positive Pool

Section 5 showed that training-time negative sampling has a direct and monotonic effect on source bias: moving from in-batch only to standard sampling to hard-neg only progressively increases artifact imbalance ΔA and strengthens bias (Table 3). The in-batch-only configuration serves as a useful *mechanism probe* because it removes all mined hard negatives and thus approximates the idealized condition $\Delta A \approx 0$. However, this setup is stricter than what typical retrieval systems would use in practice. To explore a more practical alternative, we examine a variant that *retains hard negatives* while reducing the stylistic gap between positive and negative pools.

To bridge this gap, we evaluate a more realistic variant that keeps hard negatives while reducing the stylistic gap between positives and negatives. For each annotated positive passage, we run BM25 over the positive pool and select the highest-ranked non-ground-truth passages as hard negatives. These negatives are still semantically challenging but remain stylistically similar to positives, since both are drawn from the same answer-like distribution. Intuitively, this construction preserves hard negative difficulty while shrinking ΔA.

Table 19 compares three settings for a representative relevance-supervised retriever: (1) the "Standard" configuration (in-batch + mined hard negatives), (2) the "In-batch only" setting from Section 5, and (3) the "Positive-pool hard negs" configuration (in-batch + BM25 negatives over positives). We report average ΔNDSR@5 and NDCG@5 across all 14 datasets, and the full per-dataset metrics for the positive-pool variant are

provided in Table 20.

Empirically, "In-batch" only yields the strongest reduction in source bias but also a drop in NDCG@5, consistent with prior observations that mined hard negatives help retrieval effectiveness. The "Positive-pool hard negs" setting achieves a similar suppression of bias as the in-batch-only condition, yet recovers much of the NDCG@5 lost in that extreme setup.

These results sharpen the role of negative sampling. First, *hard negatives themselves are not the cause of source bias*: models trained with positive-pool hard negatives still see difficult negatives but remain weakly biased. Instead, bias tracks the stylistic alignment between positive and negative pools: when both are drawn from similar stylistic distributions (small ΔA), source bias is substantially reduced, even in the presence of hard negatives. Together with the NQ-only supervision results in Appendix K.2, this provides additional evidence that artifact imbalance in supervision, rather than the use of hard negatives by itself, is the primary driver of source bias.