Masking or Mitigating? Deconstructing the Impact of Query Rewriting on Retriever Biases in RAG

Anonymous ACL submission

Abstract

Dense retrievers in retrieval-augmented generation (RAG) systems exhibit systematic biases—including brevity, position, literal matching, and repetition biases—that can compromise retrieval quality. Query rewriting techniques are now standard in RAG pipelines, yet their impact on these biases remains unexplored. We present the first systematic study of how query enhancement techniques affect dense retrieval biases, evaluating five methods across six retrievers. Our findings reveal that simple LLM-based rewriting achieves the strongest aggregate bias reduction (54%), yet fails under adversarial conditions where multiple biases combine. Mechanistic analysis uncovers two distinct mechanisms: simple rewriting reduces bias through increased score variance, while pseudo-document generation methods achieve reduction through genuine decorrelation from bias-inducing features. However, no technique uniformly addresses all biases, and effects vary substantially across retrievers. Our results provide practical guidance for selecting query enhancement strategies based on specific bias vulnerabilities. More broadly, we establish a taxonomy distinguishing query-document interaction biases from document encoding biases, clarifying the limits of query-side interventions for debiasing RAG systems.

1 Introduction

Retrieval-Augmented Generation (RAG) systems have become foundational for grounding large language models in external knowledge, enabling applications from question-answering to document analysis across diverse domains (Lewis et al., 2020; Lu et al., 2022; Jiang et al., 2023; Gao et al., 2023b; Asai et al., 2024). These systems typically employ dense retrievers (Karpukhin et al., 2020; Thakur et al., 2021) to identify relevant passages from document collections, which are then provided as context for response generation (Oche et al., 2025).

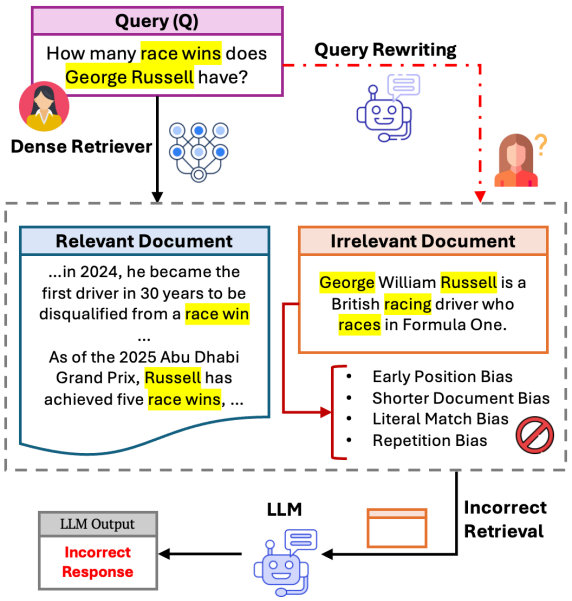

Figure 1: **Paper Overview.** Dense retrievers are susceptible to various biases, leading to incorrect retrieval and erroneous LLM outputs. Can query rewriting help? We systematically evaluate its effects on retrieval bias.

Despite their effectiveness, recent work has revealed that dense retrievers exhibit systematic biases that can compromise retrieval quality (Ram et al., 2023; Coelho et al., 2024). Fayyaz et al. (2025) demonstrated that retrievers consistently favor documents with surface-level characteristics—brevity, early answer positioning, lexical overlap, and entity repetition—over documents containing actual factual evidence. The ideal solution would be dense retrieval models inherently robust to such biases and capable of prioritizing semantic relevance. However, achieving this through architectural modifications or training objective changes remains challenging. A more tractable first step is understanding whether these biases persist under query enhancement techniques (Ma et al., 2023), which are standard in modern RAG pipelines.

Production RAG systems routinely employ query transformations including LLM-based rewrit-

ing for better semantic matching and pseudo-document generation approaches like HyDE (Gao et al., 2023a) and Query2Doc (Wang et al., 2023). While these techniques improve retrieval effectiveness, their interaction with systematic biases remains unexplored: do they amplify existing biases by generating surface-level features that retrievers favor, or do specific strategies mitigate particular bias types? Understanding this interaction is critical for deploying robust and fair RAG systems, particularly in high-stakes domains such as medical, sociotechnical, and enterprise settings (Ryan et al., 2025; Wang et al., 2025; Goyal et al., 2025; Shimgekar et al., 2025).

In this work, **we conduct the first systematic investigation into how query rewriting techniques affect retrieval biases**. We evaluate five enhancement methods across six retrievers using the controlled framework of Fayyaz et al. (2025). This framework's treatment-control methodology, which uses paired documents that differ only in specific surface characteristics while maintaining factual equivalence, enables us to isolate and measure the effect of query rewriting on individual bias types rather than confounding bias reduction with overall retrieval improvements. We find that query enhancement effects are highly differential: simple LLM rewriting achieves the strongest aggregate bias reduction (54%) but fails under adversarial conditions, while pseudo-document methods provide more modest but mechanistically robust improvements. We establish a taxonomy distinguishing *interaction biases* amenable to query-side interventions from *encoding biases* requiring retriever-level modifications, providing practical guidance for bias-aware RAG deployment.

2 Related Works

Bias in Dense Retrieval. Dense retrieval has emerged as a powerful alternative to traditional sparse methods like BM25 for information retrieval tasks (Karpukhin et al., 2020; Thakur et al., 2021), showing strong performance on retrieval benchmarks. However, despite these advances, recent work has begun to uncover systematic biases in how dense retrievers encode and rank documents, including position biases (Coelho et al., 2024) and lexical biases (Ram et al., 2023). Similarly, BehnamGhader et al. (2023) show that Dense Passage Retrieval (DPR) models fail to retrieve statements requiring reasoning beyond surface-level similarity. Most relevant to our work, Fayyaz et al. (2025) introduced ColDER, a controlled benchmark for systematically measuring various types of retrieval biases that biases persist across multiple dense retriever architectures and can compound adversarially. *Our work extends this line of research by investigating whether and how query-side interventions can mitigate these documented biases.*

Query Enhancement in RAG. RAG systems (Lewis et al., 2020; Gao et al., 2023b) have become foundational for knowledge-intensive NLP tasks, with query rewriting now a standard pipeline component (Ma et al., 2023). Several query enhancement techniques have been proposed to bridge the semantic gap between queries and documents. HyDE (Gao et al., 2023a) generates pseudo-documents and retrieves based on their embeddings rather than the original query; Query2Doc (Wang et al., 2023) concatenates LLM-generated pseudo-documents to expand queries; and various other simple LLM-based query rewriting techniques have been proposed in prior work (Jagerman et al., 2023; Mao et al., 2024; Kim et al., 2025). While these methods improve retrieval effectiveness, their impact on systematic retrieval biases remains unexplored. As RAG systems are increasingly deployed in high-stakes domains, understanding how retrieval biases propagate through these pipelines becomes critical. *Our work provides the first systematic analysis of how query enhancement techniques interact with known retrieval biases.*

3 Preliminaries, Data, and Methods

3.1 Problem Formulation

We investigate whether and how query rewriting techniques affect systematic biases in dense retrievers. We adopt the treatment-control ColDeR benchmark contributed by Fayyaz et al. (2025), which uses carefully constructed document pairs using the Re-DocRED dataset (Tan et al., 2022) that differ only in specific surface characteristics while maintaining factual equivalence.

Given a query q and a pair of documents (D_1, D_2) where D_1 contains a bias-inducing characteristic and D_2 serves as the control, we measure retrieval preferences using the similarity scores assigned by a dense retriever $\mathcal{S}$. This controlled design enables us to isolate the effect of query rewriting on individual bias types, separating bias mitigation from general retrieval performance improvements. Then, for each document pair, we compute

the difference in retrieval scores:

$$\delta = \mathcal{S}(q, D_1) - \mathcal{S}(q, D_2)$$

where $\mathcal{S}(q, D)$ represents the retriever's similarity score between the generated embeddings for the query q and document D. Positive δ values indicate bias toward surface-level characteristics, while values near zero suggest factual grounding. We apply different query enhancement techniques to transform the original query q into q', and measure how this transformation affects the distribution of δ across query-document pairs. For each bias type, we collect $N = 200$ paired documents $\{(D_1, D_2)\}_i^N$ along with associated queries. We now describe each bias type in detail.

3.2 Bias Types

We evaluate four systematic biases identified by Fayyaz et al. (2025) in dense retrievers:

(i) **Brevity Bias:** Retrievers prefer shorter documents over longer ones containing equivalent evidence. Document D_1 contains the same factual information as D_2 but with significantly fewer tokens.

(ii) **Literal Bias:** Retrievers prioritize exact lexical matches over semantic equivalence. Document D_1 uses identical or highly overlapping terminology with the query, while D_2 expresses the same information using paraphrases or synonyms.

(iii) **Position Bias:** Retrievers favor content appearing early in documents over information at later positions. In document pairs, D_1 places the relevant information at the beginning while D_2 positions it toward the end.

(iv) **Repetition Bias:** Retrievers prefer documents with repeated entity mentions. Document D_1 contains multiple mentions of key entities, while D_2 mentions them fewer times despite containing equivalent factual content.

We refer the reader to Fayyaz et al. (2025) for rigorous definitions of these bias types.

3.3 Dense Retrievers

We evaluate six state-of-the-art dense retrievers representing diverse architectural approaches:
(1) Contrastive Bi-Encoders: Contriever-MSMARCO (Izacard et al., 2021), COCO-DR Base MSMARCO (Yu et al., 2022), Dragon RoBERTa (Lin et al., 2023), DRAGON+ (Lin et al., 2023)
(2) Generative Pretraining: RetroMAE MS-MARCO FT (Xiao et al., 2022)
(3) Token-Level Interaction: ColBERTv2 (Santhanam et al., 2022)

This diverse set represents the major paradigms in modern dense retrieval—contrastive learning, diverse data augmentation, and late interaction architectures—enabling us to assess whether query rewriting effects on biases are architecture-specific or systemic across different retrieval approaches.

3.4 Query Enhancement Techniques

We evaluate four query enhancement approaches that represent the spectrum of techniques used in modern RAG systems.
(1) LLM-based Query Rewriting: We use an LLM to reformulate the original query for better semantic matching. Given query q, we prompt the model to produce the rewritten query q_{rewrite}.
(2) HyDE: Following Gao et al. (2023a), we leverage the Hypothetical Document Embeddings (HyDE) approach, which uses an LLM to generate pseudo-documents that might hypothetically contain the answer to a query rather than using the query directly for retrieval. For a given query q, we prompt a language model $\mathcal{M}$ to generate a pseudo-document $\hat{d} = \mathcal{M}(\text{prompt}(q))$ where $\text{prompt}(q)$ instructs the model to generate a passage that would contain the answer to query q. The generated hypothetical document $\hat{d}$ is then embedded and used as the query representation for retrieval. The key insight of HyDE is that hypothetical documents often provide richer semantic context than short queries, potentially capturing vocabulary, style, and relational patterns that better align with actual target documents.
(3) Query2Doc (Q2D): We adopt the Query2Doc approach introduced by Wang et al. (2023), which augments queries with LLM-generated pseudo-documents rather than replacing them entirely. For a given query q, we prompt a language model $\mathcal{M}$ to generate a pseudo-document $\hat{d}$ that might be relevant to the query. This generated passage is then concatenated with the original query to form an expanded query representation $\hat{q}_{\text{Q2D}} = q \oplus \hat{d}$ where $\oplus$ denotes concatenation. The key distinction from HyDE is that Query2Doc preserves the original query signal while enriching it with generated context. Since our retrieval scenario is inherently zero-shot, we do not utilize few-shot ex-

emplars for generation of pseudo-documents like the original Query2Doc paradigm. This augmented query q_{Q2D} is then embedded and used for retrieval.

In line with recent efforts toward domain-specific LLMs in RAG systems (Zhang et al., 2024), we test whether domain adaptation can improve query rewriting for bias mitigation. Specifically, we continually pretrain $\mathcal{M}$ on the original Re-DocRED document for a given document pair from ColDER, before applying HyDE or Query2Doc. Specifically, we use Low-Rank Adaptation (LoRA) (Hu et al., 2022) to adapt the model to document d by minimizing:

$$\mathcal{L}_{\text{CPT}} = -\sum_{i=1}^{|d|} \log P_\theta(w_i \mid w_{<i}, d)$$

where θ represents the LoRA parameters and w_i denotes the i^{th} token in document d. The hypothesis is that this adapted model $\mathcal{M}_{\text{domain}}$ familiar with domain-specific vocabulary and factual patterns may generate pseudo-documents that better align with actual evidence, potentially reducing reliance on surface-level features. We use LoRA with rank $r = 8, \alpha = 16$, and train for 10 epochs per document. This gives us our final two techniques:
(4) HyDE with Continual Pretraining (HyDE-CPT): We use the domain-adapted model $\mathcal{M}_{\text{domain}}$ to use to generate psuedo-document $\hat{d}_{\text{CPT}} = \mathcal{M}_{\text{domain}}(\text{prompt}(q))$ following the HyDE prompt, and then use the embedding of $\hat{d}_{\text{CPT}}$ for retrieval.
(5) Query2Doc with Continual Pretraining (Q2D-CPT): We concatenate the original query to the pseudo-document generated by the domain-adapted model $\mathcal{M}_{\text{domain}}$ to obtain $\hat{q}_{\text{Q2D-CPT}} = q \oplus \hat{d}_{\text{CPT}}$ and use its embedding for retrieval.

To ensure that the effects we observe are robust across the LLM used for query enhancement, we test open-source models from two family of models: Gemma 3 (Team et al., 2025) (`google/gemma-3-12b-it`) and Qwen3 (Yang et al., 2025) (`Qwen/Qwen3-4B-Instruct-2507`).

3.5 Evaluation Metrics

For each bias type and query enhancement technique, we measure retrieval preferences using paired t-tests (Ross and Willson, 2017) on score differences $\Delta = [\delta_1, \ldots, \delta_N]$ across all query-document pairs. We report *mean* $|t|$*-statistic*, the absolute value of the t-statistic measuring the strength of bias, as well as the percentage decrease in mean $|t|$-statistic compared to the vanilla query. Furthermore, we compute $p-$values to check for statistical significance of the bias at $\alpha = 0.05$ with Bonferroni correction (Weisstein, 2004). Lower $|t|$-statistic values indicate reduced bias, with values approaching zero suggesting factual grounding rather than reliance on surface characteristics.

Method	Mean $\lvert t\rvert$	Sig. Biases	Reduction
Baseline	8.72 ± 5.32	21/24	–
Rewrite	4.02 ± 2.17	13/24	+53.9%
HyDE	6.95 ± 4.73	20/24	+20.3%
HyDE-CPT	6.78 ± 4.79	19/24	+22.3%
Q2D	6.15 ± 5.96	16/24	+29.5%
Q2D-CPT	6.07 ± 5.72	17/24	+30.4%

Table 1: Summary of retrieval bias across query enhancement methods. Mean $|t|$-statistic averaged across all retrievers and bias types. Sig. Biases: number of retriever-bias combinations showing significant bias ($p < 0.05$ after Bonferroni correction). Reduction: percentage decrease in $|t|$ compared to Baseline.

4 Results

We now present results for query enhancements using Gemma-3-12B-IT, and defer the results for Qwen3-4B-Instruct to Appendix B since we observe very similar trends overall.

4.1 Overall Trend

Table 1 summarizes retrieval bias across all query enhancement methods, aggregated over six retrievers and four bias types (24 total retriever-bias combinations). All query enhancement techniques reduce bias relative to the baseline, though the magnitude varies substantially. Most strikingly, simple LLM-based query rewriting achieves the largest reduction (53.9%), cutting the mean $|t|$-statistic from 8.72 to 4.02 and reducing the number of statistically significant biases from 21 to 13 out of 24 combinations. The pseudo-document generation methods—HyDE and Query2Doc—provide more modest improvements, with reductions ranging from 20.3% to 30.4%. Query2Doc variants outperform HyDE variants, suggesting that preserving the original query signal while augmenting with generated content is more effective than replacing the query entirely. Continual pretraining (CPT) offers marginal gains over the base methods (HyDE-CPT: +2.0% over HyDE; Q2D-CPT: +0.9% over Q2D), indicating that domain adaptation primarily improves factual grounding rather than bias mitigation. However, these aggregate statistics mask

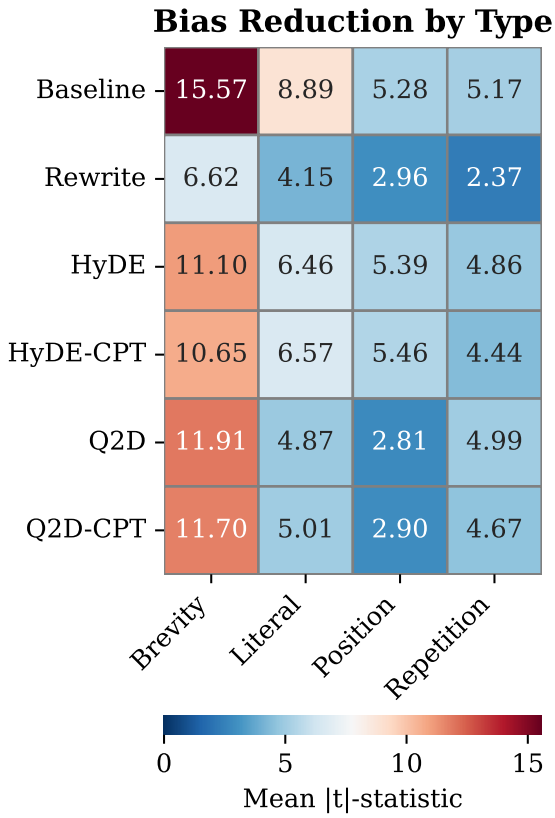

Figure 2: **Mean $|t|$-statistic by bias type and method (averaged across 6 retrievers).** Lower values indicate reduced bias. All $|t|$ values were averaged irrespective of statistical significance.

important variation across bias types that we examine in detail below.

4.2 Reduction in Retrieval Bias

Figure 2 presents the mean $|t|$-statistic across bias types and query enhancement methods, averaged over all retrievers. We observe several patterns:

Simple rewriting is surprisingly effective. Contrary to our intuition, simple LLM-based query rewriting achieves the strongest overall bias reduction, decreasing the average $|t|$-statistic from 8.72 to 4.02. This simple baseline outperforms more sophisticated pseudo-document generation methods across all four bias types. The result suggests that much of the bias in dense retrieval may stem from the surface-level characteristics of user queries themselves—their brevity, specific lexical choices, or syntactic patterns—which even basic reformulation can address.

Pseudo-document methods show bias-specific trade-offs. HyDE and Query2Doc show notably different bias reduction profiles. HyDE reduces brevity bias substantially (15.57 → 11.10) but slightly *exacerbates* position bias (5.28 → 5.39). We hypothesize that HyDE-generated documents, which replace the query entirely, may introduce their own positional patterns that interact adversely with document encodings. In contrast, Query2Doc—which preserves the original query while augmenting it—achieves the strongest position bias reduction of any method (5.28 → 2.81, a 47% decrease). This suggests that retaining the original query signal provides an anchoring effect that prevents the introduction of new biases.

Brevity bias remains the most severe. Across all methods, brevity bias exhibits the highest $|t|$-statistics, indicating that dense retrievers' preference for shorter documents is particularly robust to query-side interventions. Even the best-performing method (*Rewrite*) leaves a residual bias of 6.62, compared to near-complete mitigation for position bias (2.81 with *Q2D*). This asymmetry suggests that brevity bias may be more deeply encoded in document representations, requiring retriever-level interventions rather than query modifications alone.

Literal matching responds well to most techniques. All query enhancement techniques substantially reduce literal matching bias, with an average reduction of 39.12%. This suggests that these techniques dilute exact lexical overlap between short queries and documents, shifting retrieval from lexical toward semantic matching.

Repetition bias proves resistant to pseudo-document methods. While simple rewriting cuts repetition bias by 54% (5.17 → 2.37), HyDE and Query2Doc show minimal improvement (5.17 → 4.86 and 4.99, respectively). This asymmetry suggests that LLM-generated pseudo-documents may inadvertently repeat query terms—a natural consequence of generating text conditioned on answering the query—thereby perpetuating rather than mitigating repetition bias. See Appendix A for examples of generated pseudo-documents which confirms the same.

4.3 Variation Across Dense Retrievers

Figure 3 reveals the heterogeneity in how different retrievers respond to various query enhancement techniques.

Retriever-specific effects. We observe that simple query rewriting achieves consistent bias reduction across all six retrievers, with mean $|t|$-statistics ranging from 2.07 (CoCoDR) to 5.03 (Col-

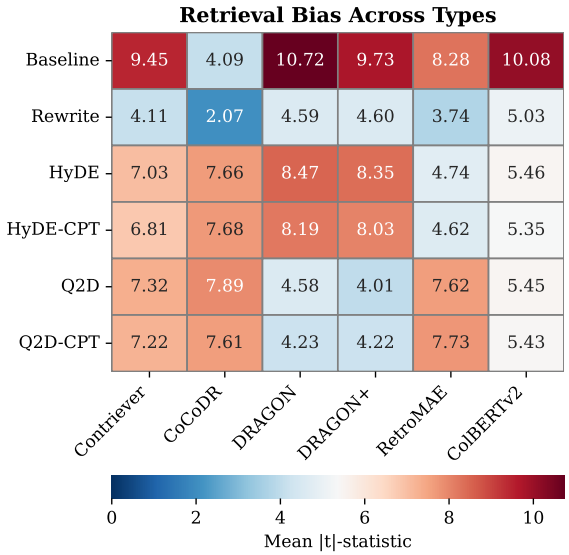

Figure 3: **Mean $|t|$-statistic by retriever and query enhancement method, averaged across all four bias types.** Lower values indicate reduced bias. Simple rewriting consistently reduces bias, while pseudo-document methods show differential effects.

BERTv2). This consistency stands in stark contrast to the pseudo-document methods, which exhibit highly retriever-dependent effects. For example, Q2D dramatically reduces bias for DRAGON and DRAGON+ (10.72 $\rightarrow$ 4.58 and 9.73 $\rightarrow$ 4.01, respectively), yet *increases* bias for CoCoDR (4.09 $\rightarrow$ 7.89) and RetroMAE (8.28 $\rightarrow$ 7.62). Similarly, HyDE substantially harms CoCoDR (4.09 $\rightarrow$ 7.66), nearly doubling its bias level. These findings suggest that the interaction between pseudo-document generation and retriever architecture is non-trivial and potentially unpredictable.

Architectural patterns. Several architecture-specific trends emerge from these results. The DRAGON variants—both trained with diverse augmentation strategies—respond exceptionally well to Query2Doc, achieving bias levels comparable to simple rewriting (4.01–4.58 vs. 4.59–4.60). We hypothesize that these retrievers, having been exposed to diverse query-document pairings during training, are better equipped to leverage the augmented context that Query2Doc provides. In contrast, CoCoDR exhibits anomalous behavior: despite having the lowest baseline bias (4.09), it is the only retriever for which all pseudo-document methods increase bias. This may reflect CoCoDR's continuous contrastive distillation objective, which could make it sensitive to the distributional shift introduced by LLM-generated content. ColBERTv2's late interaction architecture yields remarkably stable results across all enhancement methods (5.03–5.46), suggesting that token-level matching may be inherently more robust to query formulation changes than single-vector representations.

Method	RetroMAE	Contriever	CoCoDR
Baseline	0.0%	0.8%	2.0%
Rewrite	0.0%	0.8%	3.6%
HyDE	2.8%	4.0%	20.0%
HyDE-CPT	2.8%	5.2%	26.0%
Q2D	2.8%	3.2%	18.0%
Q2D-CPT	3.2%	3.6%	19.2%
Best improv.	—	6.5×	13×

Table 2: **FOIL dataset accuracy across retrievers.** CoCoDR shows the strongest response to query enhancement, with HyDE-CPT achieving 26% accuracy (13× over baseline). Simple rewriting provides minimal improvement across all retrievers.

4.4 Effect of Query Enhancement on Interplay Between Bias Types

While Section 4 demonstrated that query rewriting techniques reduce individual bias types, a natural question arises: *do these improvements transfer to scenarios where multiple biases act concurrently?*

To investigate this, we evaluate on the FOIL subset of ColDeR, which presents an adversarial set of 250 document pairs where document D_1 exploits multiple bias-inducing features (brevity, answer position, lexical overlap, repetition) while lacking the answer, and D_2 contains the correct evidence embedded within unrelated context. We report results for RetroMAE, Contriever, and CoCoDR here, to capture the spectrum of performances.

From Table 2, we see that simple query rewriting provides minimal improvement on FOIL despite achieving the largest reduction in individual bias metrics (Section 4). In contrast, pseudo-document generation methods show substantial gains—on CoCoDR, HyDE achieves 10× improvement (20.0%) and Query2Doc achieves 9× improvement (18.0%) over baseline. These methods fundamentally alter the query representation through semantic expansion and length normalization, thereby providing more robust resistance to adversarial bias combinations. CPT provides additional benefit over vanilla techniques, with HyDE-CPT achieving the strongest performance at 26.0% (13× over baseline), which suggests that document-specific fine-tuning addresses not just relevance matching but also enhances bias robustness.

These findings motivate a deeper investigation:

what distinguishes methods that genuinely reduce bias sensitivity from those that merely obscure it? We address this question through a deeper mechanistic analysis in the following section.

5 Mechanistic Analysis of Bias Reduction

We now investigate why query-enhancement techniques might succeed or fail at mitigating biases. We conduct an analysis to measure how rewriting affects the correlation between retrieval scores and bias-inducing features.

If a retriever exhibits bias toward a particular document feature (e.g., shorter length, answer position), we expect retrieval scores to correlate with that feature independent of relevance. Our intuition is that effective bias mitigation should reduce this correlation, and we operationalize this intuition by computing the Spearman correlation ρ between retrieval scores and bias-inducing features across the ColDER benchmark.

For each query q (original or rewritten) and document d, we compute the retrieval score $\mathcal{S}(q, d)$ and extract the corresponding bias feature $f(q, d)$:

(i) **Brevity bias:** Document length in tokens, $f(q, d) = |d|$.

(ii) **Literal matching bias:** Jaccard similarity (Jaccard, 1901) between query and document term sets. Let $\mathcal{T}(q)$ and $\mathcal{T}(d)$ denote the sets of unique terms in query q and document d, respectively. Then, we define:

$$f(q, d) = \frac{|\mathcal{T}(q) \cap \mathcal{T}(d)|}{|\mathcal{T}(q) \cup \mathcal{T}(d)|}$$

(iii) **Position bias:** Normalized answer position within the document, where $\text{pos}(a, d)$ denotes the character offset of answer span a in document d:

$$f(q, d) = \frac{\text{pos}(a, d)}{|d|} \in [0, 1]$$

(iv) **Repetition bias:** Average term frequency of query terms within the document. Let $\text{tf}(t, d)$ denote the number of occurrences of term t in document d:

$$f(q, d) = \frac{1}{|\mathcal{T}(q)|} \sum_{t \in \mathcal{T}(q)} \text{tf}(t, d)$$

We then compute $\rho_c = \text{Spearman}(\mathcal{S}_c, f)$ for each query condition c. A retriever with no systematic bias would yield $\rho \approx 0$, while higher absolute correlations indicate stronger bias. We quantify bias reduction as:

$$\Delta\rho = |\rho_{\text{original}}| - |\rho_{\text{rewritten}}| \quad (1)$$

where $\Delta\rho > 0$ indicates that rewriting reduced the retriever's sensitivity to the bias-inducing feature.

Method	Brevity	Literal	Position	Repetition
Baseline $\lvert\rho_q\rvert$	0.36	0.43	0.11	0.21
Rewrite	+10%	−7%	−373%	−112%
HyDE	+56%	+20%	−2%	−59%
HyDE-CPT	+49%	+19%	+54%	−9%
Q2D	+48%	+36%	−55%	−13%
Q2D-CPT	+53%	+32%	+23%	−21%

Table 3: **Feature-score decorrelation analysis.** $|\rho_q|$ represents baseline Spearman correlation for query q. Values show percentage reduction in $|\rho|$ after rewriting query q (positive = reduced sensitivity, negative = increased sensitivity). Results averaged across all 6 dense retrievers studied.

Findings: Table 3 presents the mean correlation reduction $\Delta\rho$ across six dense retrievers. We observe several key patterns:

Brevity bias. All rewriting methods substantially reduce correlation with document length. Pseudo-document methods achieve the largest gains (HyDE: +56%, Q2D: +48%), consistent with our hypothesis that generating longer text shifts the query representation away from favoring brief documents. This finding aligns with Section 4, where all methods reduced brevity bias.

Literal matching bias. Query2Doc variants achieve substantially larger reductions in literal bias compared to HyDE. We attribute this to a key mechanistic difference: Query2Doc preserves original query terms while adding semantic context, effectively *diluting* rather than *replacing* lexical signals. HyDE's complete query replacement may introduce new lexical artifacts from the generated pseudo-document. Notably, simple rewriting slightly *increases* literal bias (−7%), suggesting that paraphrasing without content expansion actually reinforces lexical overlap as the rewritten query may use more "retrieval-friendly" vocabulary.

Repetition bias. All approaches *increase* sensitivity to repetition, with simple rewriting showing the most severe degradation. This represents a systematic failure that explains the resistance of repetition bias to pseudo-document methods. Analysis of generated pseudo-documents shows that LLMs naturally produce repetitive text patterns when answering queries, which correlates with documents

containing repeated query terms. The $|t|$-statistic improvements likely reflect increased score variance rather than true debiasing.

Position bias. Position bias reveals the strongest mechanistic differences. Simple rewriting greatly position sensitivity, yet achieved strong $|t|$-statistic reduction in Section 4. This paradox suggests that simple rewriting reduces position bias through increased retrieval noise rather than principled decorrelation. In contrast, HyDE-CPT achieves genuine decorrelation. Only Q2D-CPT combines strong $|t|$-statistic reduction with positive decorrelation, suggesting that continual pretraining is essential for robust position bias mitigation.

Reconciling $|t|$-statistics and decorrelation: These findings reveal a crucial insight that *reducing overall bias (as measured by $|t|$-statistics) does not imply reducing sensitivity to bias-inducing features*. We identify two distinct mechanisms:

(i) Variance-based reduction: Simple rewriting achieves strong $|t|$-statistic improvements while *increasing* feature-score correlations for position and repetition biases. The mechanism appears to be increased retrieval score variance—noisier scores reduce the statistical power to detect systematic preferences, lowering $|t|$-statistics without addressing underlying bias sensitivity. This also explains why simple rewriting fails entirely on FOIL (Table 2); when biases act together, the underlying sensitivities may compound further.

(ii) Decorrelation-based reduction: Pseudo-document methods, particularly the CPT variants, achieve bias reduction through principled decorrelation from bias-inducing features. Specifically, HyDE-CPT and Q2D-CPT show positive decorrelation for brevity, literal matching, *and* position biases simultaneously. This more robust mechanism explains their gains on FOIL, where genuine insensitivity to bias features provides resistance to adversarial combinations.

To further understand why simple rewriting shows contrasting effects, we analyzed the linguistic transformations applied by the LLM-based rewriting (Table 4 in Appendix C). Our key observation is that these transformations are predominantly *syntactic rather than semantic*, introducing sufficient variance to reduce the paired t-statistic without actually reducing the retriever's sensitivity to bias-inducing features.

6 Discussion

Implications for researchers. Our findings suggest a taxonomy of retrieval biases based on their responsiveness to query-side interventions. *(i) Query-document interaction biases*—such as literal matching—arise from how queries and documents are compared and can be mitigated through query transformation. *(ii) Document encoding biases*—particularly position bias—appear embedded in document representations and persist regardless of query formulation, likely requiring retriever-level modifications. Additionally, our decorrelation analysis reveals that aggregate metrics can obscure different underlying mechanisms, and we recommend future work also report feature-score correlations. Finally, the increase in repetition sensitivity across all methods represents a systematic failure mode that warrants further investigation.

Implications for practitioners. When selecting query enhancement techniques, practitioners should consider specific bias vulnerabilities rather than aggregate metrics. Simple rewriting achieves strong overall reduction but fails when multiple biases act together. Critically, enhancement effects are also retriever-dependent, so techniques should be validated on specific retrievers before deployment. For high-stakes domains, we recommend CPT variants despite computational overhead, as our findings reveal that decorrelation-based reduction generalizes better to adversarial conditions than variance-based reduction.

7 Conclusion

In this work, we present the first systematic study of how query enhancement techniques affect dense retrieval biases in RAG systems. Our evaluation reveals that while all techniques reduce aggregate bias, they operate through fundamentally different mechanisms. Simple rewriting achieves strong overall reduction but increases sensitivity to bias-inducing features; pseudo-document methods with continual pretraining achieve more robust improvements through genuine decorrelation. We establish a taxonomy distinguishing query-document interaction biases, which yield to query-side interventions, from document encoding biases, which likely require retriever-level modifications. Our findings provide practical guidance for deploying bias-aware RAG systems, and highlight the continued need for retrieval paradigms that prioritize semantic relevance over superficial characteristics.

8 Limitations

(1) Benchmark and retriever scope: Our evaluation relies on the ColDER benchmark, which provides controlled document pairs derived from Re-DocRED. While this controlled setup enables precise bias measurement, it may not fully capture the distribution of biases in real-world retrieval scenarios where multiple subtle biases co-occur in less structured ways. Additionally, we evaluate only dense bi-encoder and late-interaction retrievers, primarily because of their widespread utility and effectiveness today. Therefore, our findings may not generalize to sparse retrievers (e.g., BM25), hybrid systems, or emerging retrieval paradigms such as generative retrieval.

(2) Query enhancement coverage: We evaluate five query enhancement techniques using two open-source LLMs (Gemma-3 and Qwen3). Other methods such as Step-Back Prompting, multi-query fusion, or chain-of-thought augmented retrieval may exhibit different bias profiles. However, our key contribution is not tied to a specific query-enhancement type. Furthermore, proprietary LLMs with stronger instruction-following capabilities might produce pseudo-documents with different characteristics, potentially affecting bias outcomes, which we could not evaluate in this work due to cost constraints.

(3) Downstream evaluation: Our analysis focuses on retrieval-stage biases measured through score differences and correlations. We do not evaluate how these biases propagate to downstream generation quality in end-to-end RAG systems. A retrieval bias that appears significant in isolation may be attenuated, or amplified, by the downstream language model, and understanding this interaction remains an important direction for future work.

Ethical Considerations

Our work aims to understand and mitigate biases in dense retrieval systems, which we view as a net positive for the responsible and fair deployment of RAG systems. However, we acknowledge that detailed characterization of retrieval biases could theoretically be misused to craft adversarial documents that exploit these biases to manipulate retrieval rankings. However, we believe the benefits of transparency outweigh this risk, as awareness of these biases enables practitioners to implement appropriate safeguards.

More broadly, the biases we study have implications for fairness in information access for users of modern generative search systems. For example, brevity bias may systematically disadvantage comprehensive sources, and literal matching bias may favor keyword-filled content over semantically rich sources. These effects are particularly concerning in high-stakes domains such as legal, sociotechnical, and medical information retrieval, where biased retrieval could lead to incomplete or skewed information being surfaced to end users. Our findings therefore underscore the importance of bias-aware evaluation in retrieval system development.

References

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen-tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2023. Can retriever-augmented language models reason? the blame game between the retriever and the language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15492–15509, Singapore. Association for Computational Linguistics.

João Coelho, Bruno Martins, Joao Magalhaes, Jamie Callan, and Chenyan Xiong. 2024. Dwell in the beginning: How language models embed long documents for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–377, Bangkok, Thailand. Association for Computational Linguistics.

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, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Agam Goyal, Xianyang Zhan, Yilun Chen, Koustuv Saha, and Eshwar Chandrasekharan. 2025. MoMoE: Mixture of moderation experts framework for AI-assisted online governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12645–12660, Suzhou, China. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and 1 others. 2023. Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses. *arXiv preprint arXiv:2312.15883*.

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.

Daehui Kim, Deokhyung Kang, Jonghwi Kim, Sangwon Ryu, and Gary Lee. 2025. GuRE:generative query REwriter for legal passage retrieval. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 424–438, Suzhou, China. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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. *arXiv preprint arXiv:2302.07452*.

Shuai Lu, Nan Duan, Hojae Han, Daya Guo, Seung-won Hwang, and Alexey Svyatkovskiy. 2022. ReACC: A retrieval-augmented code completion framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6227–6240, Dublin, Ireland. Association for Computational Linguistics.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. RaFe: Ranking feedback improves query rewriting for RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 884–901, Miami, Florida, USA. Association for Computational Linguistics.

Agada Joseph Oche, Ademola Glory Folashade, Tirthankar Ghosal, and Arpan Biswas. 2025. A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions. *arXiv preprint arXiv:2507.18910*.

Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. What are you token about? dense retrieval as distributions over the vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2481–2498, Toronto, Canada. Association for Computational Linguistics.

Amanda Ross and Victor L Willson. 2017. Paired samples t-test. In *Basic and advanced statistical tests: Writing results sections and creating tables and figures*, pages 17–19. Springer.

Michael J Ryan, Danmei Xu, Chris Nivera, and Daniel Campos. 2025. Enronqa: Towards personalized rag over private documents. *arXiv preprint arXiv:2505.00263*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Soorya Ram Shimgekar, Ruining Zhao, Agam Goyal, Violeta J Rodriguez, Paul A Bloom, Hari Sundaram, and Koustuv Saha. 2025. Detecting early and implicit suicidal ideation via longitudinal and information environment signals on social media. *arXiv preprint arXiv:2510.14889*.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Baiqiang Wang, Qian Lou, Mengxin Zheng, and Dongfang Zhao. 2025. Pir-rag: A system for private information retrieval in retrieval-augmented generation. *arXiv preprint arXiv:2509.21325*.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.

Eric W Weisstein. 2004. Bonferroni correction. *https://mathworld. wolfram. com/*.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. Coco-dr: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. *arXiv preprint arXiv:2210.15212*.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*.

A Examples of Generated Query Enhancements

In this section we provide examples of query enhancements across methods.

Vanilla Queries

1. When was The Private Life of Helen of Troy published?

2. Where is Los Angeles located in or next to a body of water?

3. Where was Cesare Mori born?

LLM Rewritten Queries

1. What is the publication date of The Private Life of Helen of Troy?

2. What body of water is Los Angeles located near or in?

3. What is the birthplace of Cesare Mori?

HyDE Queries

1. The “Private Life of Helen of Troy” was published in 1920 by the renowned author, Robert Graves. This groundbreaking work, a reimagining of the Trojan War legend, offered a fresh perspective on Helen’s character and the events surrounding the war.

2. Los Angeles is located on the southern coast of California, a state situated on the Pacific Ocean. The city is situated on the shores of the Pacific Ocean, with the Pacific Ocean bordering its western side. The city is also bordered by the Los Angeles Basin, a large inland basin that is partially surrounded by mountains.

3. Cesare Borgia, the Italian Renaissance figure, died in Rome, Italy on 18th of January 1507. He was a powerful and controversial figure who served as Cardinal and later as the Duke of Valentinois. His death was a significant event in the history of the Borgia family and the Italian Renaissance.

Query2Doc Queries

1. When was The Private Life of Helen of Troy published?
The “Private Life of Helen of Troy” was published in 1920 by the renowned author, Robert Graves. This groundbreaking work, a reimagining of the Trojan War legend, offered a fresh perspective on Helen’s character and the events surrounding the war.

2. Where is Los Angeles located in or next to a body of water?

Los Angeles is located on the southern coast of California, a state situated on the Pacific Ocean. The city is situated on the shores of the Pacific Ocean, with the Pacific Ocean bordering its western side. The city is also bordered by the Los Angeles Basin, a large inland basin that is partially surrounded by mountains.

3. Where was Cesare Mori born?
Cesare Borgia, the Italian Renaissance figure, died in Rome, Italy on 18th of January 1507. He was a powerful and controversial figure who served as Cardinal and later as the Duke of Valentinois. His death was a significant event in the history of the Borgia family and the Italian Renaissance.

B Results for Qwen3

To verify that our findings are not specific to the choice of language model used for query enhancement, we replicate our main experiments using Qwen3-4B-Instruct as an alternative to Gemma-3-12B-IT. Figure 4 presents the complete bias analysis across all retrievers and bias types.

The results show consistent patterns, with simple rewriting achieveing substantial bias reduction on individual metrics, while HyDE and Query2Doc show similar moderate improvements. This consistency across model architectures and scales (4B vs. 12B parameters) suggests that our findings reflect fundamental properties of query transformation strategies rather than artifacts of a specific LLM's generation characteristics.

C Further Analysis of Simple LLM-Based Query Rewriting

Metric	Baseline	Rewrite
Avg. query length (words)	7.3	8.3
Length change (words)	—	+1.0
Term preservation (%)	100	57
New terms introduced	—	3.8
Unique vocabulary	420	489
Entity preservation (%)	100	72.5

Table 4: **Query transformation characteristics across enhancement methods.** Simple rewriting makes minimal syntactic changes (high term/entity preservation). This explains why rewriting reduces measured bias variance but fails on adversarial FOIL examples.

Table 4 reveals that simple LLM-based rewriting produces minute changes, with queries increase by only 1 word on average (7.3→8.3 words), preserving 57% of original terms while introducing 3.8 new terms. The transformations are therefore predominantly *syntactic rather than semantic*. For example, *"When was X published?"* becomes *"What is the publication date of X?"* The key entity names (e.g., *"Lake Ewauna," "Miami Sound Machine"*) remain intact in 72.5% of cases.

This explains the paradox we observed in our results in Section 4 and Section 5: syntactic reformulation introduces sufficient variance to reduce the paired t-statistic—a measure of *consistency* in bias direction—without actually reducing the retriever's sensitivity to bias-inducing lexical features. Therefore, when multiple biases compound in the FOIL setting, this surface-level variance fails to provide robustness.

D Compute Resources

All experiments on open-source models were run on internal organization GPU servers equipped with 2xNVIDIA H100 and 3xNVIDIA A40.

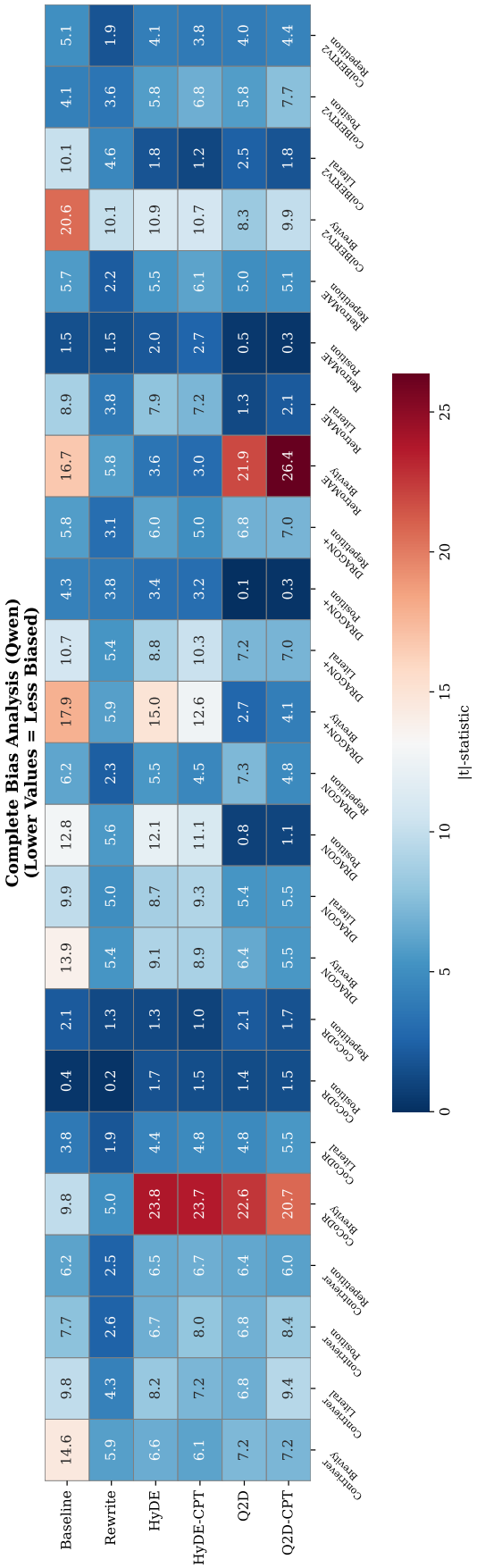

Figure 4: **Complete bias analysis using Qwen3-4B-Instruct for query enhancement.** Each cell shows the $|t|$-statistic measuring retrieval bias strength across retrievers and biases. Lower values indicate reduced bias. Results demonstrate consistent patterns with Gemma-3-12B-IT, confirming generalizability of our findings.