Collapse of Dense Retrievers:
Short, Early, and Literal Biases Outranking Factual Evidence

Anonymous ACL submission

Abstract

Dense retrieval models are commonly used in Information Retrieval (IR) applications, such as Retrieval-Augmented Generation (RAG). Since they often serve as the first step in these systems, their robustness is critical to avoid failures. In this work, by repurposing a relation extraction dataset (e.g. Re-DocRED), we design controlled experiments to quantify the impact of heuristic biases, such as favoring shorter documents, in retrievers like Dragon+ and Contriever. Our findings reveal significant vulnerabilities: retrievers often rely on superficial patterns like *over-prioritizing document beginnings, shorter documents, repeated entities, and literal matches*. Additionally, they tend to overlook whether the document contains the query's answer, lacking deep semantic understanding. Notably, *when multiple biases combine, models exhibit catastrophic performance degradation*, selecting the answer-containing document in less than 3% of cases over a biased document without the answer. Furthermore, we show that *these biases have direct consequences for downstream applications like RAG*, where retrieval-preferred documents can mislead LLMs, resulting in a 34% performance drop than not providing any documents at all. [1]

1 Introduction

Retrieval-based language models have demonstrated strong performance on a range of knowledge-intensive NLP tasks (Lewis et al., 2020; Asai et al., 2023; Gao et al., 2024). At the core of these models is a retriever that identifies relevant context to ground the generated output. Dense retrieval methods such as Contriever (Izacard et al., 2021)—where passages or documents are stored as learned embeddings—are especially appealing for their scalability across large knowledge bases and handling lexical gaps (Ni et al., 2022; Shao et al.,
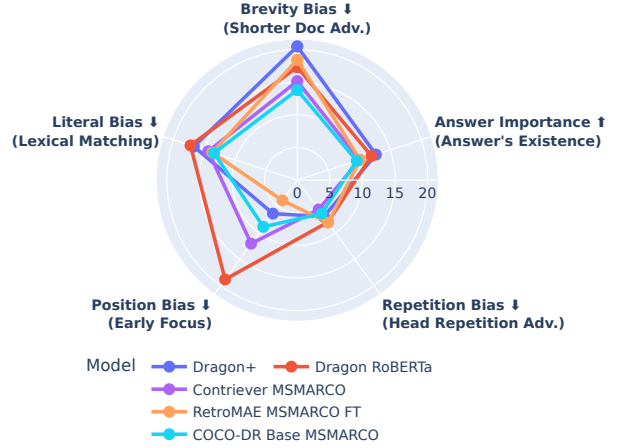
---

[1]Code and data will be released upon paper acceptance.

Figure 1: Paired t-test statistics comparing retriever scores between document pairs (D_1 vs. D_2) across five evaluation aspects. Document pairs are designed for controlled experiments shown in Table 1. Positive values indicate a retriever's preference for the more biased document in each bias scenario, while for answer importance, they reflect a preference for answer-containing documents. The results show that retrieval biases often outweigh the impact of answer presence.

2024), compared to alternatives like BM25 (Robertson and Zaragoza, 2009) or ColBERT(Khattab and Zaharia, 2020). Yet despite their widespread use, relatively little is understood about how these dense models encode and organize information, leaving key questions about their robustness against potential biases unanswered. Existing evaluations of retrieval models often focus on downstream task performance, as seen in benchmarks like BEIR (Thakur et al., 2021), without probing the underlying behavior of retrievers. Some studies have analyzed specific issues in information retrieval (IR) models, such as position bias (Coelho et al., 2024) or lexical overlap (Ram et al., 2023).

In this work, we systematically study multiple biases—both individually and in combination—for the first time. To enable fine-grained control

	Document 1 (Higher Query Document Similarity Score) - D_1	Document 2 (Lower Query Document Similarity Score) - D_2
Answer Impact	**Query:** What is the sister city of *Leonessa*? **Document:** *Leonessa* is twinned with the French town of *Gonesse* . Its population in 2008 was around 2,700 . Situated in a small plain at the foot of Mt. Terminillo	**Query:** What is the sister city of *Leonessa*? **Document:** *Leonessa* is a town and comune in the far northeastern part of the Province of Rieti in the Lazio region of central Italy . Its population in 2008 was around 2,700 . Situated in a small plain at the foot of Mt. Terminillo .....
Position Bias	**Query:** Which country is *Wonyong Sung* a citizen of? **Document:** *Wonyong Sung* (born 1950s) , *South Korean* professor of electronic engineering Won - yong is a Korean masculine given name ..... People with this name include : Kang Won - yong (1917 – 2006) ..... , South Korean swimmer	**Query:** Which country is *Wonyong Sung* a citizen of? **Document:** Won - yong is a Korean masculine given name ..... People with this name include : .... Jung Won - yong (born 1992) , South Korean swimmer *Wonyong Sung* (born 1950s) , *South Korean* professor of electronic engineering
Literal Bias	**Query:** When was *Seyhun* born? **Document:** *Seyhun* , (*August 22 , 1920* – May 26 , 2014) was an Iranian architect , sculptor , painter , scholar and professor . He studied fine arts at	**Query:** When was *Seyhun* born? **Document:** *Houshang Seyhoun* , (*August 22 , 1920* – May 26 , 2014) was an Iranian architect , sculptor , painter , scholar and professor . He studied fine arts at .....
Brevity Bias	**Query:** What series is *Lost Verizon* part of? **Document:** " *Lost Verizon* " is the second episode of *The Simpsons* ' twentieth season .	**Query:** What series is *Lost Verizon* part of? **Document:** " *Lost Verizon* " is the second episode of *The Simpsons* ' twentieth season . It first aired on the Fox network in the United States on October 5 , 2008 . Bart becomes jealous of his friends and their cell phones . Working at a golf course , Bart takes the cell phone of Denis Leary .....
Repetition Bias	**Query:** Where was *James Paul Maher* born? **Document:** Born in *Brooklyn , New York* , *Maher* graduated from St. Patrick 's Academy in Brooklyn . *James Paul Maher* (November 3 , 1865 – July 31 , 1946) was a U.S. Representative from New York . *Maher* was elected as a Democrat to the Sixty - second and to the four succeeding Congresses (March 4 , 1911 – March 4 , 1921) .	**Query:** Where was *James Paul Maher* born? **Document:** Born in *Brooklyn , New York* , *Maher* graduated from St. Patrick 's Academy in Brooklyn . Apprenticed to the hatter 's trade , he moved to Danbury , Connecticut in 1887 and was employed as a journeyman hatter . He became treasurer of the United Hatters of North America in 1897 .
Foil vs. Evide.	**Query:** Who is the publisher of *Assassin 's Creed Unity*? **Document:** " *Assassin 's Creed Unity* " " *Assassin 's Creed Unity* " *Assassin 's Creed Unity* received mixed reviews upon its release .	**Query:** Who is the publisher of *Assassin 's Creed Unity*? **Document:** Isa is a town and Local Government Area in the state of Sokoto in Nigeria . It shares borders with ..... *Assassin 's Creed Unity* is an action - adventure video game developed by Ubisoft Montreal and published by *Ubisoft*. Isa is a town and Local Government Area in the state of Sokoto in Nigeria . It shares borders with .....

Table 1: Examples from our framework highlighting Evidence, *Head Entity*, and *Tail Entity*. In all cases, retrieval models favor Document 1 over Document 2, assigning higher retrieval scores accordingly. (Explained in §3.3)

over document structure and factual positioning, we repurpose a document-level relation extraction dataset (Re-DocRED (Tan et al., 2022)).

We first investigate biases individually, identifying tendencies such as an **over-prioritization of document beginnings, document brevity, repetition of matching entities, and literal matches at the expense of ignoring answer presence.** Our statistical approach, illustrated in Figure 1, allows for comparative analysis across different biases. Additionally, we explore the interplay between these biases and propose an adversarial benchmark that combines multiple vulnerabilities.

The results of our investigation reveal concerning patterns in current retriever architectures. **When exposed to multiple interacting biases, even top-performing models exhibit dramatic degradation**, selecting the answer-containing document over the foil document—filled with biases—less than 3% of the time. Moreover, we demonstrate that **these biases can be exploited to manipulate Retrieval-Augmented Generation (RAG)**, causing retrievers to favor misleading or adversarially constructed documents, which misguide LLMs into using incorrect information and ultimately degrade its performance.

2 Related Work

Benchmarking in Information Retrieval Popular benchmarks like BEIR (Thakur et al., 2021; Guo et al., 2021; Petroni et al., 2021; Muennighoff et al., 2023) have played a crucial role in evaluating retrieval models across diverse datasets and tasks. In addition to general IR benchmarks, domain-specific benchmarks such as COIR (Li et al., 2024) for code retrieval and LitSearch (Ajith et al., 2024) for scientific literature search address retrieval challenges in specialized domains. While these benchmarks have advanced the evaluation of IR models, they primarily focus on downstream performance rather than conducting systematic analyses of biases inherent in retrieval systems.

Information Retrieval Model Analysis Prior work in information retrieval has explored various dimensions of retrieval performance, including positional biases (Coelho et al., 2024). Studies have also examined how dense retrievers exhibit biases towards common entities and struggle with OOD scenarios (Sciavolino et al., 2021). Furthermore, analysis by projecting representations to the vocabulary space has shown that supervised dense retrievers tend to learn relying heavily on lexical overlap during training (Ram et al., 2023). Similarly, BehnamGhader et al. (2023) has indicated that Dense Passage Retrieval (DPR) models often

fail to retrieve statements requiring reasoning beyond surface-level similarity. Furthermore, neural IR models have been shown to exhibit over-penalization for extra information, where adding a relevant sentence to a document can unexpectedly decrease its ranking (Usuha et al., 2024). Additionally, Reichman and Heck (2024) takes a mechanistic approach to analyze the impact of DPR fine-tuning, showing that while fine-tuned models gain better access to pre-trained knowledge, their retrieval capabilities remain constrained by the pre-existing knowledge in their base models. Further, MacAvaney et al. (2022) provides a framework for analyzing neural IR models, identifying key biases and sensitivities in these models.

Adversarial Attacks in Information Retrieval
Numerous studies have explored various dimensions of robustness in information retrieval, including aspects related to adversarial robustness (Liu et al., 2024). Adversarial perturbations, for instance, have been shown to significantly degrade BERT-based rankers' performance, revealing their brittleness to subtle modifications (Wang et al., 2022). Existing retrieval attack methods primarily encompass corpus poisoning (Lin et al., 2024; Zhong et al., 2023), backdoor attacks (Long et al., 2024), and encoding attacks (Boucher et al., 2023).

While previous work has analyzed some retrieval biases, most studies focus on task-specific supervised models and a single aspect in isolation. Our work provides a comprehensive comparative analysis of popular retrieval models across multiple dimensions of vulnerability. We systematically investigate how these biases interact and affect the retrieval capabilities of dense retrievers. By re-purposing a relation extraction dataset, we gain precise control over factual information in documents, enabling a rigorous evaluation of retrieval robustness. This multi-dimensional approach provides a nuanced understanding of the strengths and weaknesses of dense retrievers.

3 Experiments

3.1 A Framework for Identifying Biases in Retrievers

To gain fine-grained control over the facts present in a document, we take a novel approach by repurposing a relation extraction dataset that provides relation-level fact granularity. This enables a structured analysis of retrieval biases by explicitly linking queries to individual factual statements.

Model	pooling	nDCG@10	Recall@10
Dragon RoBERTa	cls	**0.55**	**0.75**
Dragon+	cls	0.54	0.74
COCO-DR Base MSMARCO	cls	0.50	0.71
Contriever MSMARCO	avg	0.50	0.71
RetroMAE MSMARCO FT	cls	0.48	0.68
Contriever	avg	0.25	0.41

Table 2: Models' performance on NQ dataset with test set queries and 2,681,468 corpus size.

One such dataset is DocRED (Yao et al., 2019), a relation extraction dataset constructed from Wikipedia and Wikidata. DocRED consists of human-annotated triplets (head entity, relation, tail entity)—for example, (Albert Einstein, educated at, University of Zurich). However, DocRED suffers from a significant false negative issue, as many valid relations are missing from the annotations. To address this, we use Re-DocRED (Tan et al., 2022), a re-annotated version of DocRED that recovers missing facts, leading to more complete and reliable annotations.

To construct a retrieval dataset from Re-DocRED, we map each relation to a query template. For example, for the relation "educated at," we use the template "Where was {Head Entity} educated?" This transformation allows us to systematically examine how retrievers handle different types of factual queries.

The answers to these queries are the tail entities found in the evidence sentences provided by the dataset. For our analysis, we ensure that each query has a single evidence sentence (S_{ev}) within the original document ($S \in D_{orig}$) that contains both the head and tail entities. This constraint makes the sentence self-contained, allowing for precise control over the document structure in subsequent sections. We also introduce the notation S_{-t}^{+h} for sentences in D_{orig} that contain the head entity but not the tail entity, and S_{-t}^{-h} for sentences that do not contain either entity. In each of the following sections, we will use this notation to construct a pair of document sets, D_1 and D_2, enabling a systematic investigation of retrieval score variations and potential biases. As a result, for each of our six analysis settings, we compile 250 queries, each with a single corresponding gold document, based on the test and validation sets of Re-DocRED.

3.2 Models Performance & Bias Discovery

First, we evaluate several dense retrievers on the NQ dataset (Kwiatkowski et al., 2019), comparing their performance using nDCG@10 and Re-
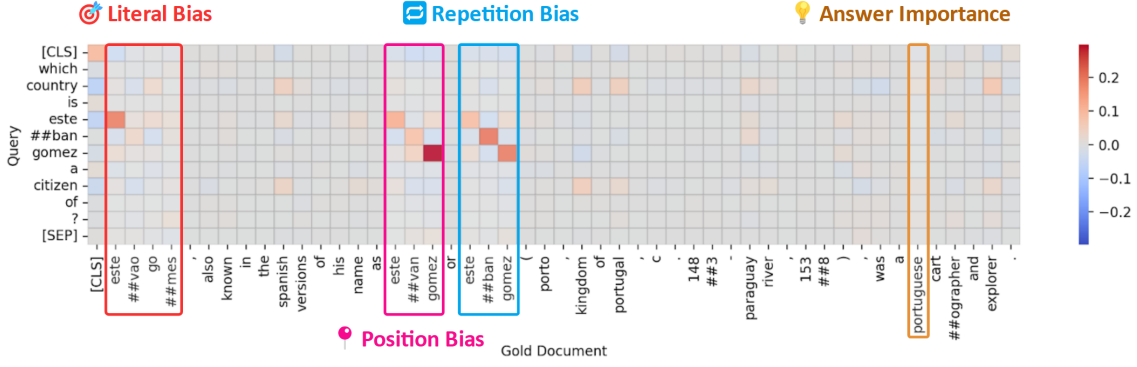
Figure 2: Visualization of the contribution of each query and document token to the final retrieval score using DecompX. Literal Bias reflects the model's preference for exact word matches, such as failing to match "esteban goemz" with "estevao gomes." Position Bias indicates a preference for entities earlier in the document receiving more attention. Repetition Bias shows that repeating an entity multiple times increases its score. Lastly, Answer Importance demonstrates that the query's answer entity receives less attention compared to head entity matches.

call@10 metrics. Table 2 shows that Dragon models lead in performance, and the significant improvement of fine-tuned Contriever over its unsupervised counterpart highlights the importance of supervision and task-specific adaptation. Models also differ in pooling mechanisms, with Contriever using average pooling and others using CLS pooling. For details, refer to the appendix A.1.

In our preliminary analysis, we utilized DecompX (Modarressi et al., 2023, 2022), a method that decomposes the representations of encoder-based models such as BERT into their constituent token-based representations. By applying DecompX to the embeddings generated by dense retrievers, we obtain decomposed representations for both the query and the document. Instead of using the original embeddings, we compute the similarity score via a dot product of the decomposed vectors. This approach enables us to visualize the contribution of each query and document token to the final similarity score as a heatmap (Figure 2), revealing biases in token-level interactions.

In our preliminary error analysis of 60 retrieval failure examples, we identified potential biases and limitations in the models (Table A.3). Figure 2 highlights some of these biases, such as Literal Bias, where the term "esteban gomez" fails to match "estevao gomez," reflecting a preference for exact matching. In subsequent sections, we design experiments and perform statistical tests to evaluate these observed biases.

3.3 Bias Types in Dense Retrieval

The following experiments are meticulously designed to control for all other factors and biases, isolating the specific bias under evaluation.

3.3.1 Answer Importance

An effective retrieval model should accurately identify the query's intent. It should retrieve relevant documents that address the query, rather than just matching entities. To assess whether dense retrieval models truly recognize the presence of answers or merely focus on entity matching, we developed a controlled experimental framework. Our experimental design contrasts two carefully constructed document types. **1. Document with Evidence**: Contains a leading evidence sentence with both the head entity and the tail entity (answer). **2. Document without Evidence D_2**: Contains a leading sentence with only the head entity but no tail.

$$
\begin{aligned}
D_1 &\coloneqq S_{ev} + \sum_{S_{-t}^{-h} \in D_{orig}} S_{-t}^{-h} \\
D_2 &\coloneqq S_{-t}^{+h} + \sum_{S_{-t}^{-h} \in D_{orig}} S_{-t}^{-h}
\end{aligned}
\tag{1}
$$

Here, S_{-t}^{+h} is another sentence from D_{orig} that replaces the original evidence S_{ev} while containing the head entity but not containing the tail entity to isolate the impact of answer presence. The remainder of both documents consists of neutral sentences $S_{-t}^{-h} \in D_{orig}$, carefully filtered to exclude any sentences containing similar head relations or tail entities. This ensures the answer information appears exclusively in the leading sentence of the evidence document. We strategically positioned the key sentences at the beginning of both documents to mitigate potential position bias effects, which we analyze in subsequent sections. An example of this setup is presented in Table 1 (Answer Impact).
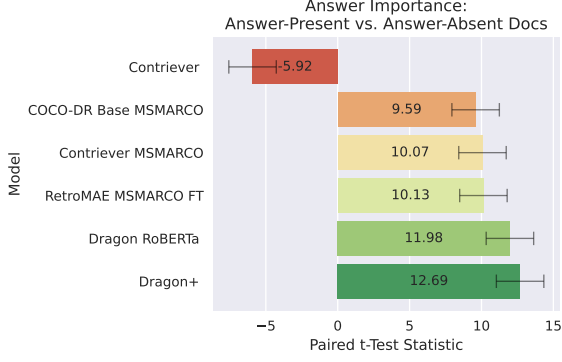
4

Figure 3: Paired t-test statistics comparing dot product similarity between the first sentence containing both head and tail (Answer) entities versus only the head entity, with 95% CI error bars. Higher values indicate recognition of the answer's importance.
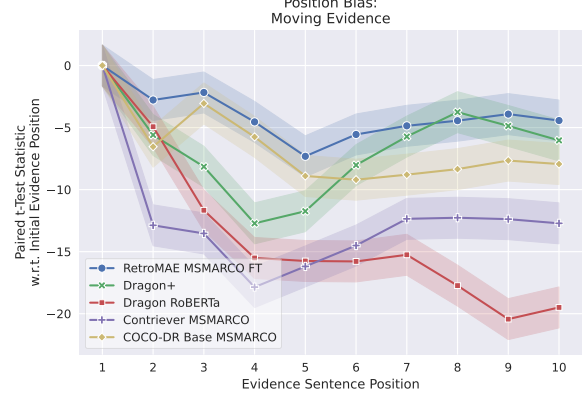

Figure 4: Paired t-test statistics comparing the effect of moving the evidence sentence position within the document to keeping it in the first position. Negative values indicate a bias towards the beginning of the document.

To quantify the models' ability to distinguish between these document types, we employ **Paired t-Test**[2] to analyze the difference in similarity scores. The t-test statistic (t) is calculated as:

$$t = \frac{\bar{d}}{SE(\bar{d})} = \frac{\text{Average Difference}}{\text{Standard Error}} \quad (2)$$

where $\bar{d} = mean\left(R(D_1) - R(D_2)\right)$ is the mean difference between paired observations[3], and $SE(\bar{d})$ is the standard error of these differences[4]. A positive t-statistic indicates higher scores for D_1 documents, while negative values suggest a preference for D_2 documents. In this scenario, positive values are desirable as they indicate the model prefers D_1 which contains the answer over D_2 which does not.

As shown in Figure 3, our analysis reveals variations across models. Dragon+ and Dragon-RoBERTa demonstrate superior tail recognition, achieving the highest positive t-statistics. In contrast, Contriever exhibits poor performance, yielding negative t-statistics that indicate a failure to properly distinguish answer-containing passages.

The vanilla Contriever's underwhelming performance can be attributed to its unsupervised training methodology, which differs from models trained on MS MARCO (Bajaj et al., 2018). While MS MARCO provides supervised training with explicit query-passage relevance labels, Contriever employs unsupervised contrastive learning. It generates positive pairs through data augmentation from document segments and derives negative examples implicitly via in-batch sampling from other texts. This training approach, while efficient for general text representation, appears insufficient for developing the fine-grained discrimination needed to understand query intent in retrieval tasks.

3.3.2 Position Bias

Position bias refers to the preference of retrieval models for information located in specific positions within a document, typically favoring content at the beginning over content appearing later. This bias is problematic as it may lead to the underrepresentation of relevant information that is positioned deeper within documents, thus reducing the overall retrieval quality and fairness.

Our analysis reveals a strong positional bias in dense retrievers, with models consistently prioritizing information at the beginning of documents. As shown in Figure 4, we conducted paired t-tests comparing retrieval scores when the evidence sentence is placed at different positions to scores when it is placed at the document's beginning $(R(D_i) - R(D_1))$.

$$
\begin{aligned}
\boldsymbol{D_1} &:= \boldsymbol{S_{ev}} + {}^1S_{-t}^{-h} + {}^2S_{-t}^{-h} + {}^3S_{-t}^{-h} + ... + {}^nS_{-t}^{-h} \\
\boldsymbol{D_2} &:= {}^1S_{-t}^{-h} + \boldsymbol{S_{ev}} + {}^2S_{-t}^{-h} + {}^3S_{-t}^{-h} + ... + {}^nS_{-t}^{-h} \quad (3) \\
\boldsymbol{D_3} &:= {}^1S_{-t}^{-h} + {}^2S_{-t}^{-h} + \boldsymbol{S_{ev}} + {}^3S_{-t}^{-h} + ... + {}^nS_{-t}^{-h}
\end{aligned}
$$

To ensure fairness, the examples were curated so that the remaining content was free of any evidence or head entity (S_{-t}^{-h}) like the last section. This design ensured that the evidence's position

---

[2]Using ttest_rel function of SciPy (Virtanen et al., 2020).

[3]R is the retriever's score

[4]$SE = \frac{\sigma}{\sqrt{n}}$

Q1	D1	Q2	D2	Contriever MSMARCO	Dragon+
long	long	long	short	+21.05	+21.04
		short	long	+22.04	+13.40
short	short	long	short	+4.62	+9.04
		short	long	+14.37	+16.62

Table 3: Paired t-test statistics (p-values < 0.05) comparing retrieval scores between exact name matches ($Q1$-$D1$) and variant name pairs ($Q2$-$D2$). Positive statistics indicate a preference for exact literal matches over semantically equivalent name variants (e.g., "US"-"US" over "US"-"United States"). (All models in Table A.7.)

was the sole factor under evaluation. The consistently negative t-statistics across models in Figure 4 confirm a strong bias favoring content at document beginnings.[5] This bias is most pronounced in Dragon-RoBERTa and Contriever-MSMARCO, which show the most negative t-statistics, indicating severe degradation in recognizing evidence further into the document. While Dragon+ and RetroMAE perform better, their negative t-statistics still confirm position bias in these models.

These findings align with recent research by Coelho et al. (2024), who demonstrated that positional biases emerge during the contrastive pretraining phase and worsened through fine-tuning on MS MARCO dataset with T5 (Raffel et al., 2020) and RepLLaMA (Ma et al., 2023) models. This can significantly impact retrieval performance when relevant information appears later in documents.

3.3.3 Literal Bias

Retrieval models should ideally recognize semantic equivalence across different surface forms of the same entity. For instance, a robust model should understand that "Gomes" and "Gomez" refer to the same person, or that "US" and "United States" represent the same entity. However, our analysis reveals that current models exhibit a strong bias toward exact literal matches rather than semantic matching.

In our dataset, each head entity can be represented by multiple alternative names. To investigate literal bias, we created different combinations of query and document by replacing all head entities with the shortest or longest name variants as illustrated in Table 1 (Literal Bias). For example, an entity might be represented as "NYC" (shortest) or "New York City" (longest), allowing us to test

how the model performs when matching different combinations of these representations.

Table 3 presents the paired t-test statistics comparing different combinations of name selections in queries and documents. The results consistently show positive statistics when Query 1 and Document 1 contain similar name representations. For our subsequent analysis of bias interplay, we specifically examine the comparison between two scenarios (Figure A.1): one where both query and document use the shortest name variant (short-short) versus cases where the query uses the short name but the document contains the long name variant (short-long). This corresponds to +14.37 and +16.62 in Table 3 for Contriever and Dragon+, respectively.[6]

3.3.4 Brevity Bias

Brevity bias refers to the tendency of retrievers to favor concise text, such as a single evidence sentence, over longer documents that include the same evidence alongside additional context. This bias is problematic because retrievers may favor a shorter, non-relevant document over a relevant one. We will discuss this potential hazard further in Section 3.5.

Here, we performed paired t-tests to compare the similarity scores of queries with two sets of documents: **(1)** *Single Evidence*, consisting of only the evidence sentence, and **(2)** *Evidence+Document*, consisting of the evidence sentence followed by the rest of the document. The examples are carefully selected to ensure the evidence sentence includes both the head and tail entity and the rest of the document contains no repetition of the head entity or additional evidence.

$$\begin{aligned} D_1 &\coloneqq S_{ev} \\ D_2 &\coloneqq S_{ev} + \sum_{S_{-t}^{-h} \in D_{orig}} S_{-t}^{-h} \end{aligned} \quad (4)$$

Figure 1 and A.4, illustrate the paired t-test statistics, where significant positive values indicate a strong bias toward brevity, as models assign higher scores to concise texts (D_1) than to longer ones with the same evidence (D_2). This behavior likely stems from the way dense passage retrievers compress document representations. Most retrievers use either a mean-pooling strategy or a [CLS] token-based method. Both methods struggle with integrating useful evidence into the representation

---

[5]Fig. 1 shows the impact of evidence placement (beginning vs. end), detailed in Appendix A, with an example in Table 1.

[6]We avoid long-long combinations to control for confounding effects, as they span multiple tokens and may introduce repetition bias due to token overlap
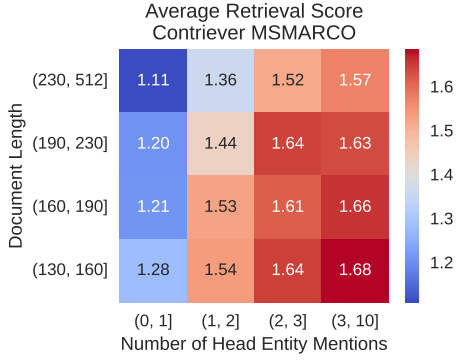
Figure 5: The average retrieval score of Contriever MS-MARCO increases with head entity repetitions but decreases with document length (all models in Figure A.5).

Model	Accuracy	Paired t-Test Statistic	p-value
Contriever	0.4%	-34.58	< 0.01
RetroMAE MSMARCO FT	0.4%	-41.49	< 0.01
Contriever MSMARCO	0.8%	-42.25	< 0.01
Dragon RoBERTa	0.8%	-36.53	< 0.01
Dragon+	1.2%	-40.94	< 0.01
COCO-DR Base MSMARCO	2.4%	-32.92	< 0.01

Table 4: The accuracy and paired t-test comparing a foil document (exploiting biases but lacking the answer) to a second document with evidence embedded in unrelated sentences. All retrieval models perform extremely poorly (<3% accuracy), highlighting their inability to distinguish biased distractors from genuine evidence.

when unrelated content is present, leading to a "pollution effect." As a result, the additional context in longer documents dilutes the importance of the evidence, causing retrievers to favor concise input.

3.3.5 Repetition Bias

Repetition bias refers to the tendency of retrieval models to prioritize documents or passages with repetitive content, particularly repeated mentions of head entities present in the query. This bias is problematic as it may skew retrieval results toward redundant or verbose documents, undermining the goal of surfacing concise and diverse information.

To analyze repetition bias, we conducted an experiment evaluating the average retrieval dot product score of the models for samples with varying document lengths and head entity repetitions (Figure 5 and A.5). A key concern is that longer documents naturally have a higher chance of lexical overlap with the query, as they may contain more repeated mentions of the head entity. This makes it difficult to disentangle the effects of document length from the number of entity repetitions. Therefore, we structure our analysis to separately examine these two factors. Our findings (Figure 5) reveal that the retrieval score increases with the number of head entity mentions, indicating a preference for documents with repeated entities. Conversely, the retrieval score decreases as document length increases, suggesting that longer documents are penalized despite potential relevance. Figure A.5 in the appendix generalizes these observations across all models. This experiment highlights the trade-off between repetition and document length, emphasizing the need for retrieval systems to balance these factors to mitigate bias.

We further explored this phenomenon through the results shown in Figures 1 and A.3. Here, we performed paired t-tests to compare the dot product similarity scores of queries with two sets of documents: (1) *More Heads*, comprising an evidence sentence and two sentences containing head mentions but no tails, and (2) *Fewer Heads*, comprising an evidence sentence and two sentences without head or tail mentions from the document (Table 1).

$$
\begin{aligned}
D_1 &:= S_{ev} + S_{-t}^{+h} + S_{-t}^{+h} \\
D_2 &:= S_{ev} + S_{-t}^{-h} + S_{-t}^{-h}
\end{aligned}
\tag{5}
$$

Positive paired t-test values indicate higher similarity for sentences with more head mentions (Figure A.3). The results strongly suggest that the model favors sentences with repeated heads, confirming the presence of repetition bias.

3.4 Interplay Between Bias Types

To understand how different biases interact and amplify retrieval model weaknesses, we conduct a systematic analysis using a controlled 250-sample dataset across all experiments. This consistent sample size ensures comparability of paired t-test statistics across bias types and provides a robust basis for evaluating their interplay.

As illustrated in Figure 1, the paired t-test results reveal that brevity bias, literal bias, and position bias are the most problematic for dense retrievers. In contrast, repetition bias, while still detrimental, exhibits a relatively lower impact, suggesting that models are slightly more robust against this type of bias. Answer importance demonstrates an acceptable distinction between evidence-containing and no-evidence documents. However, the scores are not as strong as one would expect from models designed for accurate answer retrieval, highlighting the need for further improvement in this area.

To further investigate the compounded effects of multiple biases, we conducted another experiment

MODEL	Poison Doc	Foil Doc	No Doc	Evidence Doc
gpt-4o-mini	32.0%	44.0%	52.0%	**88.0%**
gpt-4o	30.8%	62.8%	64.8%	**93.6%**

Table 5: RAG accuracy when using different document versions as references. The poisoned document, preferred by retrievers 100% of the time (Table A.4), results in worse performance than providing no document, highlighting the impact of retriever biases on RAG.

that combines several bias types into a single challenging setup. In this experiment, we created two document types. **1) Foil Document with Multiple Biases:** This document contains multiple biases, such as repetition and position biases. It includes two repeated mentions of the head entity in the opening sentence, followed by a sentence that mentions the head but not the tail (answer). So it does not include the evidence. **2) Evidence Document with Unrelated Content:** This document includes four unrelated sentences from another document, followed by the evidence sentence with both the head and tail entities. The document ends with the same four unrelated sentences. An example is shown in Table 1 (Foil vs. Evide.).[7]

$$
\begin{aligned}
\mathbf{D_1} &:= 2 \times h + S_{-t}^{+h} \\
\mathbf{D_2} &:= 4 \times \tilde{S}_{-t}^{-h} + \mathbf{S_{ev}} + 4 \times \tilde{S}_{-t}^{-h}
\end{aligned}
\tag{6}
$$

Table 4 presents the accuracy (proportion of times the model prefers D_2 over D_1), paired t-test statistics, and p-values. The results are striking: all models exhibit extremely poor performance, with accuracy dropping below 3%. The paired t-test statistics are highly negative across all models, indicating a consistent preference for foil documents over the correct evidence-containing ones. This outcome highlights the severity of bias interplay and its detrimental impact on model reliability. Furthermore, a sufficient number of biased documents can potentially cause the model to select all top-k documents from only biased results.

3.5 Impact on RAG

To assess the impact of the identified vulnerabilities on RAG systems, we use GPT-4o models (OpenAI et al., 2024) and provide them with different versions of the reference document for a given query. Additionally, we construct a poisoned document by modifying the foil document from §3.4, introducing a poisoned evidence sentence (Table A.5).

Specifically, we generate this sentence using GPT-4o by replacing the tail entity with a contextually plausible but entirely incorrect entity. This approach ensures that the poisoned document both exploits the previously discussed retrieval biases and contains an incorrect answer to the query.[8]

$$
\begin{aligned}
\mathbf{D_1} &:= 2 \times h + S_{-t}^{+h} + S_{+PoisonTail}^{+h} \\
\mathbf{D_2} &:= 4 \times \tilde{S}_{-t}^{-h} + \mathbf{S_{ev}} + 4 \times \tilde{S}_{-t}^{-h}
\end{aligned}
\tag{7}
$$

Table 5 reports the RAG accuracy,[9] showing that, as expected, providing the evidence document enables the LLM to achieve high accuracy. However, since retrievers prefer the foil document from §3.4, which lacks evidence, LLM performance drops to levels near[10] the no-document condition. This preference is concerning, as it allows biases to be exploited, making certain documents more likely to be retrieved despite embedding incorrect information. This is evident with the poisoned document, which degrades performance even worse than presenting no document by introducing false facts. In summary, **retriever biases can mislead RAG systems by providing poisoned or non-informative documents**, ultimately harming performance.

4 Conclusions

In this work, we introduced a comprehensive framework for analyzing biases in dense retrieval models. By leveraging a relation extraction dataset (Re-DocRED), we constructed a diverse set of controlled experiments to isolate and evaluate specific biases, including literal, position, repetition, and brevity biases as well as the answer's importance.

Our findings reveal that retrieval models often prioritize superficial patterns, such as exact string matches, repetitive content, or information positioned early in documents, over deeper semantic understanding and the existence of the answer. Moreover, when multiple biases combine, retriever performance deteriorates dramatically.

Furthermore, Our analysis shows that retriever biases can undermine RAG's reliability by favoring poisoned or non-informative documents over evidence-containing ones, leading to degraded performance of LLMs. These findings underscore the need for dense retrieval models that are robust to biases and capable of prioritizing semantic relevance.

---

[7]$\tilde{S}$ are sentences from an unrelated document

[8]Despite this, retrievers prefer the poisoned document over the evidence document in 100% of cases (Table A.4).

[9]Evaluated using GPT-4o. Prompts in Table A.6

[10]Slightly lower, as the model sometimes abstains by stating, "The document does not provide information."

Limitations

Quality of the Relation Extraction Dataset   Our framework relies on a relation extraction dataset, making both annotation accuracy (precision) and completeness (recall) critical. We use Re-DocRED, which addresses annotation issues in DocRED, but it may still contain imperfections that introduce minor noise into our experiments. To mitigate this, we employ statistical tests and report error margins and p-values to ensure the robustness of our findings.

Limitations of RAG Evaluation by LLMs   In our RAG experiments, we utilized GPT-4o models and carefully designed prompts (Table A.6) to poison documents, generate answers using RAG, and evaluate the results against gold-standard answers. Although GPT-4o is one of the most advanced models available, it is not infallible and may introduce some variance in the RAG results and evaluations. Nevertheless, we believe the observed trends and findings remain valid given the model's high performance and the consistency of our experimental setup.

References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. LitSearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15068–15083, Miami, Florida, USA. Association for Computational Linguistics.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, Toronto, Canada. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

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.

Nicholas Boucher, Luca Pajola, Ilia Shumailov, Ross Anderson, and Mauro Conti. 2023. Boosting big brother: Attacking search engines with encodings. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, RAID '23, page 700–713, New York, NY, USA. Association for Computing Machinery.

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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2021. MultiReQA: A cross-domain evaluation forRetrieval question answering models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 94–104, Kyiv, Ukraine. Association for Computational Linguistics.

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

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Yichun Yin, Hao Zhang, Yong Liu, Yasheng Wang, and Ruiming Tang. 2024. Coir: A comprehensive

9

benchmark for code information retrieval models. *Preprint*, arXiv:2407.02883.

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.

Zilong Lin, Zhengyi Li, Xiaojing Liao, XiaoFeng Wang, and Xiaozhong Liu. 2024. Mawseo: Adversarial wiki search poisoning for illicit online promotion. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 388–406.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, and Maarten de Rijke. 2024. Robust information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 3009–3012, New York, NY, USA. Association for Computing Machinery.

Quanyu Long, Yue Deng, LeiLei Gan, Wenya Wang, and Sinno Jialin Pan. 2024. Whispers in grammars: Injecting covert backdoors to compromise dense retrieval systems. *Preprint*, arXiv:2402.13532.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *arXiv:2310.08319*.

Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the behavior of neural IR models. *Transactions of the Association for Computational Linguistics*, 10:224–239.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2023. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, Toronto, Canada. Association for Computational Linguistics.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, Seattle, United States. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang,

Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun GoGineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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.

Benjamin Reichman and Larry Heck. 2024. Dense passage retrieval: Is it retrieving? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13540–13553, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei W Koh. 2024. Scaling retrieval-based language models with a trillion-token datastore. In *Advances in Neural Information Processing Systems*, volume 37, pages 91260–91299. Curran Associates, Inc.

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*

11

Processing, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Kota Usuha, Makoto P. Kato, and Sumio Fujita. 2024. Over-penalization for extra information in neural ir models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4096–4100, New York, NY, USA. Association for Computing Machinery.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Yumeng Wang, Lijun Lyu, and Avishek Anand. 2022. Bert rankers are brittle: A study using adversarial document perturbations. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '22, page 115–120, New York, NY, USA. Association for Computing Machinery.

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.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. COCO-DR: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13764–13775, Singapore. Association for Computational Linguistics.

A   Appendix

A.1   Models Downstream Performance

We evaluate several dense retrievers on the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019), comparing their performance using standard retrieval metrics: nDCG@10 and Recall@10[11]. The models differ in training objectives, datasets, and pooling mechanisms, offering a comprehensive view of their retrieval capabilities in our experimental setup. Table 2 (and A.2) summarizes the results.

Dragon RoBERTa and Dragon+ (Lin et al., 2023) demonstrate the highest performances due to diverse data augmentations and multiple supervision sources, which progressively enhance their generalization.[12]

COCO-DR (Yu et al., 2022) adopts continuous contrastive learning and implicit distributionally robust optimization (DRO) to address distribution shifts in dense retrieval tasks. It exhibits moderate performance, scoring lower than Dragon models.

Contriever (Izacard et al., 2021) uses unsupervised contrastive learning but performs poorly without fine-tuning (nDCG@10: 0.25). Fine-tuning on MSMARCO significantly improves its performance (nDCG@10: 0.50), underscoring the importance of fine-tuning for robust retrieval.

RetroMAE (Xiao et al., 2022), which introduces a retrieval-oriented pre-training paradigm based on Masked Auto-Encoder (MAE), featuring innovative designs like asymmetric masking, achieves slightly lower performance (nDCG@10: 0.48) compared to fine-tuned Contriever.

The models also differ in their pooling mechanisms. Contriever uses average pooling, where token representations are averaged to form a dense vector for retrieval. In contrast, the other models use CLS pooling, where the representation of the [CLS] token is taken as the sentence embedding.

In summary, Dragon models lead in performance, and the significant improvement of fine-tuned Contriever over its unsupervised counterpart highlights the importance of supervision and task-specific adaptation in dense retrieval.

A.2   Position Bias: First vs. Last

Further evidence is provided in Figure A.2, where we compared two document variants:

1. Beginning-Evidence Document D_1: The evidence sentence is positioned at the start of the document. **2. End-Evidence Document D_2:** The same evidence sentence is positioned at the end of the document.

$$D_1 := S_{ev} + \sum_{S_{-t}^{-h} \in D_{orig}} S_{-t}^{-h}$$

$$D_2 := \sum_{S_{-t}^{-h} \in D_{orig}} S_{-t}^{-h} + S_{ev} \tag{8}$$

An example of the document pairs (Position Bias) is shown in Table 1. The resulting t-statistics (Figure 1 and A.2), where higher positive values indicate a stronger preference for evidence at the beginning (D_1) over the end (D_2), provide another clear metric of positional bias. These results serve as a foundation for our subsequent analysis in the interplay between biases section.

Model	Citation
facebook/dragon-plus-query-encoder facebook/dragon-plus-context-encoder	Lin et al. (2023)
facebook/dragon-roberta-query-encoder facebook/dragon-roberta-context-encoder	Lin et al. (2023)
facebook/contriever-msmarco	Izacard et al. (2021)
facebook/contriever	Izacard et al. (2021)
OpenMatch/cocodr-base-msmarco	Yu et al. (2022)
Shitao/RetroMAE_MSMARCO_finetune	Xiao et al. (2022)
gpt-4o-mini-2024-07-18	OpenAI et al. (2024)
gpt-4o-2024-08-06	OpenAI et al. (2024)

Table A.1: The details of the models we used in this work.

Model	Pooling	nDCG@10	Recall@10
Dragon+	cls	**0.55**	**0.63**
Dragon RoBERTa	cls	0.53	0.59
Contriever MSMARCO	avg	0.52	0.59
Contriever	avg	0.50	0.59
RetroMAE MSMARCO FT	cls	0.49	0.55
COCO-DR Base MSMARCO	cls	0.48	0.53

Table A.2: Models' performance on our refined redo-cred dataset with 7170 queries and 105925 corpus size.

Issue	Count	Percentage
Long Document	33	55%
Missing Answer	19	32%
Literal Bias	11	18%
Repetition	6	10%
Numbers	2	3%
Position Bias	2	3%

Table A.3: Preliminary findings from our annotation of 60 retrieval errors based on DecompX

---

[11]Using BEIR framework (Thakur et al., 2021)

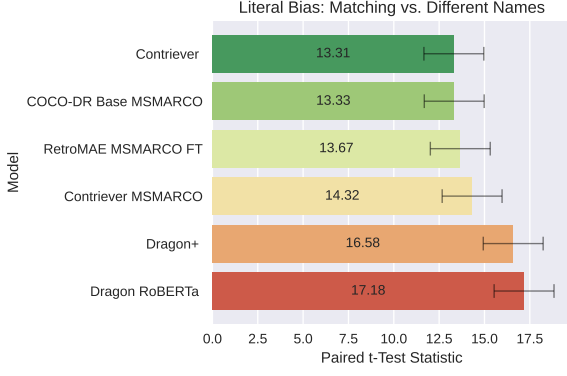[12]Dragon RoBERTa is initialized from RoBERTa and Dragon+ from RetroMAE

Figure A.1: Paired t-test statistics comparing retrieval scores between two scenarios: (1) when both query and document use the shortest name variant, and (2) when the query uses the short name but the document contains the long name variant of the same entity. Positive statistics indicate that models favor exact string matches over semantic matching of equivalent entity names.
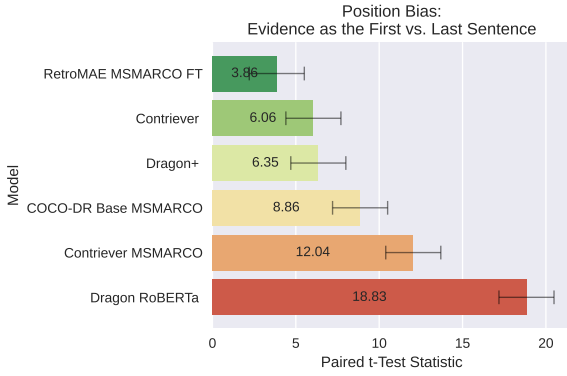

Figure A.2: Paired t-test statistics comparing document scores based on the position of the evidence sentence (beginning vs. end). Higher positive values reflect a preference for evidence at the beginning, indicating positional bias.
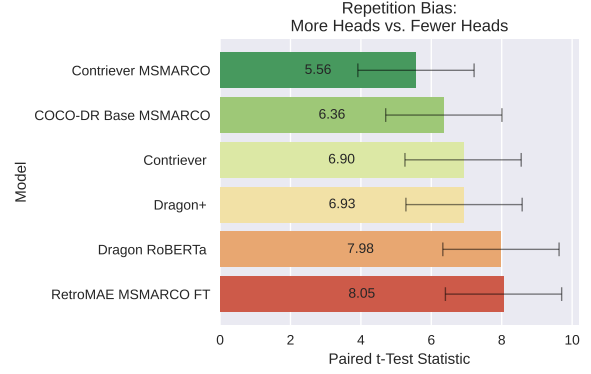

Figure A.3: Paired t-test statistics comparing the dot product similarity of queries with two sets of sentences: (1) *More Heads*, consisting of evidence and two sentences with head mentions but no tails, and (2) *Fewer Heads*, consisting of evidence and two sentences without head or tail mentions. Positive values indicate higher similarity for sentences with more heads.

Model	Accuracy	Paired t-Test Statistic	p-value
Dragon+	0.0%	-55.16	< 0.01
Dragon RoBERTa	0.0%	-49.17	< 0.01
Contriever MSMARCO	0.0%	-46.96	< 0.01
COCO-DR Base MSMARCO	0.0%	-40.19	< 0.01
RetroMAE MSMARCO FT	0.0%	-48.10	< 0.01
Contriever	1.2%	-33.60	< 0.01

Table A.4: The accuracy, paired t-test statistics, and p-values comparing a **poison document**, designed to exploit biases and having a wrong answer (tail), against a second document containing the evidence sentence embedded in the middle of eight unrelated sentences from a different document. All retrieval models perform extremely poorly (less than 2% accuracy).
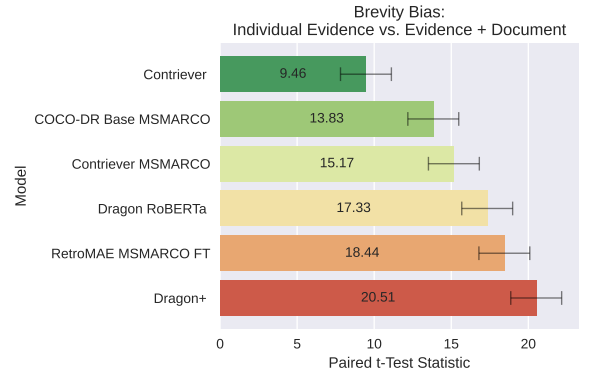

Figure A.4: Paired t-test statistics comparing scores for documents containing only the evidence sentence versus those containing the evidence plus the full document. Higher positive values indicate a stronger model bias toward brevity.

14

	Method 1 (Higher Query Document Similarity Score)	Method 2 (Lower Query Document Similarity Score)
Foil vs. Evide.	**Query:** Who is the publisher of *Assassin 's Creed Unity*? **Document:** " *Assassin 's Creed Unity* " " *Assassin 's Creed Unity* " *Assassin 's Creed Unity* received mixed reviews upon its release .	**Query:** Who is the publisher of *Assassin 's Creed Unity*? **Document:** Isa is a town and Local Government Area in the state of Sokoto in Nigeria . It shares borders with ..... *Assassin 's Creed Unity* is an action - adventure video game developed by Ubisoft Montreal and published by *Ubisoft*. Isa is a town and Local Government Area in the state of Sokoto in Nigeria . It shares borders with .....
Poison vs. Evide.	**Query:** Who is the publisher of *Assassin 's Creed Unity*? **Document:** " *Assassin 's Creed Unity* " " *Assassin 's Creed Unity* " *Assassin 's Creed Unity* received mixed reviews upon its release . *Assassin 's Creed Unity* is an action - adventure video game developed by Electronic Arts Montreal and published by *Electronic Arts*	**Query:** Who is the publisher of *Assassin 's Creed Unity*? **Document:** Isa is a town and Local Government Area in the state of Sokoto in Nigeria . It shares borders with ..... *Assassin 's Creed Unity* is an action - adventure video game developed by Ubisoft Montreal and published by *Ubisoft*. Isa is a town and Local Government Area in the state of Sokoto in Nigeria . It shares borders with .....

Table A.5: Examples from our framework for poison document and evidence document highlighting Evidence, *Head Entity*, *Tail Entity* and *Poison* replacing true tail entity. In all cases, retrieval models favor Method 1 over Method 2, assigning higher retrieval scores accordingly.

Prompt Utility	Prompt
Poisoning	In the sentence: '{evidence}', replace the entity '{tail}' with a different entity that makes sense in context but is completely different. Output only the replacement entity. replacement entity:
RAG	Answer the question based on the given document. Only give me the complete answer and do not output any other words. The following is the given document. Document: {doc} Question: {query} Answer:
RAG for No Doc	Answer the question. Only give me the answer and do not output any other words. Question: {query} Answer:
Evaluation	Query: {query} Evidence: {evidence_sentence} Gold Answer: {gold_answer} Model Answer: {model_answer} Does the Model Answer contain or imply the Gold Answer based on the evidence? YES or NO :

Table A.6: The prompts utilized for RAG.

Query Name 1	Doc Name 1	Query Name 2	Doc Name 2	COCO-DR Base MSMARCO	RetroMAE MSMARCO FT	Contriever	Contriever MSMARCO	Dragon+	Dragon RoBERTa
long	**long**	**long**	**short**	20.67	21.92	19.22	21.05	21.03	21.64
		short	**long**	23.41	23.53	21.46	22.01	13.40	7.55
		short	**short**	18.43	19.60	16.41	17.35	4.99	1.75
short	**short**	**long**	**short**	2.19	3.86	2.32	4.65	9.05	5.57
		short	**long**	13.33	13.67	13.31	14.32	16.58	17.18

Table A.7: Paired t-test statistics comparing retrieval scores between exact name matches (Q1-D1) and variant name pairs (Q2-D2). Positive statistics indicate model preference for exact literal matches over semantically equivalent name variants (e.g., preferring "US"-"US" over "US"-"United States").
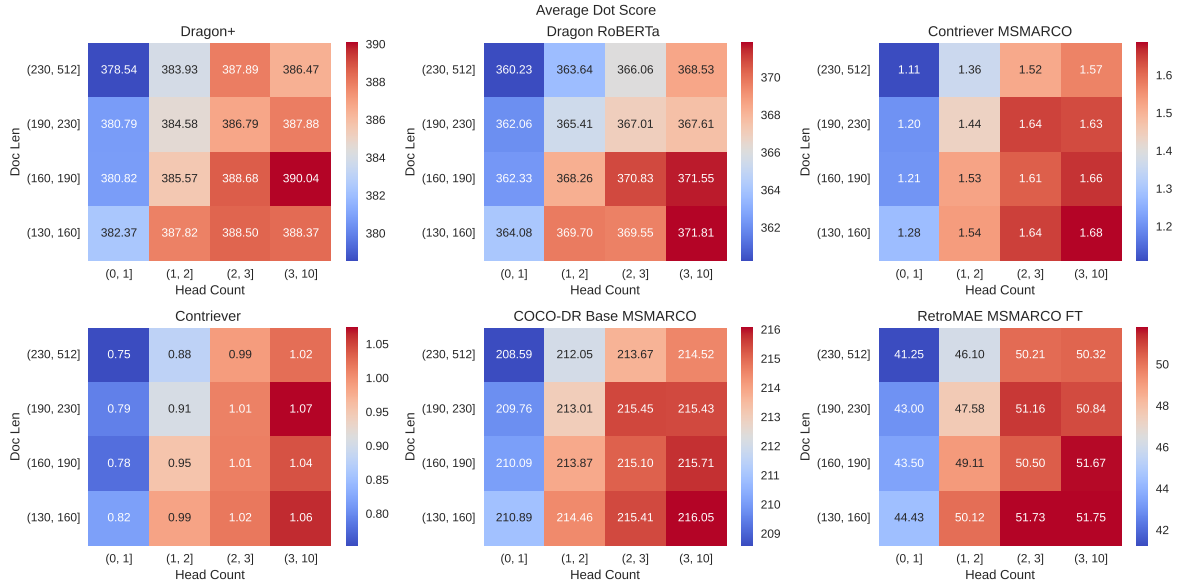
Figure A.5: The average retrieval dot product score for samples with different document lengths and head entity repetitions. (See Figure A.6 for the number of examples in each bin)
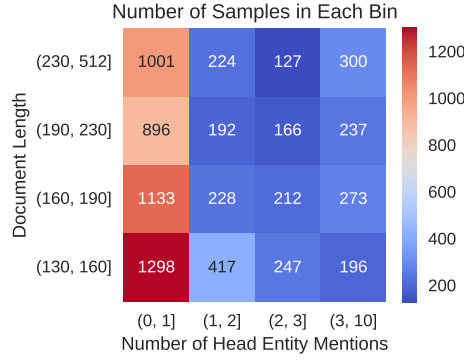

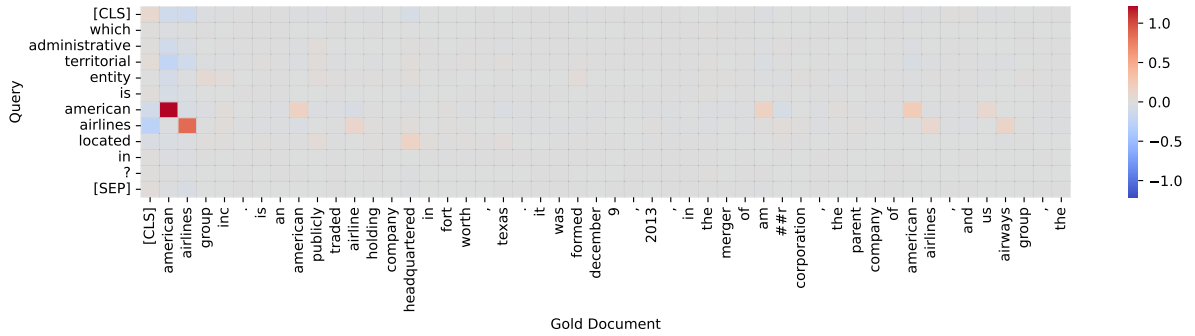Figure A.6: The number of samples in each bin of Figures A.5 and 5.

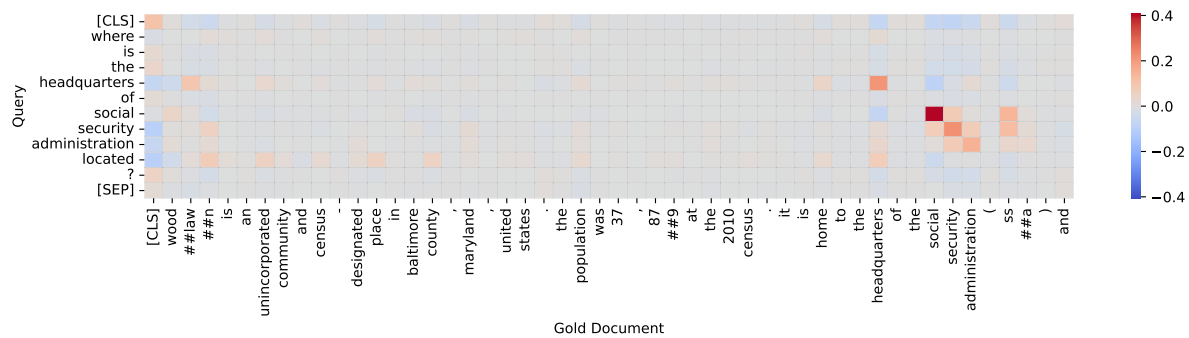Figure A.7: Visualization of token-wise effects on retriever scores using DecompX.

16

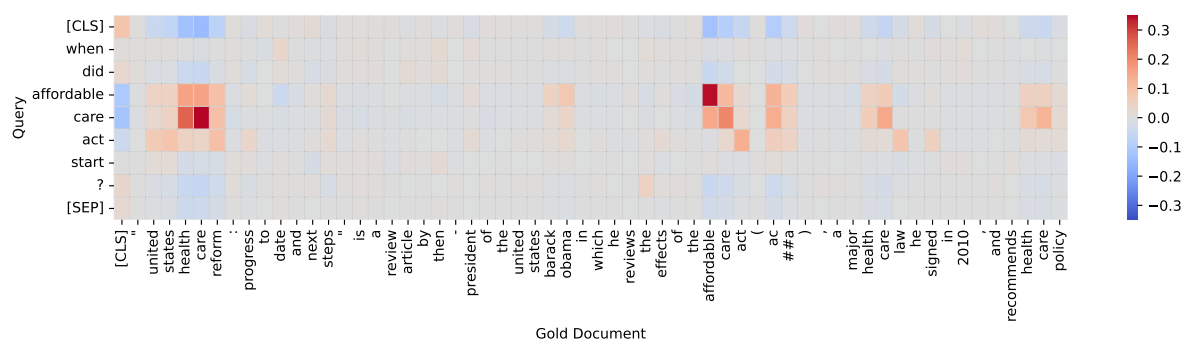Figure A.8: Visualization of token-wise effects on retriever scores using DecompX.

Figure A.9: Visualization of token-wise effects on retriever scores using DecompX.