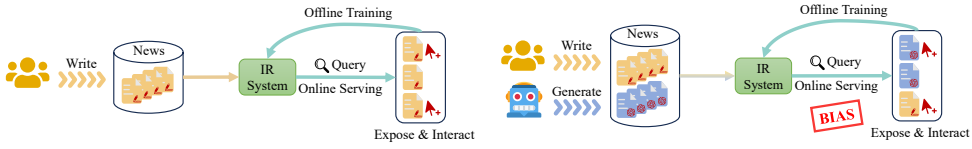
NEURAL RETRIEVERS ARE BIASED TOWARDS LLM-GENERATED CONTENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, the emergence of large language models (LLMs) has revolutionized the paradigm of information retrieval (IR) applications, especially in web search, by generating vast amounts of human-like texts on the Internet. As a result, IR systems in the LLM era are facing a new challenge: the indexed documents are now not only written by human beings but also automatically generated by the LLMs. How these LLM-generated documents influence the IR systems is a pressing and still unexplored question. In this work, we conduct a quantitative evaluation of IR models in scenarios where both human-written and LLM-generated texts are involved. Surprisingly, our findings indicate that neural retrieval models tend to rank LLM-generated documents higher. We refer to this category of biases in neural retrievers towards the LLM-generated text as the **source bias**. Moreover, we discover that this bias is not confined to the first-stage neural retrievers, but extends to the second-stage neural re-rankers. Then, in-depth analyses from the perspective of text compression indicate that LLM-generated texts exhibit more focused semantics with less noise, making them easier for neural retrievers to semantic match. To mitigate the source bias, we also propose a plug-and-play debiased constraint for the optimization objective, and experimental results show its effectiveness. Finally, we discuss the potential severe concerns stemming from the observed source bias and hope our findings can serve as a critical wake-up call to the IR community and beyond. To facilitate future explorations of IR in the LLM era, the constructed two new benchmarks and codes are available in the link `https://anonymous.4open.science/r/Source-Bias-B44E`.

1 INTRODUCTION



(a) IR in the Pre-LLM Era (b) IR in the LLM Era

Figure 1: The overview evolution of IR paradigm from the Pre-LLM era to the LLM era.

With the advent of large language models (LLMs), exemplified by ChatGPT, the field of artificial intelligence generated content (AIGC) has surged to new heights of prosperity Cao et al. (2023); Wu et al. (2023). LLMs have demonstrated their remarkable capabilities in automatically generating human-like text at scale, resulting in the Internet being inundated with an unprecedented volume of AIGC content Wei et al. (2022); Spitale et al. (2023). This influx of LLM-generated content has fundamentally reshaped the digital ecosystem, challenging conventional paradigms of content creation, dissemination, and information access on the Internet Ai et al. (2023); Zhu et al. (2023).

Meanwhile, information retrieval (IR) systems have become indispensable for navigating and accessing the Internet's vast information landscape Singhal et al. (2001); Manning (2009); Zhu et al. (2023). As illustrated in Figure 1, in the era preceding the widespread emergence of LLMs, IR systems focused on retrieving documents solely from the human-written corpus in response to users'

queries Liu et al. (2009); Li (2022); Xu & Li (2007). However, the proliferation of AIGC driven by LLMs has expanded the corpus of IR systems to include both human-written and LLM-generated texts. Consequently, this paradigm shift raises a fundamental research question: **What is the impact of the proliferation of generated content on IR systems?** We aim to explore whether existing retrieval models tend to prioritize LLM-generated text over human-written text, even when both convey similar semantic information. If this holds, LLMs may dominate information access, particularly as their generated content is rapidly growing on the Internet Hanley & Durumeric (2023).

To approach the fundamental research question, we decompose it into four specific research questions. The first question is **RQ1: How to construct an environment to evaluate IR models in the LLM era?** Given the lack of public retrieval benchmarks encompassing both human-written and LLM-generated texts, we propose an innovative and pragmatic method to create such a realistic evaluation environment without the need of costly human annotation. Specifically, we leverage the original human-written texts as the instruction conditions to prompt LLMs to generate rewritten text copies while preserving the same semantic meaning. In this way, we can confidently assign the relevant labels to LLM-generated data. Extensive empirical analysis validates the quality of our constructed environment, demonstrating its effectiveness in mirroring real-world IR scenarios in the LLM era. As a result, we introduce two new benchmarks, SciFact+AIGC and NQ320K+AIGC, tailored for IR research in the LLM era.

With the constructed environment, we further explore **RQ2: Are retrieval models biased towards LLM-generated texts?** We conduct comprehensive experiments with various representative retrieval models, ranging from traditional lexical models to modern neural models based on pretrained language models (PLMs) Guo et al. (2020); Zhao et al. (2023); Yates et al. (2021); Guo et al. (2022). Surprisingly, we uncover that neural retrievers are biased towards LLM-generated texts, i.e., tend to rank LLM-generated texts in higher positions. We refer to this as **source bias**, as the neural retrievers favor content from specific sources (i.e., LLM-generated content). Further experiments indicate that the source bias not only extends to the second-stage neural re-rankers from the first-stage retrieval but also manifests more severely. These findings corroborate the prevalence of source bias in neural retrieval models.

Then, what we are curious about is **RQ3: Why are neural retrieval models biased towards LLM-generated texts?** Inspired by the recent studies positing LLMs as lossless compressors Delétang et al. (2023), we analyze the cause of source bias from the viewpoint of text compression. Our analysis of singular values Klema & Laub (1980) in different corpora reveals that LLM-generated texts exhibit more focused semantics with minimal noise, enhancing their suitability for semantic matching. Furthermore, our in-depth perplexity analysis shows that LLM-generated texts consistently achieve lower perplexity scores, which indicates a higher degree of comprehensibility and confidence from the PLM's perspective. These observations collectively suggest that LLM-generated texts are more readily understandable to semantic match with PLM-based neural retrievers, thereby resulting in source bias.

Finally, we try to answer **RQ4: How to mitigate source bias in neural retrieval models?** To tackle this, we propose an intuitive yet effective debiased constraint. This constraint is designed to penalize biased samples during the optimization process, thereby shifting the focus of retrieval models from exploiting inherent shortcuts to emphasizing semantic relevance. Besides, our debiased constraint is model-agnostic and can be plugged and played to the ranking optimization objectives of various neural retrieval models. Furthermore, it offers the capability to control the degree of bias removal, offering the flexibility to balance the treatment between the two sources of content based on specific requirements and environmental considerations.

Last but not least, we discuss the potential emerging concerns stemming from source bias, highlighting the risk of human-written content being gradually inaccessible, especially due to the rapidly increasing LLM-generated content on the Internet Hanley & Durumeric (2023); Bengio et al. (2023). Furthermore, source bias could be maliciously exploited to manipulate algorithms and potentially amplify the spread of misinformation, posing a threat to online security. In light of these pressing issues, we hope that our findings serve as a resounding wake-up call to all stakeholders involved in IR systems and beyond.

In summary, the contributions of this paper are as follows:

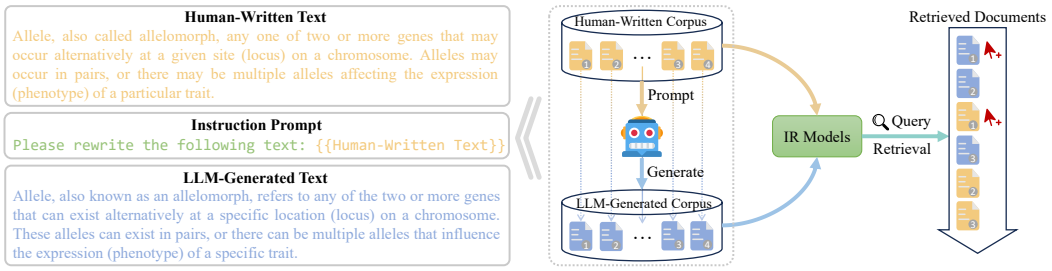


Figure 2: The overall paradigm of the proposed evaluation framework for IR in the LLM era.

(1) We introduce a more realistic paradigm of IR systems considering the growing prosperity of AIGC, where the retrieval corpus consists of both human-written and LLM-generated texts. We then uncover a new inherent bias in both neural retrieves and re-rankers preferring LLM-generated content, termed as source bias.

(2) We provide an in-depth analysis and insights of source bias from a text compression perspective, which indicates that LLM-generated texts maintain more focused semantics with minimal noise and are more readily comprehensible for neural retrievers.

(3) We propose a debiased constraint to penalize the biased samples during optimization, and experimental results demonstrate its effectiveness in mitigating source bias in different degrees.

(4) We also provide two new benchmarks, SciFact+AIGC and NQ320K+AIGC, which contain both high-quality human-written and various LLM-generated corpus and corresponding relevant labels. We believe these two benchmarks can serve as valuable resources for facilitating future research.

2 RQ1: Environment Construction

With the increasing usage of LLMs in generating texts (e.g., paraphrasing or rewriting), the corpus of IR systems includes both human-written and LLM-generated texts nowadays. Constructing an IR dataset in the LLM era typically involves two steps: collecting both human-written and LLM-generated corpora and then employing human evaluators to annotate relevance labels for each query-document pair. Given that LLM-generated content is currently unidentifiable Sadasivan et al. (2023) and the significant cost of human annotation, we introduce a natural and practical framework for quantitatively evaluating retrieval models in the LLM era, as shown in Figure 2.

To better align with real-world scenarios, the evaluation environments should meet the following three essential criteria. **Firstly**, it is imperative to distinguish between human-written and LLM-generated texts within the corpus. **Secondly**, we need access to relevance labels for LLM-generated data in response to queries. **Thirdly**, each human-written text should better have a corresponding LLM-generated counterpart with the same semantics, ensuring the most effective and fair evaluation.

2.1 Notation

Formally, in the Pre-LLM era, given a query $q \in \mathcal{Q}$ where $\mathcal{Q}$ is the set of all queries, the traditional IR system aims to retrieve a list of top-K relevant documents $\{d^{(1)}, d^{(2)}, \ldots, d^{(K)}\}$ from a corpus $\mathcal{C}^H = \{d_1^H, d_2^H, \ldots d_N^H\}$ which consists of N human-written documents. However, in the era of LLMs, there is also LLM-generated text in the corpus. To evaluate the IR models in the LLM era, we also create an additional corpus $\mathcal{C}^G = \{d_1^G, d_2^G, \ldots, d_N^G\}$ where each document is generated by a LLM, e.g., d_1^G can be created by ChatGPT by constructing a prompt that asks ChatGPT to rewrite d_1^H while preserving its original semantics. Consequently, given a query q, the objective of a retriever in the LLM era is to return the top-K relevant documents from the mixed corpus $\mathcal{C} = \mathcal{C}^H \bigcup \mathcal{C}^G$.

2.2 Constructing IR Datasets in the LLM Era

In this section, we prompt LLMs to rewrite human-written corpus to build two new standard retrieval datasets: SciFact+AIGC and NQ320K+AIGC. These two new datasets can serve as valuable resources to facilitate future research of IR in the LLM era.

Table 1: Statistics of the constructed two datasets. Avg. Doc / Query means the average number of relevant documents per query.

Dataset	# Test Queries	# Avg. Query Length	Human-Written Corpus			Llama2-Generated		
			# Corpus	Avg. Doc Length	Avg. Doc / Query	# Corpus	Avg. Doc Length	Avg. Doc / Query
SciFact+AIGC	300	12.38	5,183	201.81	1.1	5,183	192.66	1.1
NQ320K+AIGC	7,830	9.24	109,739	199.79	1.0	109,739	174.49	1.0

Human-Written Corpus. We first choose two widely used retrieval datasets written by humans in the Pre-LLM era as the seed data: SciFact and NQ320K. SciFact[1] Wadden et al. (2020) dataset aims to retrieve evidence from the research literature containing scientific paper abstracts for fact-checking. NQ320K[2] Kwiatkowski et al. (2019) is based on the Natural Questions (NQ) dataset from Google, where the documents are gathered from Wikipedia pages, and the queries are natural language questions. Following the practice in BEIR benchmark Thakur et al. (2021), we process these two datasets in a standard format: corpus $\mathcal{C}^H$, queries $\mathcal{Q}$, and labels $\mathcal{R}^H = \{(q_m, d_m^H, r_m)\}_{m=1}^M$, where M is the number of labeled query-document pairs in the corpus.

LLM-Generated Corpus. For the LLM-generated corpus, we repurpose the original human-written corpus as our seed data and instruct LLMs to rewrite each given text from the human-written corpus. As the written text generated by LLM carries almost the same semantic information as the original human-written text, we can assign the same relevance labels to new query-document pairs as those assigned to the original query-document pairs.

Our instruction is straightforward: "*Please rewrite the following text: {{human-written text}}*", as illustrated in the left part of Figure 2. This straightforward instruction enables LLMs to generate text without too many constraints while maintaining semantic equivalence to the original human-written text. Specifically, we choose Llama2 Touvron et al. (2023) to rewrite each seed human-written corpus, as Llama2 is the most widely-used open-sourced LLM.

As a result, we can obtain two corresponding LLM-generated corpora with SciFact and NQ320K as seed data. After that, we extend the original labels of query and human-written text $\mathcal{R}^H = \{(q_m, d_m^H, r_m)\}_{m=1}^M$ to get the corresponding label of LLM-generated text $\mathcal{R}^G = \{(q_m, d_m^G, r_m)\}_{m=1}^M$. We will validate the quality of the datasets in the following section. Combining each original human-written corpus $\mathcal{C}^H$ with its corresponding LLM-generated corpus $\mathcal{C}^G$, original queries $\mathcal{Q}$, and labels $\mathcal{R}^H \bigcup \mathcal{R}^G$, we can create two new datasets, denoted as SciFact+AIGC and NQ320K+AIGC. Table 1 summarize the statistics of the proposed two datasets.

2.3 Statistics and Quality Validation of Datasets

For the LLM-generated texts, a pivotal consideration is whether they faithfully preserve the underlying semantics of the corresponding human-written corpus. If they indeed do so, we then can confidently assign them the same relevance labels as the labels of their corresponding original human-written texts given each query.

Semantic-based Statistics and Analysis. We first leverage the OpenAI embedding model[3] to acquire semantic embeddings for both the LLM-generated and human-written texts. We then calculate the cosine similarity of semantic embeddings between the LLM-generated text and their corresponding human-written counterparts. The results, as shown in Figure 3, also indicate a high degree of similarity, with most values exceeding 0.95, affirming the faithful preservation of semantics in LLM-generated text. Hence, for each query-document pair (q, d^G), we can confidently assign the relevant label r to be the same as that of (q, d^H).

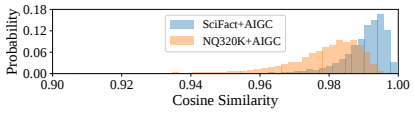


Figure 3: Distribution of cosine similarity of semantic embedding between LLM-generated and human-written corpora.

[1] https://allenai.org/data/scifact

[2] https://ai.google.com/research/NaturalQuestions

[3] text-embedding-ada-002:https://platform.openai.com/docs/guides/embeddings

Table 2: Performance comparison of retrieval models on the sole human-written or Llama2-generated corpus on SciFact+AIGC and NQ320K+AIGC datasets. For brevity, we omit the percent sign '%' of ranking metrics in subsequent tables and figures.

Model Type	Model	Corpus	SciFact+AIGC						NQ320K+AIGC					
			NDCG@1	NDCG@3	NDCG@5	MAP@1	MAP@3	MAP@5	NDCG@1	NDCG@3	NDCG@5	MAP@1	MAP@3	MAP@5
Lexical	TF-IDF	Human-Written	42.0	49.5	52.7	40.7	47.1	49.0	12.2	15.8	16.8	12.2	14.9	15.5
		LLM-Generated	43.0	49.8	52.6	40.8	47.5	49.2	9.4	12.6	13.9	9.4	11.8	12.5
	BM25	Human-Written	46.0	54.2	56.3	43.8	51.5	52.8	12.9	16.3	17.6	12.9	15.5	16.2
		LLM-Generated	46.3	53.6	55.3	44.1	51.1	52.2	11.9	15.3	16.5	11.9	14.5	15.1
Neural	ANCE	Human-Written	38.7	44.3	46.5	36.3	41.9	43.3	50.6	60.0	62.2	50.6	57.7	58.9
		LLM-Generated	41.0	46.0	48.2	37.8	43.5	45.0	49.3	58.8	61.2	49.3	56.5	57.8
	BERM	Human-Written	37.0	42.1	44.2	34.7	39.7	41.3	49.2	58.3	60.4	49.2	56.1	57.3
		LLM-Generated	40.7	44.5	46.2	37.7	42.3	43.5	48.4	57.5	59.8	48.4	55.3	56.5
	TAS-B	Human-Written	52.7	58.1	60.2	49.9	55.6	57.2	53.4	63.0	65.4	53.4	60.7	62.0
		LLM-Generated	50.7	57.0	58.9	48.0	54.6	55.9	51.9	62.3	64.7	51.9	59.8	61.1
	Contriever	Human-Written	54.0	61.8	63.2	51.4	58.9	60.0	58.2	68.4	70.3	58.2	65.9	67.0
		LLM-Generated	55.7	62.0	64.8	52.9	59.5	61.5	57.1	67.5	69.8	57.1	64.9	66.2

Retrieval Performance Evaluation. To further validate the accuracy of the relevance label assignments, we conduct an evaluation of retrieval models on the human-written corpus and the LLM-generated corpus, respectively. The following representative retrieval models are adopted in the experiments: (1) Lexical Retrieval Models: **TF-IDF** Sparck Jones (1972) and **BM25** Robertson et al. (2009) and (2) Neural Retrieval Models: **ANCE** Xiong et al. (2020), **BERM** Xu et al. (2023), **TAS-B** Hofstätter et al. (2021), **Contriever** Izacard et al. (2021).

The results on each sole source corpus on the proposed two new benchmarks are presented in Table 2. It is evident that all retrieval models exhibit no significant performance discrepancies in terms of various ranking metrics between the human-written and LLM-generated corpora across all datasets. This observation reinforces the confidence in the quality of our newly constructed datasets. Note that in our constructed datasets, LLMs were instructed to rewrite human-written texts based solely on the original human-written text, without any query-related input, thereby preventing the additional query-specific information during rewriting.

3 RQ2: Uncovering Source Bias

In this section, we conduct extensive experiments on the constructed datasets to explore the source bias from various aspects. With the constructed simulated environment, we first introduce the evaluation metrics to quantify the severity of source bias. We then conduct experiments with different retrieval models on both the first-stage retrieval and the second-stage re-ranking.

3.1 Evaluation Metrics for Source Bias

To quantitatively explore source bias, we calculate ranking metrics, targeting separately either human-written or LLM-generated corpus. Specifically, for each query, an IR model produces a ranking list that comprises documents from mixed corpora. We then calculate top-K Normalized Discounted Cumulative Gain (NDCG@K) and Mean Average Precision (MAP@K), for $K \in \{1, 3, 5\}$, independently for each corpus source. When assessing one corpus (e.g., human-written), documents from the other (e.g., LLM-generated) are treated as non-relevant, though the original mixed-source ranking order is maintained. This approach allows us to independently assess the performance of IR models on each corpus source.

To better normalize the difference among different benchmarks, we also introduce the relative percentage difference as follows:

$$\text{Relative } \Delta = \frac{\texttt{Metric}_{\text{Human-written}} - \texttt{Metric}_{\text{LLM-generated}}}{\frac{1}{2}\left(\texttt{Metric}_{\text{Human-written}} + \texttt{Metric}_{\text{LLM-generated}}\right)} \times 100\%,$$

where the $\texttt{Metric}$ can be NDCG@K and MAP@K. Note that Relative $\Delta > 0$ means retrieval models rank human-written texts higher, and Relative $\Delta < 0$ indicates LLM-generated texts are ranked higher. The greater the absolute value of Relative Δ, the greater the ranking performance difference on two corpora.

3.2 Bias in Neural Retrieval Models

In our assessment of various retrieval models on SciFact+AIGC and NQ320K+AIGC datasets, we observe distinct behaviors when evaluating against human-written and LLM-generated corpora, as reported in Table 3. Our key findings are as follows:

Table 3: Performance comparison of retrieval models for mixed human-written and LLM-generated corpora on SciFact+AIGC and NQ320K+AIGC dataset. The numbers indicate that retrieval models rank human-written documents in higher positions than LLM-generated documents (i.e., Relative $\Delta > 0\%$). Conversely, the numbers mean retrieval models rank LLM-generated documents in higher positions than human-written documents (i.e., Relative $\Delta \leq 0\%$). The intensity of the color reflects the extent of the difference. In the subsequent tables, we will continue with this color scheme.

Model Type	Model	Target Corpus	SciFact+AIGC						NQ320K+AIGC					
			NDCG@1	NDCG@3	NDCG@5	MAP@1	MAP@3	MAP@5	NDCG@1	NDCG@3	NDCG@5	MAP@1	MAP@3	MAP@5
Lexical	TF-IDF	Human-Written	22.0	36.9	39.7	21.2	33.0	34.7	7.1	11.0	12.3	7.1	10.0	10.8
		LLM-Generated	17.0	33.8	37.2	16.2	29.5	31.5	3.4	8.1	9.4	3.4	7.0	7.7
		Relative Δ	25.6	8.8	6.5	26.7	11.2	9.7	70.5	30.4	26.7	70.5	35.3	33.5
	BM25	Human-Written	26.7	40.3	44.4	25.7	36.7	39.1	7.2	11.6	12.9	7.2	10.6	11.3
		LLM-Generated	21.0	38.8	41.5	19.6	34.3	35.9	6.1	10.9	11.9	6.1	9.7	10.3
		Relative Δ	23.9	3.8	6.8	26.9	6.8	8.5	16.5	6.2	8.1	16.5	8.9	9.3
Neural	ANCE	Human-Written	15.3	30.1	32.7	14.2	26.2	27.7	22.2	41.2	44.6	22.2	36.9	38.8
		LLM-Generated	24.7	35.8	37.7	23.3	32.4	33.6	29.1	45.9	49.0	29.1	42.0	43.8
		Relative Δ	-47.0	-17.3	-14.2	-48.5	-21.2	-19.2	-26.9	-10.8	-9.4	-26.9	-12.9	-12.1
	BERM	Human-Written	16.3	30.2	31.8	15.7	26.5	27.5	18.6	37.5	40.7	18.6	33.1	34.9
		LLM-Generated	23.7	34.1	36.4	21.7	30.8	32.2	31.6	47.0	50.0	31.6	43.5	45.1
		Relative Δ	-37.0	-12.1	-13.5	-32.1	-15.0	-15.7	-51.8	-22.5	-20.5	-51.8	-27.2	-25.5
	TAS-B	Human-Written	20.0	40.2	43.1	19.5	35.2	36.9	25.7	45.4	48.8	25.7	40.9	42.8
		LLM-Generated	31.7	44.8	47.5	29.7	41.1	42.7	27.6	46.5	50.0	27.6	42.2	44.2
		Relative Δ	-45.3	-10.8	-9.7	-41.5	-15.5	-14.6	-7.1	-2.4	-2.4	-7.1	-3.1	-3.2
	Contriever	Human-Written	24.0	43.7	47.8	23.3	38.8	41.2	25.9	48.5	51.9	25.9	43.3	45.3
		LLM-Generated	31.0	47.8	50.5	29.6	43.2	44.8	32.5	51.9	55.4	32.5	47.5	49.4
		Relative Δ	-25.5	-9.0	-5.5	-23.8	-10.7	-8.4	-22.6	-6.8	-6.5	-22.6	-9.3	-8.7

Lexical models prefer human-written texts. Lexical models like TF-IDF and BM25 show a tendency to favor human-written texts over LLM-generated texts across most ranking metrics in both datasets. A plausible explanation for this phenomenon lies in the term-based distinctions between text generated by LLMs and human-written content. Additionally, the queries are crafted by humans and thus exhibit a style more closely aligned with human-written text.

Neural retrievers are biased towards LLM-generated texts. Neural models, which rely on semantic matching with PLMs, demonstrate a pronounced preference for LLM-generated texts, often performing over 30% better on these compared to human-written texts. These findings suggest an inherent bias in neural retrievers towards LLM-generated text, which we named the **source bias**. This source bias may stem from PLMs-based neural retrievers and LLMs sharing similar Transformer-based architectures Vaswani et al. (2017) and pretraining approaches, leading to potential exploitation of *semantic shortcuts* in LLM-generated text during semantic matching. Additionally, LLMs seem to semantically compress information in a manner that makes it more comprehensible to neural models. A deeper exploration into the causes of source bias is presented in the following section.

3.3 Bias in Re-Ranking Stage

In typical IR systems, there are two primary stages of document filtering: the first stage retrieval, and the subsequent second stage re-ranking. While we have revealed the presence of the source bias in the first stage, a natural pivotal research question remains: does this bias also manifest in the re-ranking stage? To delve into this, we select two representative and state-or-the-art re-ranking models: **MiniLM** Wang et al. (2020) and **monoT5** Nogueira et al. (2020) to rerank the top-100 document list retrieved by a first-stage BM25 model. The results on the SciFact+AIGC dataset with Llama-generated corpus and ChatGPT-generated corpus are presented in Table 4. From the results, while even the first-stage retrievers (BM25) may exhibit a preference for human-written content, the second-stage re-rankers once again demonstrate a bias in favor of LLM-generated content. Remarkably, the bias in re-ranking models appears to be more severe, as evidenced by the relative percentage difference of 67.3% and 59.4% in NDCG@1 for monoT5, respectively. These findings further confirm the pervasiveness of source bias in neural ranking models that rely on PLMs, regardless of the retrieval stage or re-ranking stage.

Table 4: Bias evaluation of re-ranking models on SciFact+AIGC dataset. The re-ranking methods rerank the top-100 retrieved hits from a first-stage BM25 model.

Metrics	Target Corpus	Llama2-generated			ChatGPT-generated		
		BM25	+MiniLM	+monoT5	BM25	+MiniLM	+monoT5
NDCG@1	Human-Written	26.7	21.3	19.7	24.3	18.3	21.3
	LLM-Generated	21.0	32.7	39.7	24.3	35.7	39.3
	Relative Δ	23.9	-42.2	-67.3	0.0	-64.4	-59.4
NDCG@3	Human-Written	40.3	42.8	45.9	38.5	41.4	46.4
	LLM-Generated	38.8	47.8	52.9	40.2	50.1	54.2
	Relative Δ	3.8	-11.0	-14.2	-4.3	-19.0	-15.5
NDCG@5	Human-Written	44.4	46.9	49.0	42.7	45.6	48.9
	LLM-Generated	41.5	50.2	54.7	42.7	53.0	56.1
	Relative Δ	6.8	-6.8	-11.0	0.0	-15.0	-13.7
MAP@1	Human-Written	25.7	20.8	18.9	23.7	17.9	20.5
	LLM-Generated	19.6	30.8	37.8	23.1	33.8	37.8
	Relative Δ	26.9	-38.8	-66.7	2.6	-61.5	-59.3
MAP@3	Human-Written	36.7	37.5	39.7	34.8	35.8	40.3
	LLM-Generated	34.3	43.6	48.9	35.8	45.9	50.0
	Relative Δ	6.8	-15.0	-20.8	-2.8	-24.7	-21.5
MAP@5	Human-Written	39.1	40.0	41.6	37.3	38.3	41.7
	LLM-Generated	35.9	45.0	50.1	37.3	47.6	51.4
	Relative Δ	8.5	-11.8	-18.5	0.0	-21.7	-20.8

4 RQ3: THE CAUSE OF SOURCE BIAS

In this section, we delve deeper into why neural retrieval models exhibitsource bias. Our objective is to determine whether the LLM-generated texts, characterized by reduced noise and more concentrated semantic topics, are inherently easier for neural retrieval models to semantically match. We conduct a series of analyses from the perspective of text compression and provide valuable insights.

4.1 VIEWPOINT FROM TEXT COMPRESSION

We first explore the cause of source bias from a compression perspective, drawing inspiration from recent studies that suggest LLMs are lossless compressors Delétang et al. (2023). We hypothesize that LLMs efficiently focus on essential information, minimizing noise during generation, in contrast to human-written texts, which may include more diverse topics and incidental noise. To verify this, we employ Singular Value Decomposition (SVD) Klema & Laub (1980) to compare topic concentration and noise in human-written and LLM-generated texts. The dimension of the SVD corresponds to the maximum number of topics, and the singular value associated with each topic represents its strength. High singular values predominantly capture primary topic information, whereas low singular values indicate noise.

Specifically, we utilize OpenAI embedding model to obtain embedding matrices for each corpus in the SciFact+AIGC dataset and then conduct SVD. The resulting singular values are arranged in descending order, and their comparison to the human-written corpus is visualized in Figure 4. As we can see, LLM-generated texts exhibit larger singular values at the top large singular values, while smaller singular values at the tail small singular values. This observation suggests that LLM-generated texts tend to have more focused semantics with less noise, rendering them more suitable for precise semantic matching. In contrast, human-written texts often contain a wider range of latent topics and higher levels of noise, making them harder for neural retrievers to understand. As a result, this difference in semantic concentration may contribute to the observed bias in neural retrievers.

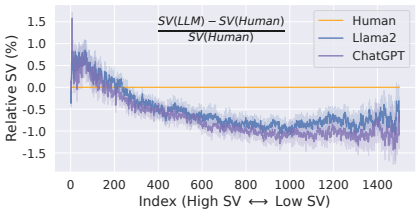


Figure 4: Comparision of the relative singular value (SV) of the different corpus after SVD. The SVs are sorted in descending order from left to right.

4.2 FURTHER ANALYSIS FROM PERPLEXITY

Considering that most modern neural retrievers are grounded on PLMs Yates et al. (2021); Guo et al. (2020); Zhao et al. (2023), such as BERT Devlin et al. (2019), Roberta Liu et al. (2019), and T5 Raffel et al. (2020), we analyze the perplexity of PLMs to further support the conclusion above from the viewpoint of compression that LLM-generated texts can be better understood by PLMs. Perplexity is an important metric for evaluating how well a language model can understand a given text Azzopardi et al. (2003); Wang et al. (2019). For a specific language model (LM) and a document $d = (d_0, d_1, \cdots, d_S)$, the log perplexity is defined as the exponentiated average negative log-likelihood of each token in the tokenized sequence of d[4]: $\text{PPL}(d) = -\frac{1}{S}\left(\sum_{s=1}^{S} \log P_{\text{LM}}(d_s|\text{context})\right)$, where S is the token length of text d and $P_{\text{LM}}(d_s)$ is the predicted likelihood of the s-th token conditioned on the context. Lower perplexity suggests more confidence and understanding of LM for text patterns, while higher perplexity implies greater uncertainty in predictions, often arising from complex or unpredictable text patterns.

Using the most widely-used LM, BERT Devlin et al. (2019), as an example, we employ it to calculate the PPL for different corpus. As BERT is not an autoregressive LM, we follow standard practices Wang et al. (2021); Wang & Cho (2019) to calculate the likelihood of each token conditioned on the other tokens, i.e.,

$$P_{\text{LM}}(d_s|\text{context}) := P_{\text{BERT}}(d_s|d_{\leq S \setminus \{s\}}).$$

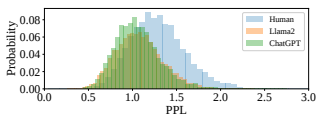


Figure 5: Comparision of the PPL of the different corpus.

[4]For simplicity, we denote the log perplexity as PPL.

The distribution of perplexity for different corpus in the SciFact+AIGC dataset is shown in Figure 5. Notably, LLM-generated texts consistently exhibit significantly lower perplexity, indicating enhanced comprehensibility and higher confidence from BERT's perspective. Consequently, PLMs-based neural retrievers can more effectively model the semantics of LLM-generated texts, leading to the observed source bias in favor of LLM-generated texts.

5 RQ4: Mitigating Source Bias

In this section, we propose a simple but effective approach to mitigate source bias by introducing a debiased constraint to the optimization objective. In this way, we can force the neural IR models to focus on modeling semantic relevance rather than the inherent semantic shortcut of the LLM-generated content.

5.1 Our Method: A Debiased Constraint

Our earlier findings of source bias indicate that neural retrievers tend to rank LLM-generated documents in higher positions. Thus, the motivation of our debiased method is straightforward, which is to force the retrieval models to focus on modeling the semantic relevance and not assign higher predicted relevance scores to the LLM-generated documents. Specifically, following the practice in Section 2.2, we first generate the corresponding LLM-generated corpus $\mathcal{C}^G$ for the original human-written training corpus $\mathcal{C}^H$. In this way, we can get the new paired training data $\mathcal{D} = \{(q_m, d_m^H, d_m^G)\}_{m=1}^M$, where each element (q_m, d_m^H, d_m^G) is a <query, human-written document, LLM-generated document> triplet. d_m^H and d_m^G are the corresponding human-written and LLM-generated relevant documents for the query q, respectively. Then we introduce the debiased constraint, which can be defined as

$$\mathcal{L}_{\text{debias}} = \sum_{(q_m, d_m^H, d_m^G) \in \mathcal{D}} \max\{0, \hat{r}(q, d^G; \Theta) - \hat{r}(q, d^H; \Theta)\} \tag{1}$$

where $\hat{r}(q, d^G; \Theta)$ and $\hat{r}(q, d^H; \Theta)$ are the predicted relevance scores of (q, d^G) and (q, d^H) by the retrieval models with parameters Θ, respectively. This constraint can penalize biased samples when the predicted relevance score of (q, d^G) is greater than that of (q, d^H).

Based on the debiased constraint defined in equation 1, we can define the final loss for training an unbiased neural retriever:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \alpha \mathcal{L}_{\text{debias}} \tag{2}$$

where the $\mathcal{L}_{\text{rank}}$ can be any common-used loss for the ranking task, e.g., contrastive loss or regression loss Zhao et al. (2023); Guo et al. (2020; 2022). And α is the debiased co-efficient that can balance the ranking performance and the degree of the source bias. The larger α indicates the greater penalty on the biased samples, leading to the retriever being more likely to rank the human-written texts in higher positions.

5.2 Results and Analysis

To evaluate the effectiveness of our proposed debiased method, we equip the debiased constraint defined in Eq. equation 1 to two representative neural retrievers: ANCE Xiong et al. (2020) and BERM Xu et al. (2023). In the experiments, we vary the debiased co-efficient α within the range of $\{1e\text{-}4, 5e\text{-}4, 1e\text{-}3, 5e\text{-}3, 1e\text{-}2\}$. The original retrieval models learned without the debiased constraint are denoted as "w/o debias". The results on the SciFact+AIGC dataset are presented in Figure 6.

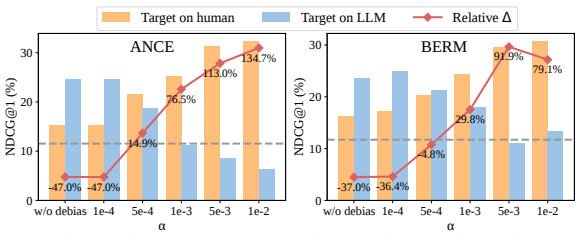
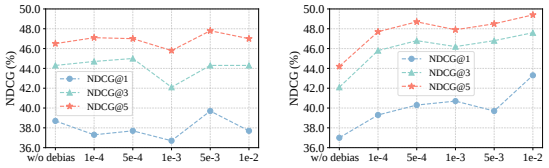


Figure 6: Performance comparison of neural retrievers on SciFact+AIGC with different co-efficient α in our proposed debiasing method. The grey dashed line represents Relative $\Delta = 0$. The results on other metrics and datasets have a similar tendency and are omitted due to space limitations.



(a) ANCE (b) BERM

Figure 7: Performance comparison of neural retrievers on only human-written SciFact dataset with different co-efficient α in our proposed debiased method.

As we can see, as the debiased co-efficient α increases, the Relative Δ gradually shifts from negative to positive across almost all metrics and mixed datasets. This trend indicates that the neural retrieval models can rank human-written text higher than LLM-generated text with large α. This can be attributed to the inclusion of our debiased constraint into the learning objective, which can penalize the biased samples and compel the retrieval models not to assign higher predicted relevance scores to LLM-generated content. Moreover, as shown in Figure 7, our method not only maintains the retrieval performance on the sole human-written corpus but also provides improvements, especially with BERM as the backbone. This improvement is likely due to the inclusion of LLM-generated samples, which might enhance the model's ability to discern relevance among similar documents.

In summary, these empirical results have demonstrated the efficacy of our proposed debiased method in mitigating source bias to different extents by adjusting the debiased coefficient α. This flexibility allows for customizing debiasing mechanisms to meet diverse perspectives and demands. Notably, the decision to maintain equality between the two content sources or favor human-written content can be tailored based on specific requirements and environmental considerations. The optimal strategy for enhancing the IR ecosystem remains an open question for further exploration.

6 Conclusion and Future Work

Conclusion. In this paper, we provide a preliminary analysis of the impact of the proliferation of generated content on IR systems, which is a pressing and emerging problem in the LLM era. We first introduce two new benchmarks, SciFact+AIGC and NQ320K+AIGC, and build an environment for evaluating IR models in scenarios where the corpus comprises both human-written and LLM-generated texts. Through extensive experiments within this environment, we uncover an unexpected bias of neural retrieval models favoring LLM-generated text. Moreover, we provide an in-depth analysis of this bias from the perspective of text compression. We also introduce a plug-and-play debiased strategy, which shows the potential to mitigate the source bias to different degrees. Finally, we discuss the crucial concerns and potential risks of this bias to the whole web ecosystem.

Discussion. Our study offers valuable insights into several promising directions for future research, including exploring source bias in other information systems (e.g., recommender and advertising systems), and examining source bias in neural models towards AIGC data across multiple data modalities, not limited to text. Moreover, with the burgeoning proliferation of LLMs and AIGC, source bias may raise significant concerns for a variety of aspects.

First, the presence of source bias poses a significant risk of gradually rendering human-written content less accessible, potentially causing a disruption in the content ecosystem. More severely, the concern is escalating with the growing prevalence of LLM-generated content onlineHanley & Durumeric (2023); Bengio et al. (2023). **Second**, there is the risk that source bias may amplify the spread of misinformation, especially considering the potential of LLMs to generate deceptive content, whether intentionally or not Chen & Shu (2023); Pan et al. (2023); Aslett et al. (2023). **Third**, source bias may be maliciously exploited to attack against neural retrieval models within today's search engines, creating a precarious vulnerability that could be weaponized by malicious actors, reminiscent of earlier web spam link attacks against PageRank Gyöngyi et al. (2004).

As discussed above, since LLMs can be readily instructed to generate texts at scale, source bias presents potential tangible and serious threats to the ecosystem of web content, public trust, and online safety. We hope this discussion will sound the alarm regarding the risks posed by source bias in the LLM era.

REFERENCES

Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. Information retrieval meets large language models: A strategic report from chinese ir community. *AI Open*, 4:80–90, 2023.

Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A Tucker. Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, pp. 1–9, 2023.

Leif Azzopardi, Mark Girolami, and Keith Van Risjbergen. Investigating the relationship between language model perplexity and ir precision-recall measures. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 369–370, 2003.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, 2023.

Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.

Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*, 2023.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. Language modeling is compression. *arXiv preprint arXiv:2309.10668*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067, 2020.

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–42, 2022.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp. 576–587, 2004.

Hans WA Hanley and Zakir Durumeric. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. *arXiv preprint arXiv:2305.09820*, 2023.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 113–122, 2021.

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

Virginia Klema and Alan Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176, 1980.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Hang Li. *Learning to rank for information retrieval and natural language processing*. Springer Nature, 2022.

Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Christopher D Manning. *An introduction to information retrieval*. Cambridge university press, 2009.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*, 2020.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

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

Vinu Sankar Sadasivan, Aounon Kumar, S. Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *ArXiv*, abs/2303.11156, 2023. URL `https://api.semanticscholar.org/CorpusID:257631570`.

Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4): 35–43, 2001.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. Ai model gpt-3 (dis) informs us better than humans. *arXiv preprint arXiv:2301.11924*, 2023.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 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*, 2021.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.

Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

Chenguang Wang, Mu Li, and Alexander J Smola. Language models with transformers. *arXiv preprint arXiv:1904.09408*, 2019.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

Xiting Wang, Xinwei Gu, Jie Cao, Zihua Zhao, Yulan Yan, Bhuvan Middha, and Xing Xie. Reinforcing pretrained models for generating attractive text advertisements. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3697–3707, 2021.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. Ai-generated content (aigc): A survey. *arXiv preprint arXiv:2304.06632*, 2023.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Jun Xu and Hang Li. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 391–398, 2007.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. BERM: Training the balanced and extractable representation for matching to improve generalization ability of dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 6620–6635, 2023.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pp. 1154–1156, 2021.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, dec 2023.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.