Neural Retrievers are Biased Towards LLM-Generated Content

Sunhao Dai
Yuqi Zhou
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
{sunhaodai,yuqizhou}@ruc.edu.cn

Liang Pang
CAS Key Laboratory of AI Safety
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
pangliang@ict.ac.cn

Weihao Liu
Xiaolin Hu
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
{weihaoliu,xiaolinhu}@ruc.edu.cn

Yong Liu
Xiao Zhang
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
{liuyonggsai,zhangx89}@ruc.edu.cn

Gang Wang
Huawei Noah's Ark Lab
Shenzhen, China
wanggang110@huawei.com

Jun Xu*
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
junxu@ruc.edu.cn

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 content 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 it easier for neural retrieval models 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 are available at https://github.com/KID-22/Source-Bias.

*Jun Xu is the corresponding author. Work partially done at Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671882

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

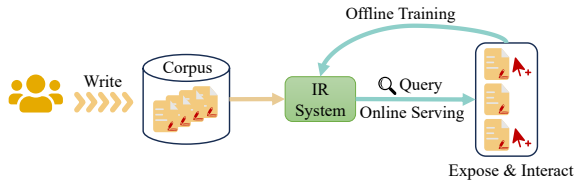
Source Bias, Information Retrieval, LLM-Generated Texts, Artificial Intelligence Generated Content

ACM Reference Format:
Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang and Jun Xu. 2024. Neural Retrievers are Biased Towards LLM-Generated Content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671882

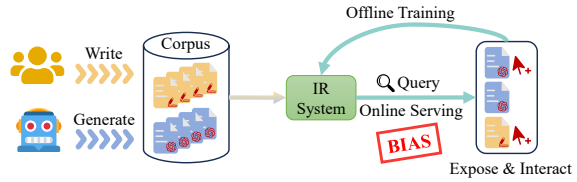
1 INTRODUCTION

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 [12, 65]. 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 [47, 64]. 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 [2, 75].

Meanwhile, information retrieval (IR) systems have become indispensable for navigating and accessing the Internet's vast information landscape [36, 45]. 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 [33, 34, 68]. 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 [25].



(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.

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 practical 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 same relevancy labels to LLM-generated data according to the original labels. 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 content?** We conduct comprehensive experiments with various representative retrieval models, ranging from traditional lexical models to modern neural models based on pretrained language models (PLMs) [22, 23, 72, 73]. 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 [17], we analyze the cause of source bias from the viewpoint of text compression. Our analysis of singular values [31] 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 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 [8, 25]. 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:

(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 relevancy labels. We believe these two benchmarks can serve as valuable resources for facilitating future research of IR in the LLM era.

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 relevancy labels for each query-document pair. Given that LLM-generated content is currently unidentifiable [42] and the significant cost of human annotation,

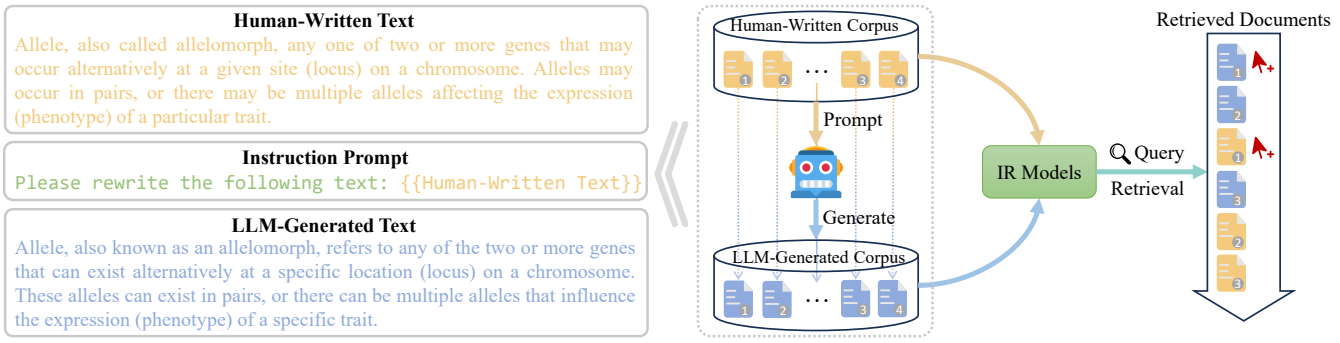


Figure 2: The overall paradigm of the proposed evaluation framework for IR in the LLM era.

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 relevancy 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 $q \in Q$ where 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 $C^H = \{d_1^H, d_2^H, \ldots d_N^H\}$ 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 LLM-generated corpus $C^G = \{d_1^G, d_2^G, \ldots, d_N^G\}$ where each document is generated by a LLM, e.g., d_1^G can be created by prompting ChatGPT to rewrite d_1^H while preserving its original semantics information. 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 $C = C^H \bigcup C^G$.

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.

2.2.1 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] [57] dataset aims to retrieve evidence from the research literature containing scientific paper abstracts for fact-checking. NQ320K[2] [32] 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 [52], we process these two datasets in a standard format: corpus C^H, queries Q, and relevancy 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 dataset.

2.2.2 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 relevancy labels to new <query, LLM-generated document> pairs as those assigned to the original labeled <query, human-written 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 [54] and ChatGPT to rewrite each seed human-written corpus, as Llama2 and ChatGPT are both the most widely-used and nearly the state-of-the-art open-sourced and closed-source LLM, respectively. We only generate texts with ChatGPT corresponding to the texts in SciFact dataset, mainly due to the significant cost involved in processing the larger NQ320K dataset.

For the LLM-generated corpus, we conduct post-processing to remove unrelated parts of the original response from LLM like "Sure, here's a possible rewrite of the text:". 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 C^H with its corresponding LLM-generated corpus C^G, original queries 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 summarizes the statistics of the proposed two datasets.

2.3 Statistics and Quality Validation of Datasets

Take the Llama2-generated data as an example, we conduct the statistics and quality validation of the constructed datasets. The analysis of ChatGPT-generated datasets shows similar observations and conclusions and is omitted due to the page limitation.

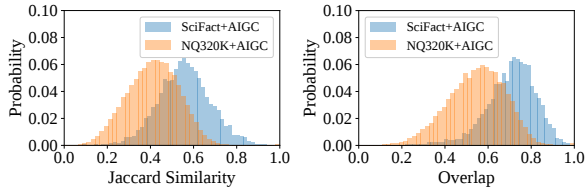
[1] https://allenai.org/data/scifact
[2] https://ai.google.com/research/NaturalQuestions

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 ; ChatGPT-Generated Corpus		
			# 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 ; 5,183	192.66 ; 203.57	1.1 ; 1.1
NQ320K+AIGC	7,830	9.24	109,739	199.79	1.0	109,739 ; –	174.49 ; –	1.0 ; –

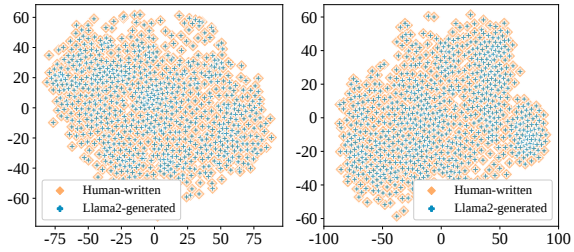
Table 2: Performance comparison of retrieval models on the sole human-written or Llama2-generated corpus on SciFact+AIGC and NQ320K+AIGC datasets. For brevity, we omit the percent sign '%' of ranking metrics in subsequent tables and figures.

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



(a) Jaccard Similarity (b) Overlap

Figure 3: Distribution of term Jaccard similarity and overlap between Llama2-generated and human-written corpora.



(a) SciFact+AIGC (b) NQ320K+AIGC

Figure 4: Semantic embedding visualization of different corpora on SciFact+AIGC and NQ320K+AIGC datasets.

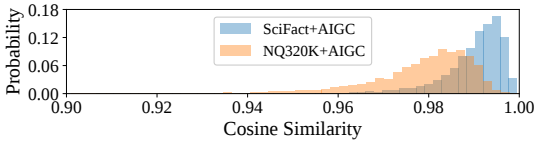


Figure 5: Distribution of cosine similarity of semantic embedding between Llama2-generated and human-written corpora.

2.3.1 Term-based Statistics and Analysis. We first analyze the term-based similarity between the LLM-generated corpus and the human-written corpus. Specifically, we compute the Jaccard similarity ($\frac{|d^G \cap d^H|}{|d^G \cup d^H|}$) and the overlap ($\frac{|d^G \cap d^H|}{|d^H|}$) between each LLM-generated document and orginal human-written document. As shown in Figure 3, both the Jaccard similarity and overlap distributions exhibit normal distribution, with peaks at about 0.6 and 0.8 for SciFact+AIGC, and about 0.4 and 0.6 for NQ320K+AIGC, respectively. These observations suggest that while there is a considerable overlap of terms between the LLM-generated text and the original human-written text, there are also distinct differences, especially noticeable in the NQ320K+AIGC dataset.

2.3.2 Semantic-based Statistics and Analysis. 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 relevancy labels as the labels of their corresponding original human-written texts given each query.

To assess this, we first leverage the OpenAI embedding model[3] to acquire semantic embeddings for both the LLM-generated and human-written texts. Subsequently, we visualize these embeddings through T-SNE [55] in Figure 4. We observe a strikingly close overlap between the Llama2-generated corpus and the human-written corpus in the latent space. This observation strongly suggests that these LLM-generated corpora adeptly preserve the original semantics. Moreover, we delve into the cosine similarity of semantic embeddings between the LLM-generated text and their corresponding human-written counterparts. The results, as shown in Figure 5, 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 relevancy label r to be the same as that of (q, d^H).

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

Table 3: Verification of semantics and text quality with human evaluation. The numbers in parentheses represent the proportion agreed upon by all three human annotators for each option.

SciFact+AIGC			NQ320K+AIGC		
Which document is more relevant to the given query?					
Human	LLM	Equal	Human	LLM	Equal
0.0%(0.0%)	0.0%(0.0%)	100.0%(82.0%)	2.0%(0.0%)	0.0%(0.0%)	98.0%(81.6%)
Which document exhibits higher quality by considering the following aspects: linguistic fluency, logical coherence, and information density?					
Human	LLM	Equal	Human	LLM	Equal
8.0%(0.0%)	6.0%(0.0%)	86.0%(46.5%)	4.0%(0.0%)	6.0%(0.0%)	90.0%(60.%)

2.3.3 Retrieval Performance Evaluation. To further validate the accuracy of the relevancy 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** [46] and **BM25** [41] and (2) Neural Retrieval Models: **ANCE** [67], **BERM** [70], **TAS-B** [26], **Contriever** [28].

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.

2.3.4 Human Evaluation. 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*. Moreover, to further verify this, we conduct a human evaluation. Specifically, we randomly select 50 <query, human-written document, LLM-generated document> triples from each dataset. The human annotators, comprising the authors and their highly educated colleagues, are asked to determine which document is more semantically relevant to the given query. The options are "Human", "LLM", or "Equal". During the evaluation, annotators are unaware of the source of each document. Each triple is labeled at least by three different annotators, with the majority vote determining the final label. The results in Table 3, confirm that both sources of texts have almost the same semantic relevance to the given queries, which guarantees the fairness of our following exploration of source bias.

Additionally, we also conduct further human evaluations specifically focused on text quality. The human annotators are asked to determine "Which document exhibits higher quality by considering the following aspects: linguistic fluency, logical coherence, and information density?" The notation process is the same as above, and the results are summarized in Table 3. The results indicate no significant distinction between LLM-generated and human-written content on text quality, demonstrating consistency across both sources. In fact, we also analyze the data cases and find that LLMs typically alter only parts of the vocabulary, leading to minor stylistic differences without impacting the core content, which can be further verified with these human evaluations.

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:

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

where the `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 between two sourced content.

3.2 Bias in Neural Retrieval Models

In our assessment of various retrieval models on SciFact+AIGC and NQ320K+AIGC datasets, we observe distinct phenomena when evaluating against human-written and LLM-generated corpora, as reported in Table 4. Our key findings are as follows:

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, as evident in Figure 3. 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 [56] 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

Table 4: Performance comparison of retrieval models for mixed human-written and Llama2-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

Table 5: Performance comparison of different neural retrieval models for mixed human-written and ChatGPT-generated corpora on SciFact+AIGC dataset.

Model	Target Corpus	NDCG@1	NDCG@3	NDCG@5	MAP@1	MAP@3	MAP@5
TF-IDF	Human-Written	22.7	36.5	39.5	22.0	32.8	34.6
	LLM-Generated	16.7	34.9	37.1	16.0	30.2	31.4
	Relative Δ	30.5	4.5	6.3	31.6	8.3	9.7
BM25	Human-Written	24.3	38.5	42.7	23.7	34.8	37.3
	LLM-Generated	24.3	40.2	42.7	23.1	35.8	37.3
	Relative Δ	0.0	-4.3	0.0	2.6	-2.8	0.0
ANCE	Human-Written	18.0	30.8	33.8	16.5	27.2	29.0
	LLM-Generated	24.7	35.6	37.4	24.0	32.7	33.7
	Relative Δ	-31.4	-14.5	-10.1	-37.0	-18.4	-15.0
BERM	Human-Written	16.3	29.9	32.3	14.8	26.0	27.4
	LLM-Generated	22.7	32.5	35.3	21.9	29.7	31.4
	Relative Δ	-32.8	-8.3	-8.9	-38.7	-13.3	-13.6
TAS-B	Human-Written	23.0	41.5	44.4	22.2	36.9	38.6
	LLM-Generated	28.7	45.5	46.7	27.2	40.9	41.6
	Relative Δ	-22.1	-9.2	-5.0	-20.2	-10.3	-7.5
Contriever	Human-Written	24.0	44.0	47.2	23.3	39.1	41.0
	LLM-Generated	33.0	48.3	50.6	31.3	44.0	45.4
	Relative Δ	-31.6	-9.3	-7.0	-29.3	-11.8	-10.2

Table 6: 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

it more comprehensible to neural models. A deeper exploration into the causes of source bias is presented in the following section.

To strengthen our conclusion that **source bias is not limited to any specific LLM**, we extend our investigation to include ChatGPT, another widely adopted and nearly state-of-the-art LLM. We employ ChatGPT to generate a corpus using the same prompts as those utilized with Llama2 in the above experiments. Subsequently, in Table 5, we report the evaluation results on the SciFact+AIGC dataset, which contains both human-written and ChatGPT-generated texts. Once again, the results clearly indicate a bias within neural retrieval models, favoring the corpus generated by ChatGPT across all ranking metrics. This observation provides additional substantiation of the presence of source bias within these neural retrieval models.

Furthermore, we also explore the popular InstructGPT-prompts GitHub Repository, which includes several common prompts for rephrasing passages [4]. The experimental results in Appendix A show varying degrees of source bias, indicating that common prompts can easily trigger source bias with LLM-generated content. These findings highlight the notable presence of source bias in neural retrieval models towards LLM-generated content.

3.3 Bias in Re-Ranking Stage

In a typical IR system, there are two primary stages of document filtering. The first stage involves a retriever, responsible for document recall, while the second stage employs a re-ranker, which fine-tunes the ordering of documents within the initially retrieved set. While we have revealed the presence of the source bias in the first stage, a natural pivotal research question remains: does this

[4] https://github.com/kevinamiri/Instructgpt-prompts

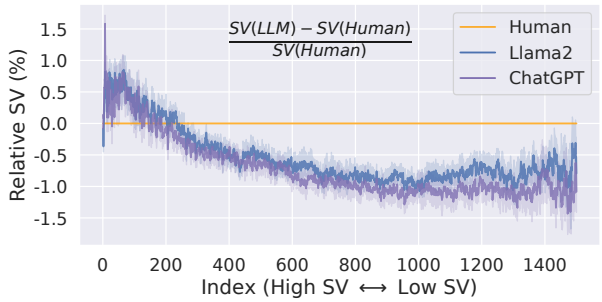


Figure 6: Comparision of the relative singular value (SV) of the different corpus after SVD. The singular values are sorted in descending order from left to right.

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** [62] and **monoT5** [37] 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 6. 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 first retrieval stage or second re-ranking stage.

4 RQ3: THE CAUSE OF SOURCE BIAS

In this section, we delve deeper into why neural retrieval models exhibit source 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 [17]. 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) [31] 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 the 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

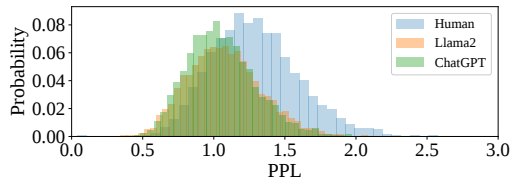


Figure 7: Comparision of the PPL of the different corpus.

in descending order, and their comparison to the human-written corpus is visualized in Figure 6. 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 source bias in neural retrievers.

4.2 Further Analysis from Perplexity

Considering that most modern neural retrievers are grounded on PLMs [23, 72, 73], such as BERT [19], Roberta [35], and T5 [40], 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 [6, 59]. 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[5]:

$$\mathrm{PPL}(d) = -\frac{1}{S}\left(\sum_{s=1}^{S} \log P_{\mathrm{LM}}(d_s|\mathrm{context})\right),$$

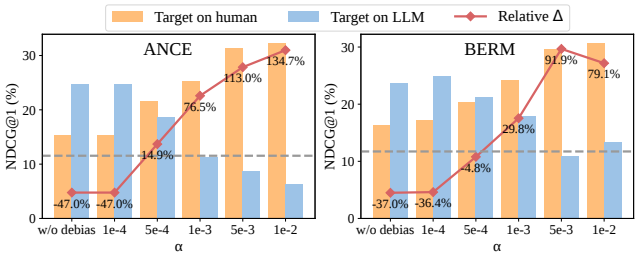
where S is the token length of text d and $P_{\mathrm{LM}}(d_s)$ is the predicted likelihood of the s-th token conditioned on the context. Lower perplexity suggests more confidence and understanding of LM for text patterns, while higher perplexity implies greater uncertainty in predictions, often arising from complex or unpredictable text patterns.

Using the most widely-used LM, BERT [19], as an example, we employ it to calculate the PPL for different corpus. As BERT is not an autoregressive LM, we follow standard practices [58, 63] to calculate the likelihood of each token conditioned on the other tokens, i.e.,

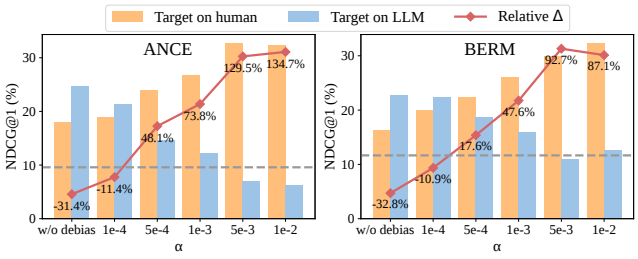
$$P_{\mathrm{LM}}(d_s|\mathrm{context}) := P_{\mathrm{BERT}}(d_s|d_{\leq S\setminus\{s\}}).$$

The distribution of perplexity for different corpus in the SciFact+AIGC dataset is shown in Figure 7. 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]For simplicity, we denote the log perplexity as PPL.



(a) Results on mixed human-written and Llama2-generated corpora



(b) Results on mixed human-written and ChatGPT-generated corpora

Figure 8: Performance comparison (NDCG@1) of neural models on SciFact+AIGC dataset with different debiased coefficient α. The grey dashed line represents Relative $\Delta = 0$.

In Appendix B, we also provide a theoretical analysis to illustrate and verify the observation in Figure 7 that LLM-generated texts have a smaller perplexity than human-written texts.

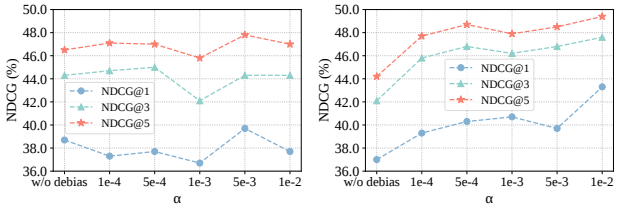
5 RQ4: MITIGATING SOURCE BIAS

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 C^G for the original human-written training corpus 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

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



(a) ANCE (b) BERM

Figure 9: Performance comparison of neural retrievers on only human-written SciFact dataset with different coefficient α in our proposed debiased method.

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 (1), we can define the final loss for training an unbiased neural retriever:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \alpha \mathcal{L}_{\text{debias}} \quad (2)$$

where the $\mathcal{L}_{\text{rank}}$ can be any common-used loss for the ranking task, e.g., contrastive loss or regression loss [22, 23, 73]. 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. (1) to two representative neural retrievers: ANCE [67] and BERM [70]. 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 8.

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 9, 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. For example, users may not mind the content's source if it is of high quality and fulfills their informational needs. However, bias becomes a significant issue when we aim to credit content providers and encourage more creation, impacting the sustainability of the content creation ecosystem. The optimal strategy for enhancing the sustainable development of the IR ecosystem remains an open question for further exploration.

6 DISCUSSION: SOUNDING THE ALARM

Through a rigorous series of experiments and thorough analysis, we have identified that neural retrieval models demonstrate clear preferences for LLM-generated texts, referred to as source bias. This bias, with the burgeoning proliferation of LLMs and AIGC, 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 online[8, 25]. **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 [5, 13, 39, 49]. **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 [24].

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.

7 RELATED WORK

Large Language Models for IR. The emergence of large language models (LLMs) [64, 71, 74] has ushered in a transformative era across various research domains, such as natural language processing (NLP) [7, 11], education [21, 38], recommender systems [15, 20], finance [27, 66], and medicine [3, 53]. In the field of IR, much effort has also been made to utilize the remarkable knowledge and capabilities of LLMs to enhance IR systems [2, 75]. In the industry community, an exemplary successful application is New Bing[6], which is an LLM-powered search assistant that adeptly extracts information from various web pages and delivers concise summarized responses to user queries, thereby improving the search experience. In the research community, there has been a proactive exploration of integrating LLMs into the IR components, including query rewriters [48, 60], retrievers [16, 69], re-rankers [14, 50], and readers [29, 43]. For a more comprehensive overview of the recent advancements in LLMs for IR, please refer to the recent survey [75].

Artificial Intelligence Generated Content. Artificial Intelligence Generated Content (AIGC) is a rapidly advancing field that involves the creation of content using advanced Generative AI (GAI) [1, 12, 65]. Unlike traditional content crafted by humans, AIGC can be generated at scale and in considerably less time [25, 47]. Recently, the development of LLMs and other GAI models has greatly improved the quality of AIGC content than before. For instance, LLMs such as ChatGPT have shown impressive abilities in generating human-like content [12, 65]. The DALL-E-3 [9], another state-of-the-art text-to-image generation system, can follow user instructions to produce high-quality images. Nevertheless, as AIGC becomes more prevalent across myriad domains, ethical concerns, and potential risks come into sharper focus [51, 61]. In fact, inevitably, the GAI models may generate content with bias and discrimination as the large training data always contain bias and toxicity [8, 18, 76]. Furthermore, researchers have found that LLMs can be manipulated into generating increasingly deceptive misinformation, posing challenges to online safety [13, 30, 49]. In addition, some recent studies indicate that training GAI models with synthetic data could result in the collapse of the next-generation models [4, 10, 44]. Thus, AIGC is a double-edged sword that requires cautious handling.

8 CONCLUSION AND FUTURE WORK

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.

Our study offers valuable insights into several promising directions for future research, including exploring source bias in other information systems (e.g., recommender systems and advertising systems) and examining source bias in neural models towards AIGC data across multiple data modalities, not limited to text. Moreover, uncovering the root cause of the source bias and thus further mitigating it are difficult but crucial research directions.

ACKNOWLEDGMENTS

This work was funded by the National Key R&D Program of China (2023YFA1008704), the National Natural Science Foundation of China (No. 62377044, 62276248, 62376275, 62076234), Beijing Natural Science Foundation (No. 4222029), Beijing Key Laboratory of Big Data Management and Analysis Methods, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, PCC@RUC, funds for building world-class universities (disciplines) of Renmin University of China, and the Youth Innovation Promotion Association CAS under Grants No.2023111. This work was supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (RUC24QSDL013). We thank all the anonymous reviewers for their positive and insightful comments.

[6]https://www.bing.com/new

REFERENCES

[1] Jorge Agnese, Jonathan Herrera, Haicheng Tao, and Xingquan Zhu. 2020. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 4 (2020), e1345.

[2] Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. 2023. Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community. *AI Open* 4 (2023), 80–90.

[3] Ian L Alberts, Lorenzo Mercolli, Thomas Pyka, George Prenosil, Kuangyu Shi, Axel Rominger, and Ali Afshar-Oromieh. 2023. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? *European journal of nuclear medicine and molecular imaging* 50, 6 (2023), 1549–1552.

[4] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2023. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850* (2023).

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

[6] Leif Azzopardi, Mark Girolami, and Keith Van Risjbergen. 2003. 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*. 369–370.

[7] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[8] 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. 2023. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688* (2023).

[9] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. 2023. Improving Image Generation with Better Captions. (2023).

[10] Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop. *arXiv preprint arXiv:2311.16822* (2023).

[11] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

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

[13] Canyu Chen and Kai Shu. 2023. Can LLM-Generated Misinformation Be Detected? *arXiv preprint arXiv:2309.13788* (2023).

[14] Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, and Jong C Park. 2023. Discrete Prompt Optimization via Constrained Generation for Zero-shot Re-ranker. *arXiv preprint arXiv:2305.13729* (2023).

[15] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*.

[16] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755* (2022).

[17] 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. 2023. Language modeling is compression. *arXiv preprint arXiv:2309.10668* (2023).

[18] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335* (2023).

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[20] Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046* (2023).

[21] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv preprint arXiv:2301.07597* (2023).

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

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

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

[25] Hans WA Hanley and Zakir Durumeric. 2023. Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. *arXiv preprint arXiv:2305.09820* (2023).

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

[27] Allen H Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research* 40, 2 (2023), 806–841.

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

[29] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299* (2022).

[30] Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2023. Disinformation Detection: An Evolving Challenge in the Age of LLMs. *arXiv preprint arXiv:2309.15847* (2023).

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

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

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

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

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

[36] Christopher D Manning. 2009. *An introduction to information retrieval*. Cambridge university press.

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

[38] Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2023. Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. *ArXiv* abs/2303.17003 (2023).

[39] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. *arXiv preprint arXiv:2305.13661* (2023).

[40] 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. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[41] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[42] Vinu Sankar Sadasivan, Aounon Kumar, S. Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-Generated Text be Reliably Detected? *ArXiv* abs/2303.11156 (2023). https://api.semanticscholar.org/CorpusID:257631570

[43] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652* (2023).

[44] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. Model dementia: Generated data makes models forget. *arXiv e-prints* (2023), arXiv–2305.

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

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

[47] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis) informs us better than humans. *arXiv preprint arXiv:2301.11924* (2023).

[48] Krishna Srinivasan, Karthik Raman, Anupam Samanta, Lingrui Liao, Luca Bertelli, and Michael Bendersky. 2022. QUILL: Query Intent with Large Language Models using Retrieval Augmentation and Multi-stage Distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 492–501.
[49] Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake News Detectors are Biased against Texts Generated by Large Language Models. *arXiv preprint arxiv:2309.08674* (2023).
[50] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. *arXiv preprint arXiv:2304.09542* (2023).
[51] Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 340–351.
[52] 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*.
[53] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.
[54] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[55] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[57] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974* (2020).
[58] Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. *arXiv preprint arXiv:1902.04094* (2019).
[59] Chenguang Wang, Mu Li, and Alexander J Smola. 2019. Language models with transformers. *arXiv preprint arXiv:1904.09408* (2019).
[60] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. *arXiv preprint arXiv:2303.07678* (2023).
[61] Tao Wang, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Zhihua Xia, and Jian Weng. 2023. Security and privacy on generative data in aigc: A survey. *arXiv preprint arXiv:2309.09435* (2023).
[62] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems* 33 (2020), 5776–5788.
[63] Xiting Wang, Xinwei Gu, Jie Cao, Zihua Zhao, Yulan Yan, Bhuvan Middha, and Xing Xie. 2021. Reinforcing pretrained models for generating attractive text advertisements. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3697–3707.
[64] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).
[65] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. 2023. Ai-generated content (aigc): A survey. *arXiv preprint arXiv:2304.06632* (2023).
[66] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).
[67] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).
[68] Jun Xu and Hang Li. 2007. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 391–398.
[69] Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2022. Match-Prompt: Improving Multi-task Generalization Ability for Neural Text Matching via Prompt Learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2290–2300.
[70] Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. 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*. 6620–6635.
[71] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712* (2023).
[72] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*. 1154–1156.
[73] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2023. Dense Text Retrieval based on Pretrained Language Models: A Survey. *ACM Trans. Inf. Syst.* (dec 2023).
[74] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
[75] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107* (2023).
[76] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).

A MORE EXPERIMENTAL RESULTS

In Table 7, we provide the results for source bias with more common prompts sourced from InstructGPT-prompts Github Repository [7]. These results indicate that common prompts can easily trigger source bias with LLM-generated content.

Table 7: Overall source bias in neural retrievers w.r.t. Relative Δ (NDCG@1) on SciFact+AIGC with mixed human-written and Llama2-generated corpora generated from different common rephrasing prompts.

Prompt	ANCE	BERM	TAS-B	Contriever
Rewrite the text below in your own words:	-7.0	-22.8	-11.2	-27.6
Paraphrase the provided text while maintaining its meaning:	-26.0	-55.3	-24.1	-13.6
Summarize the following passage in a concise manner:	-1.4	-43.3	-34.0	-32.4
Simplify the given passage while keeping the main ideas:	-29.0	-21.7	-22.5	-40.9
Rephrase the given text using alternative expressions:	-25.3	-34.7	-61.9	-18.4
Condense the following passage to focus on key points:	-19.0	-24.8	-22.8	-16.4
Briefly restate the provided text without losing its essence:	-29.6	-50.7	-41.3	-29.8
Reword the passage below to make it more succinct:	-40.6	-71.1	-54.7	-34.0
Express the following text in a different way while keeping its intent:	-50.7	-39.8	-34.9	0.0

B THEORETICAL ANALYSIS AND INSIGHTS

In Figure 7, we have compared the PPL for different corpus using the BERT model. In this section, we aim to further provide some theoretical insights into the above observations that LLM-generated texts have a smaller perplexity than human-written texts.

Without loss of generality, we define the PPL in an autoregressive manner. Let d^H denote a document written by humans, and d^G a document generated by an LLM conditioned on d^H. For a given document d and BERT model $\mathcal{B}$, PPL is calculated as

$$\mathrm{PPL}(d, \mathcal{B}) = -\frac{1}{S}\sum_{s=1}^{S} \log P_{\mathrm{BERT}}(d_s | d_{<s}).$$

Similarly, we use $\mathrm{PPL}(d, \mathcal{H})$ to represent the PPL of document d when evaluated by humans. The PPL of d^G conditioned on d^H is denoted as

$$\mathrm{PPL}(d^G \mid d^H, \mathcal{B}) = -\frac{1}{S}\sum_{s=1}^{S} \log P_{\mathrm{BERT}}(d^G_s | d^G_{<s}, d^H).$$

When evaluated by humans, we use $\mathrm{PPL}(d^G \mid d^H, \mathcal{H})$ to represent the PPL of d^G conditioned on d^H.

In the theorem below, we introduce three assumptions: Semantic Superiority, Conditional Redundancy, and Bounded Perplexity,

[7] https://github.com/kevinamiri/Instructgpt-prompts#rephrase-a-passage

to theoretically establish the sufficient conditions under which $\text{PPL}(d^G, \mathcal{B}) \leq \text{PPL}(d^H, \mathcal{B})$ holds. Semantic Superiority suggests that the perplexity of human-written texts, when evaluated by humans, is lower than when evaluated by BERT. Conditional Redundancy implies that the perplexity of d^G, given d^H, is lower than the perplexity of d^H when evaluated directly. This is intuitively true when the information added in generating from d^H to d^G doesn't exceed the original information in d^H. Bounded perplexity assumes that there exists an upper bound ϵ on the increase in perplexity when evaluating d^G directly, compared to evaluating d^G conditioned on d^H. Then we have the following theorem:

Theorem B.1. *Given the following conditions:*

- *Semantic Superiority: human beings outperform BERT in understanding human-written texts, i.e.,*
$$\text{PPL}(d^H, \mathcal{B}) - \text{PPL}(d^H, \mathcal{H}) \geq 0.$$
- *Conditional Redundancy: generating d^G from d^H adds less perplexity than d^H itself, i.e.,*
$$\text{PPL}(d^H, \mathcal{H}) - \text{PPL}(d^G \mid d^H, \mathcal{H}) \geq 0.$$
- *Bounded Perplexity: there exists a bounded non-negative difference ϵ in BERT's perplexity for d^G with or without d^H, i.e.,*
$$\text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{B}) \leq \epsilon.$$

If LLM aligns more closely with BERT than with humans when predicting d^G given d^H, such that for any $s \in [S]$,

$$\begin{aligned} &D_{\text{KL}}\left(P_{LLM}(d_s^G|d_{<s}^G, d^H) \| P_{BERT}(d_s^G|d_{<s}^G, d^H)\right) + \epsilon \\ &\leq D_{\text{KL}}\left(P_{LLM}(d_s^G|d_{<s}^G, d^H) \| P_{Human}(d_s^G|d_{<s}^G, d^H)\right), \end{aligned} \tag{3}$$

it follows that

$$\mathbb{E}_{P_{LLM}(d^G|d^H)}\left[\text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^H, \mathcal{B})\right] \leq 0.$$

In Theorem B.1, the KL divergence is used to compare the distributions of the document d^G conditioned on d^H according to the LLM, BERT model, and humans. It is worth emphasizing that inequation (3) is not the assumption on the understanding capabilities of BERT, LLM, and humans. Instead, this inequation assumes that when predicting d^G given d^H, the predictions by LLM are more closely aligned with those of BERT.

We demonstrate that, when inequation (3) is satisfied, the perplexity (evaluated by PLMs such as BERT) of d^G is lower than that of d^H. We'd like to emphasize that it is reasonable to expect that inequation (3) holds true because both LLM and BERT are Transformer-based models that use similar pretraining paradigms. The commonality in model structure and learning paradigms may lead to similar inherent biases in text prediction, making their predictions more aligned with each other.

The proof for Theorem B.1 is provided as follows:

Proof. We start by introducing the term $\text{PPL}(d^H, \mathcal{H})$:

$$\begin{aligned} &\text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^H, \mathcal{B}) \\ &\quad = \text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^H, \mathcal{H}) + \text{PPL}(d^H, \mathcal{H}) - \text{PPL}(d^H, \mathcal{B}) \\ &\quad \leq \text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^H, \mathcal{H}), \end{aligned}$$

where the last step follows from the Semantic Superiority condition.

$$\begin{aligned} \text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^H, \mathcal{B}) &\leq \text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^H, \mathcal{H}) \\ &= \text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{G}) \\ &\quad + \text{PPL}(d^G \mid d^H, \mathcal{G}) - \text{PPL}(d^H, \mathcal{H}). \end{aligned}$$

Next, we provide upper bounds of $\text{PPL}(d^G \mid d^H, \mathcal{G}) - \text{PPL}(d^H, \mathcal{H})$:

$$\begin{aligned} &\text{PPL}(d^G \mid d^H, \mathcal{G}) - \text{PPL}(d^H, \mathcal{H}) \\ &\quad = \text{PPL}(d^G \mid d^H, \mathcal{G}) - \text{PPL}(d^G \mid d^H, \mathcal{H}) \\ &\quad\quad + \text{PPL}(d^G \mid d^H, \mathcal{H}) - \text{PPL}(d^H, \mathcal{H}) \\ &\quad \leq \text{PPL}(d^G \mid d^H, \mathcal{G}) - \text{PPL}(d^G \mid d^H, \mathcal{H}), \end{aligned}$$

where the inequality follows from the Conditional Redundancy. Taking expectation on $\text{PPL}(d^G \mid d^H, \mathcal{G}) - \text{PPL}(d^G \mid d^H, \mathcal{H})$:

$$\begin{aligned} &- S\mathbb{E}_{P_{\text{LLM}}(d^G|d^H)}\left[\text{PPL}(d^G \mid d^H, \mathcal{G}) - \text{PPL}(d^G \mid d^H, \mathcal{H})\right] \\ &= \sum_{s=1}^{S} \mathbb{E}_{P_{\text{LLM}}(d^G|d^H)} \log \frac{P_{\text{LLM}}(d_s^G|d_{<s}^G, d^H)}{P_{\text{Human}}(d_s^G|d_{<s}^G, d^H)} \\ &= \sum_{s=1}^{S} \mathbb{E}_{P_{\text{LLM}}(d_{<s}^G|d^H)} \mathbb{E}_{P_{\text{LLM}}(d_s^G|d_{<s}^G, d^H)} \log \frac{P_{\text{LLM}}(d_s^G|d_{<s}^G, d^H)}{P_{\text{Human}}(d_s^G|d_{<s}^G, d^H)} \\ &= \sum_{s=1}^{S} \mathbb{E}_{P_{\text{LLM}}(d_{<s}^G|d^H)} \underbrace{D_{\text{KL}}(P_{\text{LLM}}(d_s^G|d_{<s}^G, d^H) \| P_{\text{Human}}(d_s^G|d_{<s}^G, d^H))}_{D_{\text{KL}}(P_{\text{LLM}} \| P_{\text{Human}})}. \end{aligned}$$

Similarly, $\text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{G})$ can be rewritten as:

$$\begin{aligned} &\text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{G}) \\ &\quad = \text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{B}) \\ &\quad\quad + \text{PPL}(d^G \mid d^H, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{G}). \end{aligned}$$

Taking expectation on $\text{PPL}(d^G \mid d^H, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{G})$:

$$\begin{aligned} &S\mathbb{E}_{P_{\text{LLM}}(d^G|d^H)}\left[\text{PPL}(d^G \mid d^H, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{G})\right] \\ &= \sum_{s=1}^{S} \mathbb{E}_{P_{\text{LLM}}(d_{<s}^L|d^H)} \underbrace{D_{\text{KL}}(P_{\text{LLM}}(d_s^G|d_{<s}^G, d^H) \| P_{\text{BERT}}(d_s^G|d_{<s}^G, d^H))}_{D_{\text{KL}}(P_{\text{LLM}} \| P_{\text{BERT}})}. \end{aligned}$$

Thus,

$$\begin{aligned} &\mathbb{E}_{P_{\text{LLM}}(d^G|d^H)}\left[\text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^H, \mathcal{B})\right] \\ &\leq \mathbb{E}_{P_{\text{LLM}}(d^G|d^H)}\left[\text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{B})\right] \\ &+ \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{P_{\text{LLM}}(d_{<s}^G|d^H)}(D_{\text{KL}}(P_{\text{LLM}} \| P_{\text{BERT}}) - D_{\text{KL}}(P_{\text{LLM}} \| P_{\text{Human}})). \end{aligned}$$

The final results can be derived by considering the assumptions:

$$\begin{aligned} \text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^G \mid d^H, \mathcal{B}) &\leq \epsilon \\ D_{\text{KL}}(P_{\text{LLM}} \| P_{\text{BERT}}) - D_{\text{KL}}(P_{\text{LLM}} \| P_{\text{Human}}) &\leq -\epsilon. \end{aligned}$$

From these, it follows that:

$$\mathbb{E}_{P_{\text{LLM}}(d^G|d^H)}\left[\text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^H, \mathcal{B})\right] \leq \epsilon - \epsilon = 0.$$

□