



Cocktail: A Comprehensive Information Retrieval Benchmark with LLM-Generated Documents Integration

Sunhao Dai¹, Weihao Liu¹, Yuqi Zhou¹, Liang Pang², Rongju Ruan³, Gang Wang³,
Zhenhua Dong³, Jun Xu^{1*}, Ji-Rong Wen¹

¹Gaoling School of Artificial Intelligence, Renmin University of China

²CAS Key Laboratory of AI Safety, Institute of Computing Technology, CAS

³Huawei Noah’s Ark Lab

{sunhaodai, junxu}@ruc.edu.cn, pangliang@ict.ac.cn

Abstract

The proliferation of Large Language Models (LLMs) has led to an influx of AI-generated content (AIGC) on the internet, transforming the corpus of Information Retrieval (IR) systems from solely human-written to a coexistence with LLM-generated content. The impact of this surge in AIGC on IR systems remains an open question, with the primary challenge being the lack of a dedicated benchmark for researchers. In this paper, we introduce Cocktail, a comprehensive benchmark tailored for evaluating IR models in this mixed-sourced data landscape of the LLM era. Cocktail consists of 16 diverse datasets with mixed human-written and LLM-generated corpora across various text retrieval tasks and domains. Additionally, to avoid the potential bias from previously included dataset information in LLMs, we also introduce an up-to-date dataset, named NQ-UTD, with queries derived from recent events. Through conducting over 1,000 experiments to assess state-of-the-art retrieval models against the benchmarked datasets in Cocktail, we uncover a clear trade-off between ranking performance and source bias in neural retrieval models, highlighting the necessity for a balanced approach in designing future IR systems. We hope Cocktail can serve as a foundational resource for IR research in the LLM era, with all data and code publicly available at <https://github.com/KID-22/Cocktail>.

IR benchmarks, notably MS MARCO (Nguyen et al., 2016), TREC (Craswell et al., 2020) and BEIR (Thakur et al., 2021), have exclusively utilized human-written content. However, the recent surge in Artificial Intelligence Generated Content (AIGC) facilitated by advanced Large Language Models (LLMs) has revolutionized the IR landscape (Dai et al., 2024a). This evolution has broadened the scope of IR systems, which now encompass a hybrid corpus composed of both human-written and LLM-generated content (Ai et al., 2023; Zhu et al., 2023; Dai et al., 2024a), presenting new challenges and opportunities of IR in the LLM era.

For instance, recent studies (Dai et al., 2024b; Xu et al., 2024) have shed light on a critical issue within neural retrieval models in the LLM era: a pronounced “source bias”. This bias, characterized by the preferential ranking of LLM-generated content over semantically equivalent human-written content, poses a significant threat to the IR ecosystem. Hence, the need to comprehensively understand such impact of LLM-generated content across different IR models and diverse IR domains and tasks has become more pressing, especially with the escalating prevalence of LLM-generated content (Hanley and Durumeric, 2023; Bengio et al., 2023). However, existing IR benchmarks either fail to reflect the real-world IR scenarios of the LLM era, as they solely contain human-written texts in their corpus, or provide limited datasets for exploring source bias. These shortcomings highlight the need for a comprehensive benchmark that accurately mirrors the current IR landscape, characterized by the integration of both human-written and LLM-generated texts within the corpus, to facilitate new research questions in this LLM era.

To fill this gap, we present a comprehensive benchmark tailored for IR in the LLM era, namely Cocktail, where the corpus contains both human-written and LLM-generated texts. Cocktail encompasses 16 retrieval datasets spanning differ-

1 Introduction

Information retrieval (IR) systems, as the keystone in overcoming information overload, have seen widespread application across various domains, including search engines (Li et al., 2014), question answering (Karpukhin et al., 2020), dialog systems (Chen et al., 2017), etc. A typical IR system aims at finding relevant documents or passages from a specific corpus in response to user’s queries (Li, 2022; Zhao et al., 2023a). Traditionally,

*Corresponding author.

ent domains and tasks, enabling both in-domain and out-of-domain evaluation settings. To construct these datasets, we first select 15 widely used public human-written corpora from MS MARCO (Nguyen et al., 2016), TREC (Craswell et al., 2020), and BEIR (Thakur et al., 2021). Then, based on these human-written corpora, we use Llama2 (Touvron et al., 2023) to rewrite each text to preserve semantic equivalence while introducing LLM-generated corpora. Finally, we mix the original human-written corpora and the LLM-generated corpora to get the final Cocktail corpora and assign the same relevancy label for the corresponding query-document pairs. Furthermore, to address the potential biases introduced by the inherent knowledge of LLMs during the rewriting process, we collect an additional new dataset, NQ-UTD. This new dataset comprises 80 queries and 800 documents from recent events. It serves as a critical component of Cocktail, offering an essential perspective for assessing the performance of IR systems in the context of both pre- and post-LLM era datasets.

With the benchmarked diverse datasets in Cocktail, we then conduct comprehensive evaluations of over ten state-of-the-art (SOTA) retrieval models through more than 1,000 experiments. Our analysis firstly reinforces previous findings by Dai et al. (2024b), highlighting a pervasive bias towards LLM-generated content across nearly all 16 datasets in Cocktail among all neural retrieval models. Furthermore, the results illustrated in Figure 1 reveal a distinct trade-off between ranking performance and source bias within these SOTA neural models. This observation suggests that while striving for high performance, these models may rely on inherent shortcuts, failing to grasp true semantic relevance and resulting in severe source bias. Hence, future work requires better considering a balance between performance and bias mitigation in the design of next-generation IR models.

In summary, our contributions are as follows:

(1) To the best of our knowledge, Cocktail is the first comprehensive benchmark with 16 datasets from a variety of domains and tasks tailed for IR research in the LLM era, where the corpus of each dataset contains both human-written texts and corresponding LLM-generated counterparts.

(2) We conduct extensive evaluations of state-of-the-art retrieval models using the Cocktail benchmark, assessing both retrieval accuracy and source bias. The evaluation tool, along with the codes,

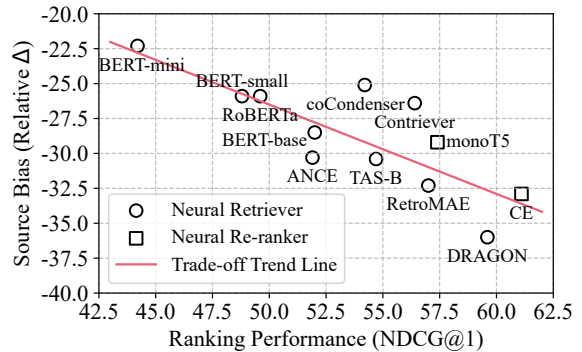


Figure 1: Ranking performance versus source bias comparison on averaged results of 16 datasets benchmarked in Cocktail. A more negative Relative Δ signifies increased source bias towards LLM-generated content. The Pearson correlation coefficient between these two axes is -0.798 (p -value < 0.05), indicating a strong negative correlation. For brevity, we omit the ‘%’ symbol of the scores in all the tables and figures.

is open-sourced, facilitating ease of adaptation for evaluating new models and datasets.

(3) Extensive empirical studies reveal a clear trade-off between ranking performance and source bias in neural retrieval models. This finding underscores the importance of achieving a suitable balance between performance improvement and bias mitigation in future IR model designs.

2 Related Work

IR meets Large Language Models. Information retrieval (IR), the keystone of information access, has now significantly been reshaped by the emergence of large language models (LLMs) (Zhao et al., 2023b; Ai et al., 2023; Dai et al., 2024a). This intersection has manifested in two pivotal ways. On the one hand, much effort has been made to utilize the advanced capabilities of LLMs to refine the whole retrieval pipeline (Zhu et al., 2023), including the integration of LLMs across various IR components, such as query rewriters (Srinivasan et al., 2022), retrievers (Wang et al., 2023), re-rankers (Sun et al., 2023b), and readers (Shi et al., 2023). On the other hand, the capacity of LLMs for generating human-like text at scale has shifted the landscape of searchable corpora, which now includes both human-written and LLM-generated texts. This evolution has introduced new challenges, most notably the emergence of source bias (Dai et al., 2024b; Xu et al., 2024; Dai et al., 2024a), where neural retrievers exhibit a preference for LLM-generated content, potentially compro-

Dataset				Train	Dev	Test			Avg. Word Length		
	Domain	Task	Relevancy	# Pairs	# Query	# Query	# Corpus	Avg. D/Q	Query	Human Doc	LLM Doc
Collected Before the Emergence of LLM (~ - 2021/04)											
MS MARCO	Misc.	Passage-Retrieval	Binary	532,663	-	6,979	542,203	1.1	6.0	58.1	55.1
DL19	Misc.	Passage-Retrieval	Binary	-	-	43	542,203	95.4	5.4	58.1	55.1
DL20	Misc.	Passage-Retrieval	Binary	-	-	54	542,203	66.8	6.0	58.1	55.1
TREC-COVID	Bio-Medical	Bio-Medical IR	3-level	-	-	50	128,585	430.1	10.6	197.6	165.9
NFCorpus	Bio-Medical	Bio-Medical IR	3-level	110,575	324	323	3,633	38.2	3.3	221.0	206.7
NQ	Wikipedia	Question Answering	Binary	-	-	3,446	104,194	1.2	9.2	86.9	81.0
HotpotQA	Wikipedia	Question Answering	Binary	169,963	5447	7,405	111,107	2.0	17.7	67.9	66.6
FiQA-2018	Finance	Question Answering	Binary	14,045	499	648	57,450	2.6	10.8	133.2	107.8
Touché-2020	Misc.	Argument Retrieval	3-level	-	-	49	101,922	18.4	6.6	165.4	134.4
CQADupStack	StackEx.	Dup. Ques.-Retrieval	Binary	-	-	1,563	39,962	2.4	8.5	77.2	72.0
DBPedia	Wikipedia	Entity-Retrieval	3-level	-	67	400	145,037	37.3	5.4	53.1	54.0
SCIDOCS	Scientific	Citation-Prediction	Binary	-	-	1,000	25,259	4.7	9.4	169.7	161.8
FEVER	Wikipedia	Fact Checking	Binary	140,079	6666	6,666	114,529	1.2	8.1	113.4	91.1
Climate-FEVER	Wikipedia	Fact Checking	Binary	-	-	1,535	101,339	3.0	20.2	99.4	81.3
SciFact	Scientific	Fact Checking	Binary	919	-	300	5,183	1.1	12.4	201.8	192.7
Collected After the Emergence of LLM (2023/11 - 2024/01)											
NQ-UTD	Misc.	Question Answering	3-level	-	-	80	800	3.7	12.1	101.1	94.7

Table 1: Statistics of all 16 datasets in Cocktail benchmark. Avg. D/Q denotes the average number of relevant documents per query.

missing the fairness and accuracy of search results. Moreover, the inclusion of LLM-generated content also raises concerns about privacy (Yao et al., 2023) and the dissemination of misinformation (Pan et al., 2023). In this paper, we focus on the second line and aim to establish a comprehensive benchmark for evaluating IR models in the LLM era, which can help understand the impact of LLM-generated content on IR systems.

Related Benchmarks. Historically, before the proliferation of LLM-generated content on the internet, several benchmarks were established for the evaluation of IR models, primarily utilizing corpora composed of human-written documents or passages. Notably, MS MARCOO (Nguyen et al., 2016) and TREC (Craswell et al., 2020) are widely used for supervised evaluation in IR research. Similarly, BEIR (Thakur et al., 2021) presents a diverse benchmark incorporating 18 datasets from various IR domains and tasks, tailored to zero-shot evaluation. Despite their contributions to advancing IR systems, these benchmarks fall short of reflecting the real-world scenarios of the current LLM era due to the absence of LLM-generated content in their corpora (Dai et al., 2024b; Xu et al., 2024; Dai et al., 2024a). This gap underscores the necessity for new benchmarks that include both human-written and LLM-generated texts, offering a more comprehensive and realistic evaluation environment to navigate the challenges and opportunities presented by the integration of LLMs into IR.

3 Benchmarking Retrieval Datasets

Cocktail establishes a comprehensive benchmark tailored for the evaluation of IR models in the LLM era, characterized by corpora containing both human-written and LLM-generated content. This benchmark aims to assess the performance and biases of existing retrieval models while encouraging the creation of future IR systems that excel in robustness and generalization across diverse scenarios in the LLM era. To achieve this, we collect and construct 16 IR datasets, incorporating 15 datasets from the pre-LLM era and one newly developed dataset to ensure a wide representation of domains and tasks. The statistics for the 16 datasets benchmarked in Cocktail are summarized in Table 1. We also list the dataset website links and the corresponding licenses in Appendix Table 5.

3.1 Dataset Construction

A key concern in constructing datasets with LLMs is the potential for LLMs to have prior knowledge about queries, which could lead to an unfair evaluation. To address this, following previous works (Dai et al., 2024b), we choose rewriting documents without incorporating query-related information rather than a full generation from LLMs with given queries, ensuring that any detected source bias is a genuine reflection of model preferences. Moreover, by avoiding full document generation based on queries, this approach also simplifies controlling the text length across different

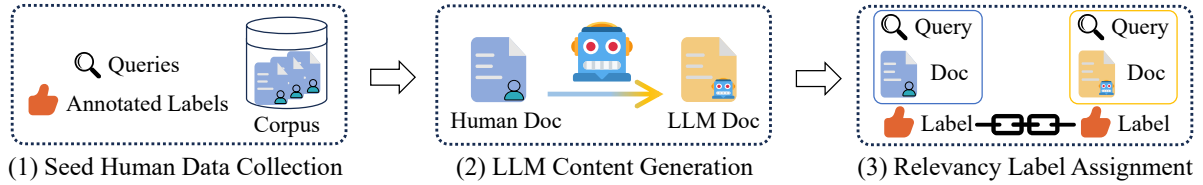


Figure 2: An overview of the dataset construction pipeline involved in Cocktail.

sources, ensuring consistency and further reducing potential biases in the evaluation. Specifically, we construct each dataset in Cocktail following a three-step process: 1) collecting seed datasets with human-written corpus and alongside relevancy labels for given queries; 2) leveraging LLMs to generate corresponding LLM-generated corpus with the human-written corpus as inputs; and 3) assigning relevancy labels to the LLM-generated content and given queries, ensuring seamless integration into the benchmark. An overview of our construction pipeline is shown in Figure 2.

Seed Human Datasets. To cover as diverse IR domains and tasks as possible, we select the widely-used 15 datasets from MS MARCO (Nguyen et al., 2016), TREC (Craswell et al., 2020), and BEIR (Thakur et al., 2021) benchmark as our human-written corpus. These datasets span six domains and eight distinct tasks, facilitating comprehensive evaluations under both in-domain and out-of-domain settings. Moreover, as suggested in Sun et al. (2023b), we also construct a new test dataset NQ-UTD to ensure that relevance annotations have not been learned by LLMs. NQ-UTD comprises 80 queries distributed across eight domains, sourcing from recent events (from Nov. 2023 to Jan. 2024). We verify that neither GPT-4 nor Gemini, two of the most advanced LLMs to date, can not answer the questions within NQ-UTD, validating no prior knowledge of these questions in LLM. For details about the collection and annotation processes of NQ-UTD, along with detailed descriptions of all 16 datasets, please refer to Appendix A.1.

LLM-Generated Corpus. To ensure a fair evaluation of the impact of incorporating LLM-generation content, we utilize the widely used llama-2-7b-chat to rewrite each human-written text without changing its semantic information, with the following instructions: “*Original Text: {{human-written text}}* Please rewrite the above given text. Your answer must be formatted as follows: *Rewritten Text: <your rewritten text>.*” This post-instruction strengthens the instruction-

following capabilities of LLMs, proving particularly effective for processing and generating responses to long input texts (Liu et al., 2023).

Relevancy Label Assignment. Upon the creation of the LLM-generated corpus for each dataset, we assign the relevancy labels from the original <query, human-written doc> pairs to the new <query, LLM-generated doc> pairs. This process is underpinned by the premise that both sources of the content preserve nearly identical semantic information, which will be further verified in the following section.

All the datasets in Cocktail are organized in a standard format (corpus, queries, qrels) akin to the BEIR benchmark (Thakur et al., 2021), facilitating ease of use and comparison. Examples for each dataset are showcased in Appendix Table 12. More details about the data processing and quality control are provided in Appendix A.2.

3.2 Dataset Statistics and Analysis

As shown in Table 1, there are minimal differences in average word length between human-written and LLM-generated documents, with the latter being marginally shorter. The detailed text length distribution, depicted in Appendix Figure 5, further confirms the negligible variance in text length. Additionally, term-based Jaccard similarity and overlap distributions between LLM-generated and human-written documents, visualized in Appendix Figure 6, show a noticeable distinction in terms despite similar semantic content.

Following practices in Dai et al. (2024b), our dataset construction process involved feeding only the original human-written text to the LLM for rewriting, without inputting any query-specific information to avoid introducing additional query-related bias. Furthermore, we employed the OpenAI embedding model¹ to obtain semantic embeddings for both text sources, comparing them through cosine similarity. These comparisons, presented in Appendix Figure 7, demonstrate a high

¹text-embedding-ada-002: <https://platform.openai.com/docs/guides/embeddings>

similarity, indicating successful semantic retention of the original human-written text in the LLM-generated texts. Additionally, evaluations of several retrieval models on sole human-written or LLM-generated corpora showed consistent performance, as seen in Appendix Table 8. Moreover, we also conduct human evaluations for quality verification for each dataset in Appendix Table 10. These observations reinforce confidence in the quality of our newly constructed datasets, suggesting that the LLM-generated content maintains semantics comparable to human-written texts for IR tasks. The detailed dataset analysis is provided in Appendix A.3.

4 Benchmarking Evaluation Protocol

Evaluation Framework. To standardize the assessment of IR models within the LLM era, we also develop a Python framework that combines user-friendliness with comprehensive evaluation capabilities. Built on the foundation of the BEIR (Thakur et al., 2021), our framework inherits its best features, including the ability to easily replicate experiments from open-sourced repositories and incorporate new models and datasets. A key innovation of our framework is its ability to conduct evaluations using either individual or mixed corpora, accommodating the mixture of human-written and LLM-generated content characteristic of the LLM era. These features make our framework an invaluable tool for advancing IR research and application in both academia and industry.

Evaluation Metrics. Aligned with the BEIR benchmark (Thakur et al., 2021), we select Normalized Discounted Cumulative Gain (NDCG@ K) as our primary metric to assess retrieval accuracy, given its robustness in capturing the effectiveness of IR systems across tasks with binary and graded relevance judgments. Following previous studies (Dai et al., 2024b; Xu et al., 2024), we choose $K = 1$ since the top-1 item in the retrieved list is most likely to be viewed and clicked by users. To provide a more comprehensive evaluation, we also report results on $K = 3$ and $K = 5$ in Appendix C.

Moreover, in the multi-sourced corpus shaped by LLMs, the evaluation of IR systems necessitates not only mere accuracy metrics but also a critical assessment of source bias (Dai et al., 2024b). This bias, as evidenced by ranking LLM-generated content higher than human-written content, poses a significant challenge in today’s IR ecosystem. To quantify and normalize source bias in differ-

ent datasets, we follow previous works (Dai et al., 2024b; Xu et al., 2024) and adopt the Relative Δ . This metric captures the relative percentage difference in NDCG scores between human-written and LLM-generated content, which is defined as:

$$\text{Relative } \Delta = \frac{\text{NDCG}_{\text{Human}} - \text{NDCG}_{\text{LLM}}}{(\text{NDCG}_{\text{Human}} + \text{NDCG}_{\text{LLM}})/2} \times 100\%,$$

where the $\text{NDCG}_{\text{Human}}$ and NDCG_{LLM} denote the NDCG scores attributed to human-written and LLM-generated content, respectively. Note that Relative $\Delta > 0$ indicates that IR models rank human-written content higher than LLM-generated content, and conversely, Relative $\Delta < 0$ indicates the opposite tendency. The absolute value of Relative Δ reflects the extent of source bias.

5 Benchmarking Retrieval Models

In this section, we delve into the evaluation and analysis of various retrieval models utilizing the constructed Cocktail benchmarked datasets.

5.1 Retrieval Models

Following the BEIR benchmark (Thakur et al., 2021), we focus on evaluating the advanced state-of-the-art transformer-based neural retrieval models. Besides the widely used **lexical retriever** BM25 (Robertson et al., 2009), our experiments include two main types of neural retrieval models:

Neural Retriever. We utilize and fine-tune the following two most commonly used pre-trained language models on the MS MARCO dataset (Nguyen et al., 2016) using the official training script² from BEIR: 1) BERT (Devlin et al., 2019); 2) RoBERTa (Liu et al., 2019). Additionally, we evaluate the performance of state-of-the-art models also trained on MS MARCO, employing officially released checkpoints: 3) ANCE (Xiong et al., 2020); 4) TAS-B (Hofstätter et al., 2021); 5) Contriever (Izacard et al., 2022); 6) coCondenser (Gao and Callan, 2022); 7) RetroMAE (Xiao et al., 2022); 8) DRAGON (Lin et al., 2023).

Neural Re-ranker. For re-ranking, we employ two state-of-the-art models with their publicly available official pre-trained checkpoints: 1) CE (Wang et al., 2020); 2) monoT5 (Raffel et al., 2020).

In our experiments, unless specified otherwise, neural re-rankers re-rank the top-100 documents

²https://github.com/beir-cellar/beir/blob/main/examples/retrieval/training/train_msmarco_v3.py

Model (→)	Lexical BM25	Neural Retrievers								Neural Re-rankers			
		BERT	RoBERTa	ANCE	TAS-B	Contriever	coCondenser	RetroMAE	DRAGON	CE	monoT5		
PLM	-	BERT	RoBERTa	RoBERTa	DistilBERT	BERT	BERT	BERT	BERT	MiniLM	T5	Average	
# Paras	-	110M	125M	125M	66M	110M	110M	110M	110M	66M	220M	All	Neural
Supervised Evaluation (In-Domain Datasets Collected in Pre-LLM Era)													
MS MARCO	38.8	51.6	51.9	52.8	54.4	55.6	55.3	55.7	59.1	<u>57.8</u>	56.2	53.6	55.0
DL19	57.8	<u>78.3</u>	76.4	72.5	71.7	72.1	76.4	77.5	<u>78.3</u>	81.0	75.2	74.3	75.9
DL20	55.3	<u>79.9</u>	76.9	73.2	72.5	73.8	79.3	76.5	82.4	73.5	77.8	74.6	76.6
Zero-shot Evaluation (Out-of-Domain Datasets Collected in Pre-LLM Era)													
TREC-COVID	67.0	67.0	62.0	68.0	65.0	64.0	70.0	72.0	75.0	<u>77.0</u>	85.0	70.2	70.5
NFCorpus	45.4	39.2	36.4	35.3	41.5	42.6	43.5	41.6	42.7	50.2	<u>49.4</u>	42.5	42.2
NQ	45.7	62.5	60.0	60.0	65.0	<u>69.6</u>	65.6	67.5	70.4	69.1	68.9	64.0	65.9
HotpotQA	84.3	75.3	62.9	72.1	84.6	88.1	82.1	88.5	89.1	93.9	<u>90.7</u>	82.9	82.7
FiQA-2018	24.4	24.7	24.9	28.9	28.1	34.0	27.2	31.8	<u>36.3</u>	35.5	38.7	30.4	31.0
Touché-2020	57.1	39.8	42.9	44.9	44.9	40.8	36.7	44.9	52.0	<u>56.1</u>	55.1	46.8	45.8
CQADupStack	28.6	26.6	26.9	31.3	23.4	34.2	33.2	31.3	<u>36.0</u>	34.7	36.5	31.2	31.4
DBPedia	36.6	55.3	50.6	48.0	56.3	62.1	56.9	56.8	59.6	<u>58.6</u>	32.9	52.2	53.7
SCIDOCS	16.2	12.3	11.2	14.5	16.0	16.6	14.5	16.2	17.9	<u>19.0</u>	19.5	15.8	15.8
FEVER	65.6	80.6	75.9	80.4	83.1	86.4	75.8	87.6	<u>88.0</u>	90.2	43.4	77.9	79.1
Climate-FEVER	25.2	26.4	27.4	28.3	32.7	31.4	28.1	32.3	34.0	<u>35.2</u>	35.3	30.6	31.1
SciFact	55.7	38.0	40.3	40.7	53.7	55.0	50.0	52.3	55.3	<u>56.3</u>	65.0	51.1	50.7
Zero-shot Evaluation (Out-of-Domain Datasets Collected in the LLM Era)													
NQ-UTD	76.9	74.4	66.9	78.8	81.9	75.6	73.1	78.8	76.9	89.4	<u>88.8</u>	78.3	78.5
Averaged Result													
Supervised	50.6	<u>69.9</u>	68.4	66.2	66.2	67.2	70.3	<u>69.9</u>	73.3	70.8	69.7	67.5	69.2
Zero-shot	48.4	47.9	45.3	48.6	52.0	53.9	50.5	54.0	<u>56.4</u>	58.9	54.6	51.8	52.2
All	48.8	52.0	49.6	51.9	54.7	56.4	54.2	57.0	<u>59.6</u>	61.1	57.4	54.8	55.4

Table 2: Overall ranking performance (NDCG@1) across all benchmarked datasets in Cocktail. The second-to-last column is the average result across all models, while the last column is the average for all neural retrieval models. The **best performed** result for each dataset is marked in bold, and the second best is underlined.

retrieved by BM25. Detailed information on the benchmarked models, including the publicly available official pre-trained checkpoints and implementation details, can be found in Appendix B.

5.2 Benchmarked Results

We conduct extensive evaluations with more than 1,000 experiments on the Cocktail benchmarked datasets. The results of retrieval accuracy and source bias across all benchmarked retrieval models and datasets are reported in Table 2 and Table 3³, respectively. Figure 1 shows average results of neural retrieval models across all datasets. From the results, we have the following key observations:

Neural models consistently exhibit source bias towards LLM-generated content. This bias is evident across neural retrieval models, spanning both in-domain and, more significantly, out-of-distribution datasets. This trend persists in data from both pre-LLM and LLM eras. Remarkably, the average Relative Δ across all neural models on the datasets surpasses -25% . These findings further support the findings of Dai et al. (2024b) and verify the widespread source bias in neural re-

trieval models, regardless of the domain or task, highlighting an urgent need to address this bias.

Stronger neural retrieval models exhibit more severe source bias. The results illustrated in Figure 1 underscore a significant trade-off faced by neural retrieval models: advancements in ranking performance often come with an increase in source bias. This trend suggests that these SOTA neural models may not fully understand semantic relevance. Instead, these models tend to leverage inherent shortcuts to enhance performance, inadvertently leading to increased bias. This phenomenon suggests that attempts to boost model performance could unintentionally magnify source bias issues, underlining the challenge of advancing model capabilities without leading to severe source bias.

Neural re-rankers generalize better but are still biased towards LLM-generated content. Neural re-rankers, while achieving superior ranking performance on most datasets and showing enhanced generalization capabilities compared to neural retrievers, are not exempt from pervasive source bias. Specifically, re-ranking models like CE and monoT5 exhibit a significant bias towards LLM-generated content, sometimes even more pronounced than that observed in neural retrievers.

³Note that SciFact was regenerated with our prompt, leading to slight result discrepancies from Dai et al. (2024b)

Model (→)	Lexical BM25	Neural Retrievers								Neural Re-rankers			
		BERT	RoBERTa	ANCE	TAS-B	Contriever	coCondenser	RetroMAE	DRAGON	CE	monoT5		
PLM	-	BERT	RoBERTa	RoBERTa	DistiBERT	BERT	BERT	BERT	BERT	MiniLM	T5	Average	
# Paras	-	110M	125M	125M	66M	110M	110M	110M	110M	66M	220M	All	Neural
Supervised Evaluation (In-Domain Datasets Collected in Pre-LLM Era)													
MS MARCO	72.2	-13.2	-18.9	1.1	-29.0	-10.8	5.1	-20.5	-18.2	-26.3	-7.8	-6.0	-13.8
DL19	108.7	-51.3	11.0	-0.8	-55.0	-21.4	-39.6	-53.9	-111.1	-18.3	-53.7	-25.9	-39.4
DL20	101.6	-76.5	-63.5	-2.5	-11.0	-10.8	-38.1	-43.7	-59.2	-25.3	-11.3	-21.8	-34.2
Zero-shot Evaluation (Out-of-Domain Datasets Collected in Pre-LLM Era)													
TREC-COVID	32.8	-62.7	-64.5	-58.8	-95.4	-87.5	-68.6	-66.7	-45.3	-64.9	-63.5	-58.6	-67.8
NFCorpus	-29.5	-30.6	-50.5	-23.2	-44.8	-99.5	-49.2	-17.3	-37.9	-66.1	-38.9	-44.3	-45.8
NQ	-17.9	-26.6	-16.7	-12.7	-41.8	-37.9	-25.0	-25.2	-47.2	-61.6	-33.6	-31.5	-32.8
HotpotQA	51.0	-1.1	-3.5	-13.6	0.2	-5.7	-5.4	5.2	-8.5	36.6	14.8	6.4	1.9
FiQA-2018	-8.2	-38.1	8.8	-33.3	-6.4	-38.3	-29.4	-52.8	-12.7	-42.3	-35.7	-26.2	-28.0
Touché-2020	-21.4	-25.6	-76.0	-36.1	-36.1	-29.8	10.9	-9.4	-66.4	-127.3	-66.4	-44.0	-46.2
CQADupStack	22.4	-45.1	-39.4	-19.8	-10.3	-22.2	-6.6	-67.1	-24.4	-30.5	-8.7	-22.9	-27.4
DBPedia	18.6	-5.4	-19.3	-25.8	2.5	4.5	-11.6	-24.6	-21.5	-13.3	5.5	-8.2	-10.9
SCIDOCS	2.5	21.1	-21.4	-9.7	-20.0	0.0	-26.2	-22.2	-16.8	-27.4	-35.9	-14.2	-15.8
FEVER	-26.2	2.5	-2.4	-87.1	-16.8	-20.4	-22.2	-1.4	-26.8	-4.4	11.5	-17.6	-16.7
Climate-FEVER	6.3	-15.2	-16.1	-109.9	-22.6	-12.7	-17.8	-15.5	-10.6	5.1	-89.0	-27.1	-30.4
SciFact	1.1	-52.6	-14.9	-29.6	-53.3	1.5	-21.6	-29.4	-38.7	-8.2	-38.2	-25.8	-28.5
Zero-shot Evaluation (Out-of-Domain Datasets Collected in the LLM Era)													
NQ-UTD	37.2	-35.5	-27.8	-22.3	-47.1	-31.4	-56.3	-73.1	-30.9	-51.9	-16.9	-32.4	-39.3
Averaged Result													
Supervised	94.2	-47.0	-23.8	-0.7	-31.7	-14.3	-24.2	-39.4	-62.8	-23.3	-24.3	-17.9	-29.2
Zero-shot	5.3	-24.2	-26.4	-37.1	-30.1	-29.2	-25.3	-30.7	-29.8	-35.1	-30.4	-26.6	-29.8
All	22.0	-28.5	-25.9	-30.3	-30.4	-26.4	-25.1	-32.3	-36.0	-32.9	-29.2	-25.0	-29.7

Table 3: Overall source bias evaluation w.r.t. Relative Δ (NDCG@1) across all benchmarked datasets in Cocktail. The **numbers** (i.e., Relative $\Delta > 0$) suggest that retrieval models generally prefer human-written content while the **numbers** (i.e., Relative $\Delta \leq 0$) indicate retrieval models prefer LLM-generated content.

Metric	BM25	+ CE	+ monoT5	DRAGON	+ CE	+ monoT5
NDCG@1	48.8	61.1	57.4	59.6 †10.8	59.8 †1.3	58.5 †1.1
Relative Δ	22.0	-32.9	-29.2	-36.0 ‡58.0	-36.5 ‡3.6	-33.7 ‡4.5

Table 4: Re-ranking results with the top-100 retrieved hits from a first-stage BM25 or DRAGON model.

This is particularly evident in datasets such as NQ and Touché-2020, further emphasizing the widespread nature of source bias in PLM-based neural retrieval models.

Bias in the first retrieval stage tends to propagate and even amplify during the re-ranking stage. This trend is particularly notable in datasets like NFCorpus and Touché-2020, where the source bias observed in the first retrieval stage persists and tends to intensify during the second-stage re-ranking. Moreover, as detailed in Table 4, enhancing the efficiency of the first-stage retrieval by replacing BM25 with DRAGON does not necessarily improve performance in the subsequent re-ranking phase. However, the source bias inherent in the first-stage retriever significantly impacts and may even magnify in subsequent re-ranking. This observation underscores the critical need for developing holistic approaches to mitigate bias throughout the

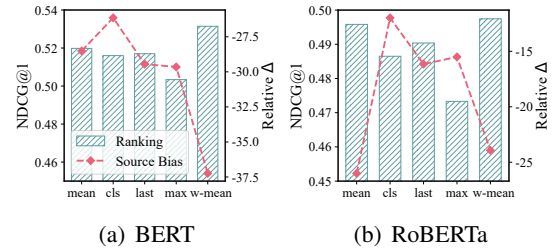


Figure 3: Results of different pooling strategies. “w-mean” denotes weighted mean pooling. A more negative Relative Δ signifies increased source bias towards LLM-generated content.

retrieval pipeline, ensuring fairness and accuracy in the whole IR system.

5.3 Further Analysis

Impact of Different Pooling Strategies. Pooling strategies in PLMs are critical for aggregating information from the token embeddings for downstream semantic matching. We explore the ranking performance and source bias on BERT and RoBERTa w.r.t. different pooling strategies, including CLS token, max token, last token (Muennighoff, 2022), mean (Reimers and Gurevych, 2019), and weighted mean pooling (Muennighoff, 2022). The averaged results of all benchmarked datasets in Cocktail are

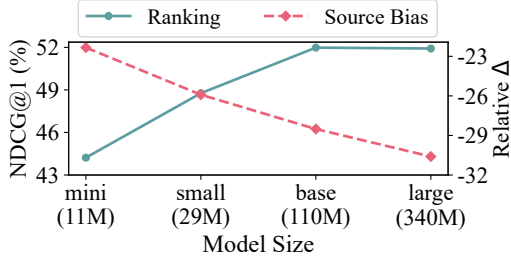


Figure 4: Comparison of different model sizes. A more **negative** Relative Δ signifies increased source bias towards LLM-generated content.

shown in Figure 3. As we can see, the ranking performance and degree of source bias vary significantly with the specific pooling strategy.

Weighted mean pooling demonstrates the most effectiveness for ranking, which can be attributed to the nuanced semantic understanding by positionally weighting tokens, thus enhancing document relevance matching. Yet, this strategy also incurs the most severe source bias, possibly because LLM-generated texts have distinctive structural or stylistic features that become more pronounced and amplified under weighted aggregation.

Conversely, max pooling, which selects the maximum value across each dimension from the token embeddings, appears to be the least effective. This could be due to its focus on the most dominant features within the text, potentially overlooking the broader contextual nuances captured by other strategies. The dominance of specific features might not always align with the relevance signals needed for accurate document ranking, explaining the lower performance and less bias.

Other strategies, such as mean, CLS token, and last token pooling, strike a balance by capturing overall text representations while still highlighting key tokens. However, the results show that while they are capable of supporting effective document ranking, they still suffer from severe source bias with Relative Δ far below -10% .

Overall, the variance in ranking performance and source bias across pooling strategies underscores the critical role of model architecture in retrieval model design. This analysis suggests that while weighted mean pooling offers enhanced ranking capabilities, it comes with a trade-off of increased source bias. Future work can explore hybrid or innovative pooling methods to balance performance with bias mitigation.

Impact of PLM Model Size. Our investigation

extends to the influence of model size on ranking performance and source bias, utilizing BERT models of varying sizes: mini, small, base, and large. As illustrated in Figure 4, an increase in model size leads to a notable enhancement in ranking performance. However, this improvement is paralleled by a rise in source bias (i.e., more negative).

The trend indicates that larger models, with their enhanced semantic understanding and processing capabilities, are more effective in judging document relevance, boosting their ranking performance. However, the increased capabilities may also make them more sensitive to the nuanced distinctions between human-written and LLM-generated texts, amplifying the source bias. Consequently, while the advanced performance of larger models is promising, it is coupled with a more severe bias risk, posing a significant challenge that merits more exploration in future research.

6 Conclusion and Future Work

This study proposes Cocktail, the first comprehensive benchmark consisting of 16 diverse datasets with mixed human-written and LLM-generated corpora. Alongside this, we present an evaluation tool equipped with standardized data formats and easily adaptable evaluation codes for a wide array of retrieval models. This tool is designed to systematically evaluate ranking performance and source bias in IR systems, both in supervised and zero-shot settings, paving the way for the development of more robust next-generation IR models.

Utilizing Cocktail, we conduct an extensive evaluation of over ten state-of-the-art neural retrieval models through more than 1,000 experiments. Our findings reveal a clear trade-off between ranking performance and source bias in neural retrieval models, underscoring the difficulty of enhancing model performance without increasing bias towards LLM-generated content. This challenge emphasizes the need for a suitable balance between performance enhancement and bias mitigation in the design of future IR models.

The newly collected and annotated NQ-UTD dataset comprises queries derived from recent events, featuring all content not yet incorporated into the pre-training knowledge of most LLMs. This characteristic renders it a valuable resource for fairly evaluating the effectiveness of LLMs in processing new, unseen data, especially for LLM-based retrieval or question-answering systems.

Limitations

We believe our proposed Cocktail benchmark is a foundational step toward advancing IR research in the LLM era. Nonetheless, our work still has several limitations for future research efforts. First, while our benchmarked 16 datasets encompass a broad range of IR domains and tasks, the IR field is continually evolving, with new areas of interest emerging regularly. Future updates to the Cocktail benchmark could benefit from including datasets from other IR domains, such as legal information retrieval (Sansone and Sperl , 2022). This expansion would not only diversify but also enhance the benchmark’s utility across more specialized areas. Second, the scale of our NQ-UTD dataset is currently limited to 800 query-passage pairs, primarily due to the high costs of human annotation. This encompasses financial costs, the time required to develop annotation guidelines, train annotators, and perform manual audits and validations. Future initiatives could focus on expanding the NQ-UTD dataset, possibly by employing LLMs to support and streamline the annotation process (Zhang et al., 2023). Such an approach could facilitate broader coverage and richer labeled query-passage pairs. Third, the construction of LLM-generated corpus in our benchmarked 16 datasets was significantly influenced by the inference costs of LLM, leading us to rely exclusively on Llama2 (Touvron et al., 2023) for generating LLM-based content. However, the real-world IR landscape is shaped by a variety of LLMs. Future research might include content generated by an array of LLMs, such as OpenAI ChatGPT and Google Gemini, to mirror the diversity of LLM-generated content more accurately. Despite these limitations, we envision the Cocktail benchmark as a valuable resource for IR research in the LLM era, offering a foundation upon which next-generation IR models can be built.

Ethics Statement

We commit to maintaining the highest ethical standards in our research, strictly adhering to ethical guidelines, applicable licenses, and laws. The 15 datasets utilized in this study are sourced from public repositories and have been employed within the bounds of their respective licenses and legal constraints. For the collected NQ-UTD dataset, we followed stringent ethical protocols, ensuring the corpus collected from online news sources was carefully screened and double-checked to remove

any personally identifiable or sensitive information. Participants involved in data annotation were thoroughly informed about the process and provided informed consent before their participation.

We recognize the potential risks posed by large language models (LLMs), such as generating harmful, biased, or incorrect content (Pan et al., 2023; Chen and Shu, 2023). The Llama2 model, used in this study for corpus generation, is not immune to these issues and may inadvertently produce misleading content. We have taken measures to minimize these risks, ensuring our benchmark aligns with ethical standards. Despite the acknowledged risks associated with LLMs, we assert that the scientific benefits and insights derived from our benchmark far outweigh potential concerns. Our resources are intended solely for scientific research, designed to foster advancements in information retrieval within this LLM era.

Acknowledgements

Jun Xu is the corresponding author. This work was funded by the National Key R&D Program of China (2023YFA1008704), the National Natural Science Foundation of China (No. 62377044, No.62376275), 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. This work was partially done at Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education.

References

- 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*, 4:80–90.
- 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*.
- Alexander Bondarenko, Maik Fr be, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, et al. 2020. Overview of touch  2020: argument retrieval. In *Experimental*

- IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 384–395. Springer.
- Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer.
- Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.
- Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.
- Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track.
- Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.
- Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024a. Unifying bias and unfairness in information retrieval: A survey of challenges and opportunities with large language models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024b. Neural retrievers are biased towards llm-generated content. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- 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*, pages 4171–4186.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leipold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.
- Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853.
- 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*.
- Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1265–1268.
- 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*, pages 113–122.
- Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.
- 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*, 7:453–466.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- Hang Li. 2022. *Learning to rank for information retrieval and natural language processing*. Springer Nature.

- Hang Li, Jun Xu, et al. 2014. Semantic matching in search. *Foundations and Trends® in Information Retrieval*, 7(5):343–469.
- Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *arXiv preprint arXiv:2302.07452*.
- Yijin Liu, Xianfeng Zeng, Fandong Meng, and Jie Zhou. 2023. Instruction position matters in sequence generation with large language models. *arXiv preprint arXiv:2308.12097*.
- 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*.
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www’18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.
- Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.
- Yikang Pan, Liangming Pan, Wenhui 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*.
- 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):5485–5551.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Carlo Sansone and Giancarlo Sperli. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.
- 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*.
- 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*, pages 492–501.
- Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2023a. Learning to tokenize for generative retrieval. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.
- 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*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- 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*.
- Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

- Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.
- 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*, 33:5776–5788.
- Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548.
- 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*.
- Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024. Invisible relevance bias: Text-image retrieval models prefer ai-generated images. *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Eric Sun, and Yue Zhang. 2023. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *arXiv preprint arXiv:2312.02003*.
- Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103.
- Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2023a. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Yujia Zhou, Jing Yao, Zhicheng Dou, Ledell Wu, and Ji-Rong Wen. 2022. Dynamicretriever: A pre-training model-based ir system with neither sparse nor dense index. *arXiv preprint arXiv:2203.00537*.
- 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*.

A Dataset Details

In this section, we provide details about the datasets in Cocktail. Table 5 lists the dataset website links and the corresponding licenses.

A.1 Detailed Description of Datasets

A.1.1 NQ-UTD

The **Natural Question-Up To Date (NQ-UTD)** dataset comprises 80 questions focusing on recent hotspot events from November 2023 to January 2024. These questions span eight domains: Sports, News, Science, Technology, Autos, Music, Movies, and Games, with approximately 10 queries per domain related to the latest events, questioning about their times, locations, or other specifics. To gather relevant passages for each query, we searched these questions on platforms like Google, X (formerly Twitter), Reddit, Wikipedia, Quora, and Facebook. Following the practice in Sun et al. (2023b), we also conducted searches using keywords to retrieve passages that are partially relevant but do not directly answer the questions. We also queried these questions with the latest state-of-the-art LLMs, gpt-4-1106-preview and gemini-pro, which have their knowledge bases updated only until April 2023. The results show their 0% accuracy in answering questions from our test set, confirming that these LLMs had no pre-knowledge about these questions. The collected 800 documents were then manually annotated for relevance by the authors and their highly educated colleagues. Each document received a relevance score: 0 for not relevant, 1 for partially relevant (mentioning some aspects of the query but not fully answering it), and 2 for relevant (providing a complete answer to the query). To ensure consistent and high-quality annotations, each document was reviewed by five individuals, with a majority vote determining the final label.

The statistics and examples of the NQ-UTD dataset are presented in Table 6 and Table 7, respectively. Note that this dataset also offers a fair evaluation of the current capabilities of the advanced LLM-based IR models. As NQ-UTD contains content not previously included in LLM training data, it can serve as a valuable resource for evaluating LLMs’ ability to process and answer queries on recent events.

A.1.2 MS MARCO

MS MARCO (Nguyen et al., 2016), developed by Microsoft Research, is a cornerstone dataset in the fields of natural language processing (NLP) and information retrieval (IR). This dataset comprises over a million user-generated questions derived from Bing search logs, tailored to advance three primary NLP tasks: document ranking, passage retrieval, and question answering. Within our Cocktail benchmark, we specifically utilize the passage retrieval subset, which contains 532K labeled query-passage pairs. A majority of the open-sourced pre-trained transformer checkpoints have been trained using this dataset. Following previous studies (Thakur et al., 2021; Lin et al., 2023), we employ the MS MARCO Dev set as the supervised evaluation test set.

A.1.3 TREC

DL19. The DL19 (TREC Deep Learning Track 2019) dataset (Craswell et al., 2020), is a key resource for exploring ad hoc ranking across extensive datasets from various domains. Our study focuses on the passage retrieval task utilizing the DL19 dataset, which comprises 43 queries and uses the MS MARCO passage corpus. This allows for its application in supervised evaluation settings.

DL20. The DL20 dataset (Craswell et al., 2021) comes from the second year of the TREC Deep Learning Track, and also focuses on ad hoc ranking through human-annotated training sets across diverse domains. We also focus on the passage ranking task featured in the DL20 dataset, which includes 54 queries and also leverages the MS MARCO passage corpus. This also allows for its application in supervised evaluation settings.

A.1.4 BEIR

If not specified, we utilize the preprocessed version of the following datasets as included in the BEIR benchmark (Thakur et al., 2021) to serve as the human-written seed datasets in our Cocktail benchmark, ensuring consistency and comparability in our evaluations.

TREC-COVID. The TREC-COVID dataset is developed by the Text REtrieval Conference (TREC) (Voorhees et al., 2021), specifically designed to address the challenges of biomedical information retrieval in the context of the COVID-19 pandemic. This dataset represents a concerted effort to harness the rapidly expanding collection of

Dataset	Official Website Link	License
MS MARCO	https://microsoft.github.io/msmarco/	MIT License
DL19	https://microsoft.github.io/msmarco/TREC-Deep-Learning-2019	CC BY 4.0 license
DL20	https://microsoft.github.io/msmarco/TREC-Deep-Learning-2020	CC BY 4.0 license
TREC-COVID	https://ir.nist.gov/covidSubmit/index.html	Dataset License Agreement
NFCorpus	https://www.cl.uni-heidelberg.de/statnlpgroup/nfcorpus/	
NQ	https://ai.google.com/research/NaturalQuestions	CC BY-SA 3.0 license
HotpotQA	https://hotpotqa.github.io	CC BY-SA 4.0 license
FiQA-2018	https://sites.google.com/view/fiqa/	
Touché-2020	https://webis.de/events/touche-20/shared-task-1.html	CC BY 4.0 license
CQADupStack	http://nlp.cis.unimelb.edu.au/resources/cqadupstack/	License 2.0 license
DBPedia	https://github.com/iai-group/DBpedia-Entity/	CC BY-SA 3.0 license
SCIDOCS	https://allenai.org/data/scidocs	GNU General Public License v3.0 license
FEVER	http://fever.ai	CC BY-SA 3.0 license
Climate-FEVER	http://climatefever.ai	
SciFact	https://github.com/allenai/scifact	CC BY-NC 2.0 license
NQ-UTD	https://github.com/KID-22/Cocktail	MIT License

Table 5: Website links and licenses for the benchmarked datasets in Cocktail. (Note: Licenses for NFCorpus, FiQA-2018, and Climate-FEVER are not provided by the authors).

Domain	# Queries	# Passages	# Relevancy Scores			# Source					
			0	1	2	Google	X	Reddit	Wikipedia	Quora	Facebook
Sports	12	120	66	17	37	60	22	10	19	9	0
News	13	130	78	14	38	68	23	11	13	15	0
Scientific	10	100	68	12	20	59	10	8	6	9	8
Technology	12	120	92	1	27	64	6	4	12	13	21
Autos	8	80	50	15	15	51	6	1	7	11	4
Music	10	100	60	18	22	67	6	8	5	14	0
Movies	7	70	41	8	21	53	2	3	9	3	0
Games	8	80	48	9	23	56	3	9	8	4	0
All	80	800	503	94	203	478	78	54	79	78	33

Table 6: Statistics of our proposed NQ-UTD dataset. The column ‘# Relevancy Scores’ represents the count of documents categorized by their respective relevance levels.

scholarly articles related to COVID-19, providing a focused resource for biomedical IR.

NFCorpus. The NFCorpus dataset (Boteva et al., 2016) serves as a detailed resource tailored for ranking tasks within the biomedical field. A distinctive aspect of NFCorpus is its meticulously curated relevance links, connecting queries directly to pertinent research articles.

NQ. The NQ (Natural Questions) dataset aims to propel advancements in natural language understanding, particularly focusing on the question-answering task (Kwiatkowski et al., 2019). This dataset is compiled from real questions submitted by users on Google search, each annotated with answers extracted from Wikipedia articles.

HotpotQA. HotpotQA (Yang et al., 2018) is a large-scale dataset that includes diverse questions posed by human participants. These questions necessitate a multi-step reasoning process, often requiring the extraction of answers from extensive Wikipedia texts and the identification of supporting evidence within them. HotpotQA is specifically

crafted to evaluate the effectiveness of models in deciphering complex, multi-hop questions and delivering precise answers.

FiQA-2018. FiQA-2018 (Financial Opinion Mining and Question Answering-2018) (Maia et al., 2018) presents a specialized question-answering challenge within the financial domain. The dataset benefits from a comprehensive corpus gathered from esteemed financial news platforms, analyst reports, and influential financial blogs.

Touché-2020. Touché-2020 dataset (Bondarenko et al., 2020) was created with the specific aim of advancing the field of argument retrieval. Its purpose is to identify argumentative content relevant to contentious queries. It contains argumentative passages from a broad spectrum of domains, which are carefully selected and annotated.

CQADupStack. CQADupStack (Hoogeveen et al., 2015) is a popular dataset for duplicate question retrieval, which aims to identify duplicate questions in community question answering (cQA) forums. This dataset includes carefully annotated

Domain	Question	Reference Answer
Sports	Who is the MVP of the first NBA In-Season Tournament?	LeBron James
Sports	Who wins 2023 FIFA Club World Cup?	Manchester City F.C.
News	Where was the 44th Gulf Cooperation Council (GCC) Summit held?	Doha, Qatar
News	Which country will take over as the rotating chair of the APEC in 2024?	Peru
Scientific	Which paper won NeurIPS2023 Test of Time Award?	Distributed Representations of Words and Phrases and their Compositionality
Scientific	Which university designed AI Coscientist?	Carnegie Mellon University
Technology	Which company released the machine learning framework MLX?	Apple
Technology	What is the maximum number of cores in Intel fifth-generation Xeon CPU?	64
Autos	When did Xiaomi announce SU7?	Dec 2023
Autos	The country with the largest automobile export volume in 2023?	China
Music	Which singer was named the 2023 Person of the Year by Time magazine?	Taylor Swift
Music	When did Les McCann pass away?	December 29, 2023
Movies	Which movie was the highest-grossing film worldwide in 2023?	Barbie
Movies	Which actress won the Best Actress award at the European Film Awards?	Sandra Wheeler
Games	Which team is the champion of the League of Legends S13 Finals?	SKT T1
Games	Which one is the best-selling games in the US 2023?	Hogwarts Legacy

Table 7: Examples of queries and reference answers from different domains of our proposed NQ-UTD dataset.

and categorized questions and answers across various domains. To maintain focus without losing the essence of generalizability, our study incorporates a subset from the English domains as our selected CQADupStack dataset.

DBPedia. The DBPedia dataset (Hasibi et al., 2017) includes structured information about entities, concepts, categories, and their relationships, gathered from Wikipedia entries. This dataset focuses on the entity retrieval task, where queries aim to retrieve relevant entities from the English DBpedia corpus dated October 2015.

SCIDOCS. The SCIDOCS dataset (Cohan et al., 2020) is a vast collection of scholarly articles from the Semantic Scholar database, enhanced with detailed metadata extraction and annotations. It serves as a robust dataset for the citation prediction task, where the objective is to identify papers cited by a given query paper title.

FEVER. The FEVER (Fact Extraction and VERification) dataset (Thorne et al., 2018) comes from the fact-checking domain, offering a curated collection of human-labeled claims sourced from Wikipedia. Each claim is meticulously classified as Supported, Refuted, or NotEnoughInfo, setting the stage for a nuanced fact-checking task.

Climate-FEVER. Climate-FEVER (Diggelmann et al., 2020) is a specialized dataset focused on the area of climate science, comprising real-world claims accompanied by evidence sentences from Wikipedia. Similar to the FEVER dataset, the primary task involves evaluating whether the provided evidence supports, refutes, or lacks sufficient infor-

mation to adjudicate each claim.

SciFact. SciFact (Wadden et al., 2020) is designed for the verification of scientific claims sourced from peer-reviewed literature, with each claim meticulously matched with corroborative or refuting evidence from related studies. The dataset comprises expert-crafted scientific claims alongside abstracts containing evidence, each annotated with labels indicating support or contradiction, and rationales explaining the basis of the evidence.

A.2 Dataset Processing and Quality Control

Given the substantial computational cost associated with rewriting documents for the 16 datasets using Large Language Models (LLMs), we adopted sampling methods in line with practices from (Zhou et al., 2022; Sun et al., 2023a). Specifically, for datasets with an original corpus size smaller than 200,000 documents, we retained the entire corpus without filtering. For datasets exceeding 100,000 documents, we retain all candidate documents that appear in the labeled data (including train, valid, and test set). If the number of documents post-filtering fell below 100,000, we supplemented the corpus with a random selection of up to 100,000 additional documents from the remaining corpus to ensure a challenging dataset size.

In refining each corpus, we only kept documents with text lengths between 10 and 2,000 words. This criterion was set because texts shorter than 10 words often consisted of empty texts or symbols, compromising data quality—this was particularly evident in the TREC-COVID dataset, which contained over 40,000 documents with empty text.

Conversely, texts longer than 2,000 words were filtered out to accommodate the context length limitation of 4,096 tokens for rewriting with Llama2, impacting only 0.05% of the data. These filters did not significantly alter the overall data distribution, with a total of less than 1% data removed, and our evaluation results align closely with those reported in the BEIR benchmark (Thakur et al., 2021) for the original datasets, as shown in Table 2.

When LLM is used for rewriting, LLM will inevitably refuse to rewrite due to the safety constraints, especially for datasets like Touché-2020, which focus on retrieving contentious viewpoints. For such LLM-generated data, we keep the same content as the human-written counterparts. This approach allows for the easy removal of refused rewrites if necessary, ensuring both the integrity and the quality of the dataset processing and control measures implemented. The statistics of the final 16 benchmarked datasets in Cocktails are summarized in Table 1.

A.3 Dataset Statistics and Analysis

Term-based Statistics. Figure 5 illustrates the distribution of text lengths, revealing minimal variation between the lengths of texts. We then calculated the Jaccard similarity and overlap between each LLM-generated document d^G and the corresponding human-written document d^H , using the following two formulas:

$$\text{Jaccard similarity} = \frac{|d^G \cap d^H|}{|d^G \cup d^H|},$$

$$\text{Overlap} = \frac{|d^G \cap d^H|}{|d^H|}.$$

The results presented in Figure 6 highlight significant differences in terms despite their ostensibly similar semantic information.

Semantic Analysis. Cosine similarity between LLM-generated and human-written documents, calculated using the OpenAI embedding model⁴, is displayed in Figure 7. The results, predominantly above 0.9, signify a high level of semantic preservation in the LLM-generated texts compared to the original human-written content. We also compare the semantics between randomly selected <human doc, LLM doc> pairs and matching pairs <human doc, LLM doc>, as shown in Table 9. The average similarity score for matching pairs was

⁴text-embedding-ada-002

0.9816, significantly higher than the 0.7135 average for random pairs. This significant statistical difference supports the close semantic alignment between LLM-generated and human-authored documents, further validating the quality of our constructed dataset. Further, evaluations of various retrieval models on sole human-written or LLM-generated corpora, as detailed in Table 8, demonstrate consistent performance across all datasets between two types of corpus. These observations across all datasets reinforce confidence in the quality of our newly constructed datasets, suggesting that the LLM-generated content maintains semantics comparable to human-written texts for IR tasks.

Quality Verification with Human Evaluation. To further validate the assigned relevancy label, we also conduct a human evaluation study. Due to cost constraints of human evaluation, we sample 20 <query, human doc, LLM doc> triples from each of the 16 datasets included in Cocktail. These triples were annotated by graduate students and our colleagues, who were asked to assess which document is more semantically relevant to the given query, with options being “human doc”, “LLM doc”, or “equal”. At least three different annotators evaluated each triple, with the majority vote deciding the final label. We summarized the human evaluation results in Table 10, with numbers in parentheses indicating the percentage of agreement among all three evaluators for each option. The results show a comparable level of semantic relevance between human-written and LLM-generated texts for the given queries, ensuring the fairness of our analysis of source bias.

A.4 Dataset Examples

For a better understanding of the datasets used in our Cocktail benchmark, we offer examples for each dataset in Table 12, showcasing a given query along with the corresponding relevant human-written document and LLM-generated document.

B Model Details

B.1 Detailed Description of Models

We select the widely used approach **BM25** (Robertson et al., 2009) as the representation of **lexical retrieval models** in our benchmark.

For **neural retrieval models** that leverage pre-trained language models to acquire semantic relationships between queries and documents, we

Model (→)		Lexical BM25	Neural Retrievers								Neural Re-rankers	
			BERT	RoBERTa	ANCE	TAS-B	Contriever	coCondenser	RetroMAE	DRAGON	CE	monoT5
MS MARCO	Human	38.6	49.9	49.8	51.4	54.2	55.3	54.1	55.2	60.2	58.3	56.8
	LLM	34.3	49.5	50.1	50.1	51.8	53.5	52.4	53.0	56.2	55.2	53.7
DL19	Human	56.2	77.1	79.5	72.1	76.4	74.4	77.9	76.4	76.7	80.2	78.7
	LLM	51.2	79.8	73.6	68.2	73.3	70.9	79.5	76.0	79.1	81.0	80.6
DL20	Human	54.6	74.7	75.3	73.5	75.9	73.2	78.7	73.5	81.8	79.6	76.9
	LLM	45.7	79.3	73.8	76.2	71.9	75.6	77.5	79.3	78.7	75.3	76.5
TREC-COVID	Human	62.0	63.0	61.0	74.0	75.0	64.0	73.0	79.0	75.0	68.0	83.0
	LLM	63.0	63.0	60.0	67.0	63.0	60.0	70.0	73.0	65.0	75.0	86.0
NFCorpus	Human	43.8	38.4	32.5	33.4	40.9	42.9	43.0	40.3	42.6	49.2	48.9
	LLM	45.8	37.5	36.1	34.1	41.0	42.3	43.5	41.8	43.2	49.7	48.4
NQ	Human	44.8	61.4	59.5	58.9	64.3	69.3	64.5	66.5	69.9	68.5	69.3
	LLM	43.5	60.5	57.8	58.1	64.0	68.1	63.8	66.3	68.9	68.0	68.0
HotpotQA	Human	83.9	74.1	61.5	70.4	84.2	88.0	80.8	88.2	89.0	94.2	91.1
	LLM	80.7	73.7	62.3	70.8	83.4	87.3	81.1	87.2	88.2	92.6	89.6
FiQA-2018	Human	23.3	23.3	24.5	27.8	28.7	31.3	27.3	30.1	35.5	33.0	40.0
	LLM	22.1	23.3	21.5	26.2	25.3	31.2	24.7	29.2	33.3	33.2	37.0
Touché-2020	Human	52.0	46.9	40.8	49.0	55.1	48.0	38.8	44.9	43.9	51.0	53.1
	LLM	57.1	46.9	49.0	41.8	38.8	37.8	27.6	44.9	55.1	58.2	51.0
CQADupStack	Human	28.1	24.3	24.3	29.9	23.1	33.6	32.9	30.5	36.0	33.9	35.7
	LLM	22.8	24.1	23.5	26.7	21.0	29.8	28.9	28.0	32.1	29.7	31.3
DBPedia	Human	34.1	54.3	49.0	46.8	56.6	61.8	57.4	57.8	60.5	60.8	37.0
	LLM	33.9	55.3	49.4	48.5	55.8	62.1	57.8	56.5	59.6	59.0	33.0
SCIDOCS	Human	16.0	13.5	11.3	14.4	15.8	16.7	14.2	16.1	17.2	18.8	19.6
	LLM	15.9	11.8	11.8	13.6	15.1	16.0	14.4	15.9	17.2	18.4	19.5
FEVER	Human	65.2	79.0	74.4	79.5	82.3	85.4	72.2	87.2	86.8	89.8	25.5
	LLM	63.0	79.1	73.3	80.0	82.3	85.6	74.5	86.5	87.1	89.2	24.2
Climate-FEVER	Human	25.9	26.8	27.2	29.5	32.1	32.8	28.1	33.6	34.3	36.0	35.9
	LLM	24.3	25.9	26.2	28.0	32.6	31.3	27.6	31.0	33.2	34.4	33.9
SciFact	Human	53.0	35.3	38.3	38.7	52.7	54.3	46.3	53.0	53.7	54.3	64.0
	LLM	56.3	35.3	39.7	39.3	50.7	53.0	46.7	51.7	52.7	53.0	63.7
NQ-UTD	Human	71.9	75.6	63.1	76.3	81.3	77.5	73.1	80.0	76.9	88.1	89.4
	LLM	73.1	75.0	68.8	77.5	81.3	76.3	71.3	76.9	78.8	89.4	88.1

Table 8: Performance comparison (NDCG@1) of retrieval models on Cocktail benchmark using the sole human-written or LLM-generated corpus.

Dataset	Matching Pairs	Random Pairs
MS MARCO	0.9802 (± 0.0138)	0.6725 (± 0.0369)
DL19	0.9798 (± 0.0145)	0.6733 (± 0.0373)
DL20	0.9817 (± 0.0133)	0.6728 (± 0.0362)
TREC-COVID	0.9875 (± 0.0088)	0.7502 (± 0.0436)
NFCorpus	0.9856 (± 0.0096)	0.7490 (± 0.0373)
NQ	0.9847 (± 0.0150)	0.6996 (± 0.0348)
HotpotQA	0.9905 (± 0.0123)	0.7073 (± 0.0366)
FiQA-2018	0.9657 (± 0.0243)	0.7216 (± 0.0372)
Touché-2020	0.9693 (± 0.0262)	0.7274 (± 0.0359)
CQADupStack	0.9583 (± 0.0336)	0.7360 (± 0.0311)
DBPedia	0.9900 (± 0.0121)	0.7023 (± 0.0373)
SCIDOCS	0.9863 (± 0.0112)	0.7444 (± 0.0352)
FEVER	0.9866 (± 0.0110)	0.6983 (± 0.0363)
Climate-FEVER	0.9873 (± 0.0107)	0.6992 (± 0.0366)
SciFact	0.9881 (± 0.0084)	0.7484 (± 0.0445)
NQ-UTD	0.9833 (± 0.0122)	0.7138 (± 0.0443)
Avg.	0.9816 (± 0.0150)	0.7135 (± 0.0376)

Table 9: Comparison of cosine similarity scores between matching and random pairs for each dataset.

select the following representative and state-of-the-art models:

BERT (Devlin et al., 2019), a foundational pre-trained language model, is commonly used as an encoder in dense retrieval systems. It was fine-tuned on the MSMARCO dataset using the official BEIR benchmark training script.

RoBERTa (Liu et al., 2019) builds on BERT’s

Dataset	Human Doc	LLM Doc	Equal
MS MARCO	0.0% (0.0%)	10.0% (0.0%)	90% (77.8%)
DL19	0.0% (0.0%)	5.0% (0.0%)	95.0% (94.7%)
DL20	5.0% (0.0%)	0.0% (0.0%)	95.0% (78.9%)
TREC-COVID	5.0% (0.0%)	5.0% (0.0%)	90% (61.1%)
NFCorpus	0.0% (0.0%)	0.0% (0.0%)	100.0% (75.0%)
NQ	0.0% (0.0%)	0.0% (0.0%)	100.0% (100.0%)
HotpotQA	0.0% (0.0%)	0.0% (0.0%)	100.0% (80.0%)
FiQA-2018	5.0% (0.0%)	0.0% (0.0%)	95.0% (68.4%)
Touché-2020	0.0% (0.0%)	0.0% (0.0%)	100.0% (70.0%)
CQADupStack	5.0% (0.0%)	5.0% (0.0%)	90% (83.3%)
DBPedia	0.0% (0.0%)	0.0% (0.0%)	100.0% (85.0%)
SCIDOCS	0.0% (0.0%)	0.0% (0.0%)	100.0% (100.0%)
FEVER	5.0% (0.0%)	0.0% (0.0%)	95.0% (84.2%)
Climate-FEVER	0.0% (0.0%)	0.0% (0.0%)	100.0% (90.0%)
SciFact	10.0% (0.0%)	5.0% (0.0%)	85.0% (64.7%)
NQ-UTD	0.0% (0.0%)	0.0% (0.0%)	100.0% (90.0%)
Avg.	2.2%	1.9%	95.9%

Table 10: Human evaluation of semantic relevance between human-written and LLM-generated documents across all datasets in the Cocktail benchmark.

success with more data and refined training techniques to enhance performance. RoBERTa was fine-tuned in the same manner to BERT, using the MSMARCO dataset.

ANCE (Xiong et al., 2020) improves dense retrieval by selecting challenging negatives across the corpus and updating the Approximate Nearest Neighbor (ANN) index asynchronously with each

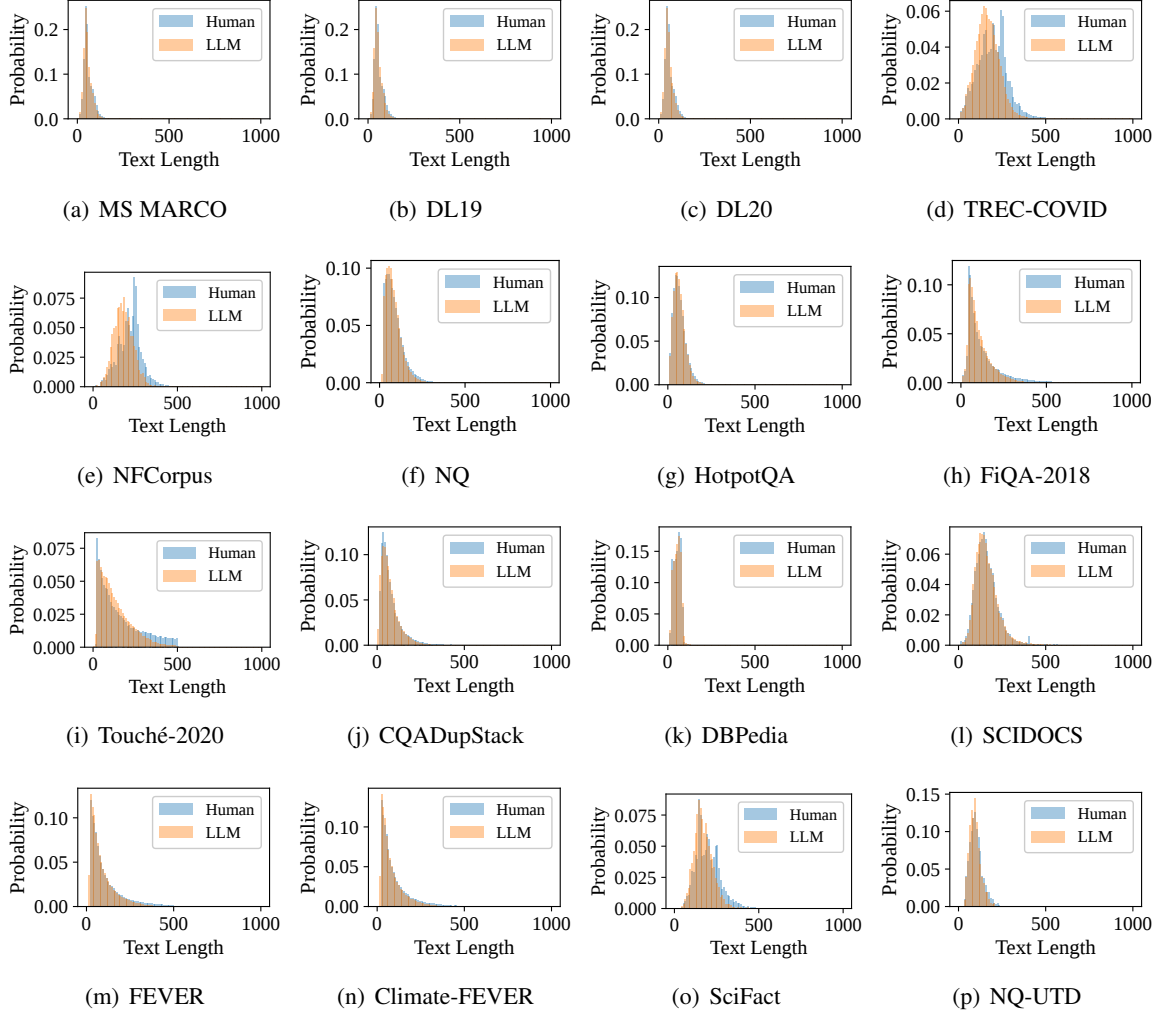


Figure 5: Distribution of text length of corpus for each dataset in Cocktail.

training iteration.

TAS-B (Hofstätter et al., 2021), a bi-encoder model, leverages balanced margin sampling for efficient query selection and dual supervision from a cross-encoder and a ColBERT model for enhanced learning.

Contriever (Izacard et al., 2022) employs contrastive learning with positive samples generated through cropping and token sampling, using MoCo to incorporate negatives from a queue of previous batches for self-supervised training.

coCondenser (Gao and Callan, 2022) undertakes a two-stage process beginning with pretraining and unsupervised contrastive loss for embedding generation, followed by supervised training with the pre-trained encoder.

RetroMAE (Xiao et al., 2022) utilizes a Masked Auto-Encoder approach for pretraining, enhancing sentence reconstruction from masked inputs to re-

fine language modeling.

DRAGON (Lin et al., 2023) applies progressive supervision and query augmentation for effective dense retrieval, suitable for both supervised and zero-shot settings.

Additionally, the following two widely used **neural re-ranking models** are also adopted in our evaluation framework:

CE (Wang et al., 2020) is a cross-encoder model pre-trained on MS MARCO dataset (Nguyen et al., 2016) through knowledge distillation from three teacher models: BERT-base (Devlin et al., 2019), BERT-large (Devlin et al., 2019), and ALBERT-large (Lan et al., 2019).

monoT5 (Raffel et al., 2020) is a sequence-to-sequence re-ranker based on T5 (Raffel et al., 2020), which is also pre-trained on the MS MARCO dataset.

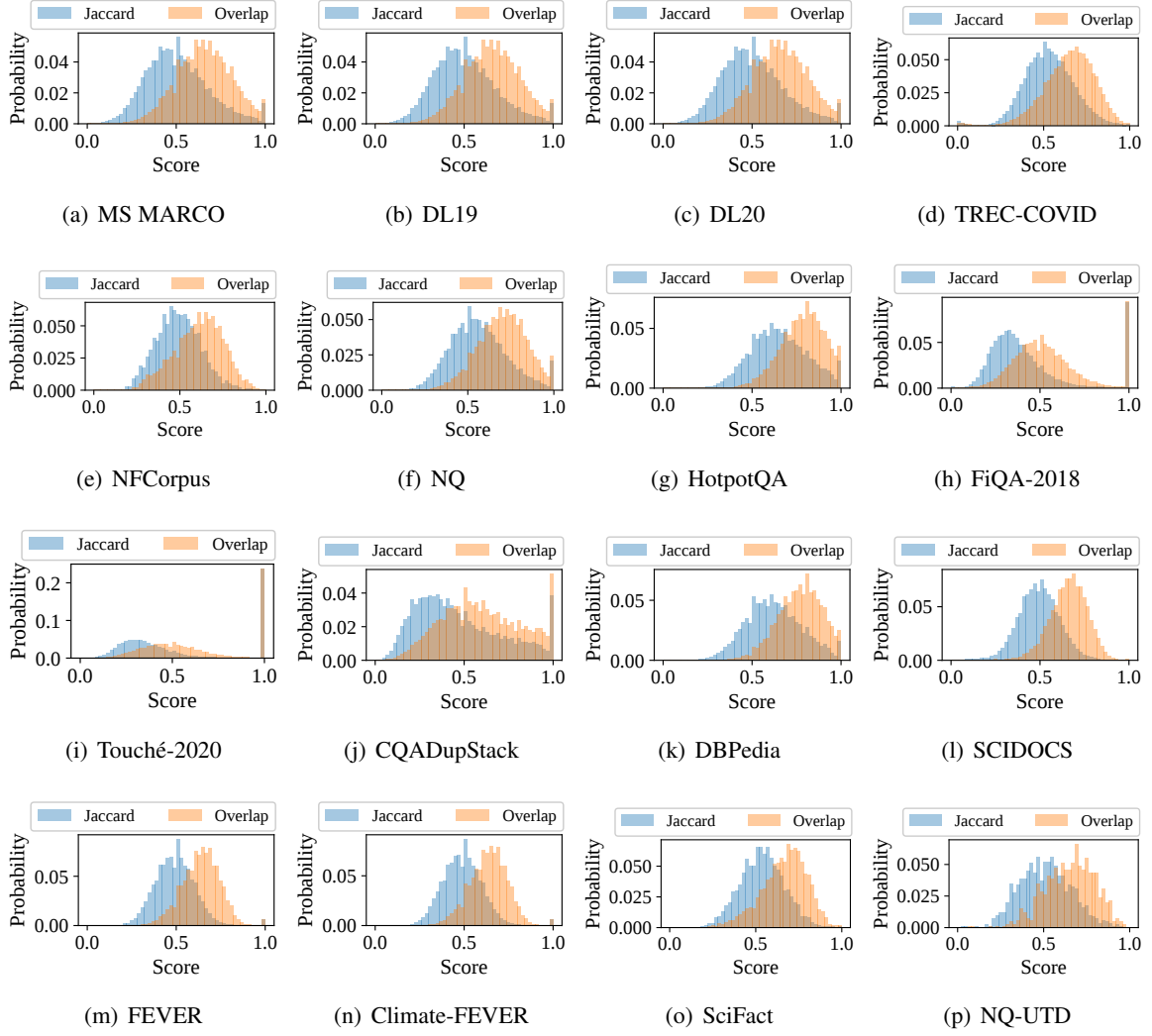


Figure 6: Distribution of term-based Jaccard similarity and overlap between LLM-generated and human-written corpora for each dataset in Cocktail.

B.2 Checkpoints and Implementation Details

For the fine-tuning of BERT and RoBERTa models, we utilized the official training script⁵ provided in the BEIR benchmark on the MSMARCO training dataset. Unless otherwise noted, mean pooling was employed as the pooling strategy, alongside cosine similarity as the scoring function. Training parameters for each model included a duration of 10 epochs, utilizing a batch size of 75, a learning rate set at $2e-5$, and the AdamW optimizer for efficient optimization. Our software framework utilizes PyTorch version 2.0.0 and the HuggingFace Transformers library version 4.31.0. All experiments were performed using approximately 100 hours on four NVIDIA RTX A6000 (48G) GPUs.

⁵https://github.com/beir-cellar/beir/blob/main/examples/retrieval/training/train msmarco_v3.py

For the other benchmarked neural retrieval models, we use the publicly available official pre-trained checkpoints, which are listed in Table 11. For all the neural models, only the first 512 word tokens of all documents are inputted.

C More Experimental Results

In addition to the findings presented through NDCG@1 in Table 2 and Table 3, we further extend our analysis to include the results for ranking performance and source bias for NDCG@3 and NDCG@5. These results are detailed in Table 13 and Table 14 for NDCG@3, and Table 15 and Table 16 for NDCG@5, respectively. Consistent with our earlier observations in Section 5.2, results on these metrics also reveal a similar trend, underscoring the robustness of our findings.

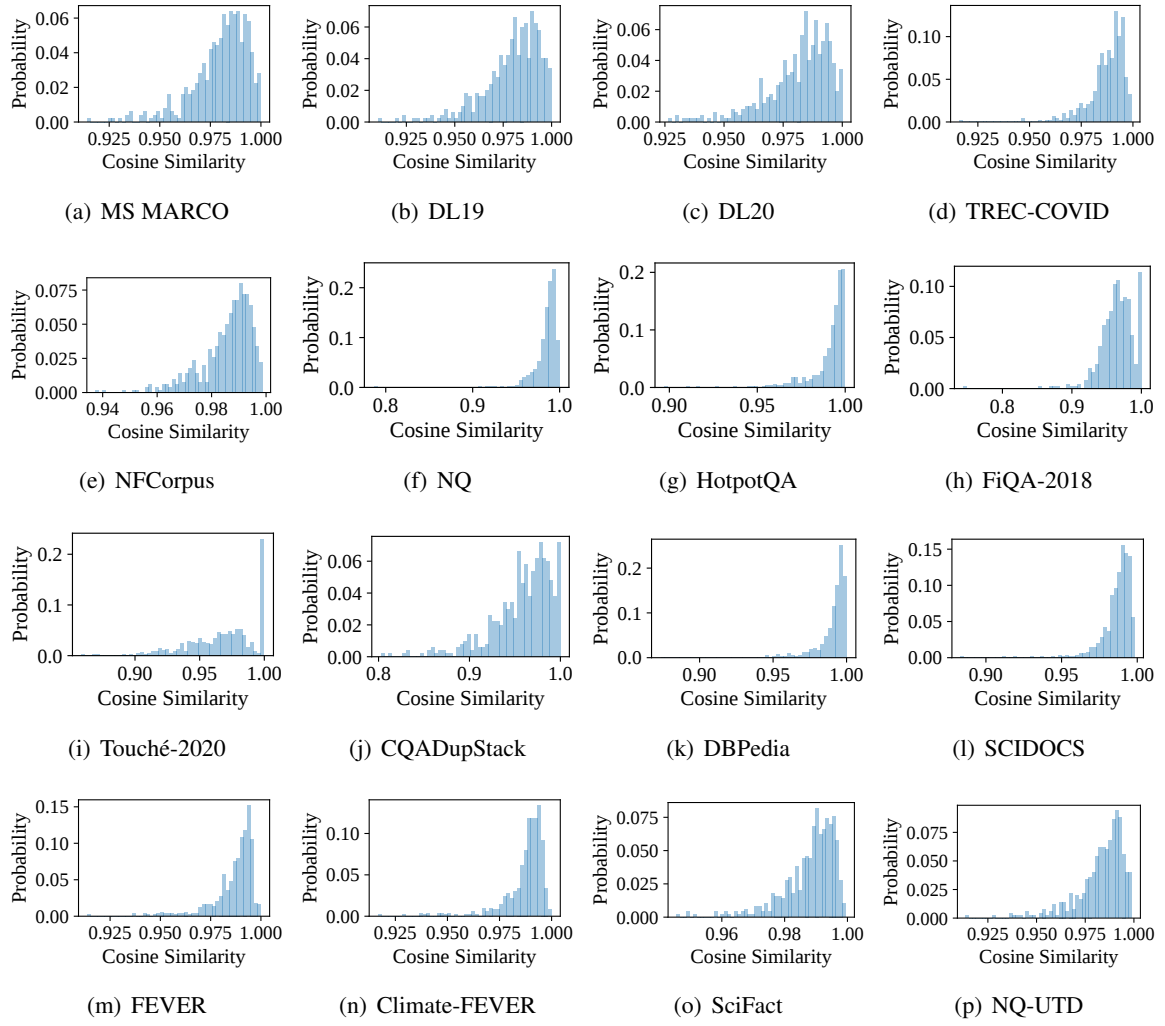


Figure 7: Distribution of cosine similarity of semantic embedding between LLM-generated and human-written corpora for each dataset in Cocktail.

Model Name	Publicly Available Link
BERT (mini)	https://huggingface.co/prajjwal1/bert-mini
BERT (small)	https://huggingface.co/prajjwal1/bert-small
BERT (base)	https://huggingface.co/bert-base-uncased
BERT (large)	https://huggingface.co/bert-large-uncased
RoBERTa	https://huggingface.co/FacebookAI/roberta-base
ANCE	https://huggingface.co/sentence-transformers/msmarco-roberta-base-ance-firstp
TAS-B	https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b
Contriever	https://huggingface.co/nthakur/contriever-base-msmarco
coCondenser	https://huggingface.co/sentence-transformers/msmarco-bert-co-condensor
RetroMAE	https://huggingface.co/nthakur/RetroMAE_BEIR
DRAGON (Query)	https://huggingface.co/nthakur/dragon-roberta-query-encoder
DRAGON (Doc)	https://huggingface.co/nthakur/dragon-roberta-context-encoder
CE	https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2
monoT5	https://huggingface.co/castorini/monot5-base-msmarco-10k

Table 11: Publicly available model checkpoints links used for evaluation in the Cocktail benchmark.

Dataset	Query	Relevant Human-Written Document	Relevant LLM-Generated Document
MS MARCO	do physicians pay for insurance from their salaries?	Active-duty physicians and their dependents will receive free or discounted health care coverage and dental coverage. Physicians in the Reserve and Guard can also participate in TRICARE Reserve Select, which is part of the Military's health care plan. Finally, physicians can receive up to \$400,000 in term life insurance coverage for only \$29 a month.	Active-duty physicians and their dependents are eligible for free or discounted healthcare coverage and dental coverage. In addition, physicians in the Reserve and Guard can enroll in TRICARE Reserve Select, which is part of the Military's healthcare plan. Furthermore, physicians can receive up to \$400,000 in term life insurance coverage for just \$29 per month.
DL19	axon terminals or synaptic knob definition	April 29, 2013. the terminal part of an axon from which a neural signal is rendered, via dispersion of a neurotransmitter, across a synapse to a nearby neuron. TERMINAL BUTTON: The terminal button is commonly referred to as the synaptic button, end button, button terminal, terminal bulb, and synaptic knob..	April 29, 2013. The terminal end of an axon, where a neural signal is transmitted to a nearby neuron through the diffusion of a neurotransmitter across a synapse. TERMINAL BUTTON: Also known as the synaptic button, end button, button terminal, terminal bulb, or synaptic knob.
DL20	average salary for dental hygienist in nebraska	In Nebraska, as the number of dental hygienists is growing, the salaries earned by dental hygienists are increasing. Dental hygienists earned a yearly mean salary of \$58,410 in 2006. They earned a yearly mean salary of \$64,440 in 2010.	In Nebraska, as the number of dental hygienists is increasing, so are their salaries. According to data from 2006, dental hygienists earned a yearly mean salary of \$58,410. By 2010, this figure had risen to \$64,440.
TREC-COVID	will SARS-CoV2 infected people develop immunity? Is cross protection possible?	Summary COVID-19 had a mild clinical course in patients with Agammaglobulinemia lacking B lymphocytes, whereas it developed aggressively in Common Variable Immune Deficiency. Our data offer mechanisms for possible therapeutic targets.	COVID-19 had a mild clinical course in patients with Agammaglobulinemia, a condition characterized by a lack of B lymphocytes, while it developed aggressively in Common Variable Immune Deficiency. Our data provide insights into potential therapeutic targets.
NFCorpus	What is Actually in Chicken Nuggets?	To study the origin and spread of Yersinia enterocolitica among pigs, fecal and blood samples were repeatedly taken on a fattening farm. A few piglets were found to be already infected on breeding farms. After the piglets were mixed, the infection spread through the whole unit. Eventually, all the pigs excreted the pathogen.	To investigate the origins and spread of Yersinia enterocolitica among pigs, fecal and blood samples were repeatedly collected from a fattening farm. Initially, a few piglets on breeding farms were found to be infected. Once these piglets were mixed, the infection rapidly spread throughout the entire unit, eventually infecting all the pigs.
NQ	how many episodes are in chicago fire season 4	The fourth season of Chicago Fire, an American drama television series with executive producer Dick Wolf, and producers Derek Haas, Michael Brandt, and Matt Olmstead, was ordered on February 5, 2015, by NBC.[1] and premiered on October 13, 2015 and concluded on May 17, 2016.[2] The season contained 23 episodes.[3]	The fourth season of Chicago Fire, a US drama television series created by executive producer Dick Wolf and producers Derek Haas, Michael Brandt, and Matt Olmstead, was greenlit by NBC on February 5, 2015, and premiered on October 13, 2015, concluding on May 17, 2016. The season consisted of 23 episodes.
HotpotQA	What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?	Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. In the film, two teenage girls cause their respective parents much concern when they start to become interested in boys. The parents' bickering about which girl is the worse influence causes more problems than it solves.	Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. In the film, two teenage girls, Corliss and her friend, cause their respective parents much distress when they start to develop romantic interests in boys. The parents' arguing over which girl is the worse influence creates more problems than it solves.
FiQA-2018	What are the ins/outs of writing equipment purchases off as business expenses in a home based business?	Keep this rather corny acronym in mind. Business expenses must be CORN: As other posters have already pointed out, certain expenses that are capital items (computers, furniture, etc.) must be depreciated over several years, but you have a certain amount of capital items that you can write off in the current tax year.	Keep this straightforward acronym in mind. Business expenses must be CLEAR: As other posters have already noted, certain expenses that are capital items (computers, furniture, etc.) must be depreciated over several years, but you have a certain amount of capital items that you can write off in the current tax year.
Touché-2020	Should teachers get tenure?	Why should teachers be laid off because of seniority? Many electives teachers got laid off. Even a Music Teacher got laid off even though she won an award for the best teacher in the valley! Ironic isn't it?	Teachers should not be laid off solely based on seniority. It is unfair that many elective teachers were let go, including a talented Music Teacher who recently won an award for being the best teacher in the valley. It is ironic that she was laid off despite her achievement.
CQADupStack	Is "a wide range of features" singular or plural?	I hope you can enlighten me. I get varying answers in Google and I need to find out which is the correct grammatical structure for these sentences. > The rest of the staff is/are on leave at the moment. > > The rest of my family is/are arriving late.	I hope you can enlighten me. I've been getting conflicting answers in Google, and I need to determine the correct grammatical structure for these sentences. The rest of the staff are on leave at the moment. The rest of my family are arriving late.
DBPedia	social network group selection	Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).	Clustering is the process of grouping a set of objects together based on their similarities, where objects within the same group (cluster) are more alike than those in other groups.
SCIDOCS	DeltaCFS: Boosting Delta Sync for Cloud Storage Services by Learning from NFS	The state machine approach is a general method for implementing fault-tolerant services in distributed systems. This paper reviews the approach and describes protocols for two different failure models—Byzantine and fail stop. Systems reconfiguration techniques for removing faulty components and integrating repaired components are also discussed.	The state machine approach is a versatile method for building fault-tolerant services in distributed systems. This paper examines the approach and outlines protocols for two distinct failure models—Byzantine and fail-stop. Additionally, the paper discusses techniques for reconfiguring systems to remove faulty components and integrate repaired ones.
FEVER	Konidela Production Company was established.	Konidela Production Company is an Indian film production company established by actor Ram Charan , son of Chiranjeevi .	Konidela Production Company is a renowned Indian film production house founded by the talented actor Ram Charan, the son of the legendary actor Chiranjeevi.
Climate-FEVER	Each of the six major past ice ages began when the atmospheric carbon dioxide content was far higher than at present.	The Geologic temperature record are changes in Earth 's environment as determined from geologic evidence on multi-million to billion (109) year time scales . The study of past temperatures provides an important paleoenvironmental insight because it is a crucial component of the climate and oceanography of the time .	The geologic temperature record reflects changes in Earth's environment as revealed through geologic evidence spanning millions to billions of years (109 years). The study of past temperatures offers valuable paleoenvironmental insights since it is a fundamental component of Earth's climate and oceanography during those time periods.
SciFact	Cataract and trachoma are the primary cause of blindness in Southern Sudan.	Background Blindness and low vision are thought to be common in southern Sudan. However, the magnitude and geographical distribution are largely unknown. We aimed to estimate the prevalence of blindness and low vision, identify the main causes of blindness and low vision, and estimate targets for blindness prevention programs in Mankien payam (district), southern Sudan.	Background blindness and low vision are believed to be prevalent in southern Sudan, but the scope and geographical distribution of these issues remain largely unexplored. Our study aimed to determine the prevalence of blindness and low vision, identify the primary causes of blindness and low vision, and estimate targets for blindness prevention programs in Mankien payam (district), southern Sudan.
NQ-UTD	Who won women's doubles champion in the 2023 Badminton World Tour Finals?	The women's doubles final of the BWF World Tour Finals 2023 concluded at the Hangzhou Olympic Sports Center on Sunday. China's top doubles pair, Chen Qingchen and Jia Yifan, defeated South Korea's Baek Ha Na and Lee So Hee 2-0 to successfully defend their title and win \$210,000 in prize money.	The women's doubles final of the BWF World Tour Finals 2023 came to a close at the Hangzhou Olympic Sports Center on Sunday. China's premier doubles pair, Chen Qingchen and Jia Yifan, outplayed South Korea's Baek Ha Na and Lee So Hee, winning the match 2-0 and successfully defending their title to collect a prize money of \$210,000.

Table 12: Examples of queries, relevant human-written documents, and relevant LLM-generated documents for each dataset in our Cocktail Benchmark.

Model (→)	Lexical BM25	Neural Retrievers								Neural Re-rankers			
		BERT	RoBERTa	ANCE	TAS-B	Contriever	coCondenser	RetroMAE	DRAGON	CE	monoT5		
PLM	-	BERT	RoBERTa	RoBERTa	DistilBERT	BERT	BERT	BERT	BERT	MiniLM	T5	Average	
# Paras	-	110M	125M	125M	66M	110M	110M	110M	110M	66M	220M	All	Neural
Supervised Evaluation (In-Domain Datasets Collected in Pre-LLM Era)													
MS MARCO	39.7	53.5	53.7	54.6	57.1	58.3	57.4	58.1	61.9	<u>59.0</u>	58.0	55.6	57.2
DL19	51.7	75.7	75.5	69.4	75.0	72.6	75.5	75.5	<u>76.6</u>	78.8	75.8	72.9	75.0
DL20	50.7	75.5	75.5	72.2	72.9	71.2	76.5	77.4	78.0	<u>76.9</u>	75.4	72.9	75.2
Zero-shot Evaluation (Out-of-Domain Datasets Collected in Pre-LLM Era)													
TREC-COVID	62.4	66.4	60.0	71.0	64.0	63.2	70.9	<u>74.3</u>	68.3	71.9	82.1	68.6	69.2
NFCorpus	44.2	36.3	32.3	33.0	39.1	42.1	41.1	39.3	41.1	48.2	<u>48.0</u>	40.4	40.1
NQ	46.2	62.5	60.0	60.3	66.3	<u>70.6</u>	66.0	67.9	71.2	68.6	69.1	64.4	66.3
HotpotQA	72.3	62.3	51.7	59.2	72.3	76.3	69.4	76.5	77.0	83.5	<u>81.4</u>	71.1	71.0
FiQA-2018	22.2	21.9	22.2	26.2	26.1	29.9	25.3	28.6	<u>33.5</u>	31.6	34.9	27.5	28.0
Touché-2020	51.4	47.0	42.1	46.8	44.7	46.3	34.3	47.1	53.4	<u>52.8</u>	52.5	47.1	46.7
CQADupStack	26.4	24.8	25.0	29.5	23.3	32.6	31.7	29.5	34.8	32.6	<u>34.4</u>	29.5	29.8
DBPedia	34.2	52.6	47.3	45.8	54.5	59.3	54.8	55.1	<u>57.7</u>	57.4	33.2	50.2	51.8
SCIDOCS	15.4	11.7	10.5	12.7	15.0	15.5	13.9	14.8	<u>16.1</u>	<u>17.4</u>	18.3	14.7	14.6
FEVER	67.5	80.5	75.8	80.7	83.8	86.5	75.5	87.7	<u>87.9</u>	89.1	44.6	78.1	79.2
Climate-FEVER	23.4	25.1	25.2	25.9	30.3	29.6	25.4	29.9	31.0	<u>32.7</u>	33.0	28.3	28.8
SciFact	<u>58.3</u>	37.9	40.9	40.8	54.0	56.4	49.9	54.5	56.7	57.1	65.4	52.0	51.4
Zero-shot Evaluation (Out-of-Domain Datasets Collected in the LLM Era)													
NQ-UTD	70.5	72.4	66.7	74.3	78.7	75.4	69.3	75.5	76.6	86.6	<u>86.2</u>	75.7	76.2
Averaged Result													
Supervised	47.4	68.2	68.2	65.4	68.3	67.4	69.8	70.3	72.2	<u>71.6</u>	69.7	67.1	69.1
Zero-shot	45.7	46.3	43.1	46.6	50.2	52.6	48.3	52.4	<u>54.3</u>	56.1	52.5	49.8	50.2
All	46.0	50.4	47.8	50.2	53.6	55.4	52.3	55.7	<u>57.6</u>	59.0	55.8	53.1	53.8

Table 13: Overall ranking performance (NDCG@3) across all benchmarked datasets in Cocktail. The second-to-last column is the average result across all models, while the last column is the average for all neural retrieval models. The **best performed** result for each dataset is marked in bold, and the second best is underlined.

Model (→)	Lexical BM25	Neural Retrievers								Neural Re-rankers			
		BERT	RoBERTa	ANCE	TAS-B	Contriever	coCondenser	RetroMAE	DRAGON	CE	monoT5		
PLM	-	BERT	RoBERTa	RoBERTa	DistilBERT	BERT	BERT	BERT	BERT	MiniLM	T5	Average	
# Paras	-	110M	125M	125M	66M	110M	110M	110M	110M	66M	220M	All	Neural
Supervised Evaluation (In-Domain Datasets Collected in Pre-LLM Era)													
MS MARCO	35.9	-4.2	-8.3	0.5	-8.5	-1.7	1.9	-5.3	-3.2	-4.0	0.9	0.4	-3.2
DL19	81.6	-50.6	-17.2	-21.0	-49.3	-21.9	-21.1	-47.6	-45.4	-19.5	-25.4	-21.6	-31.9
DL20	91.3	-59.3	-37.8	-7.7	-18.7	-10.6	-17.0	-31.1	-36.2	-21.3	-13.9	-14.8	-25.4
Zero-shot Evaluation (Out-of-Domain Datasets Collected in Pre-LLM Era)													
TREC-COVID	25.6	-51.0	-40.4	-33.8	-73.1	-62.0	-68.0	-25.9	-17.9	-66.5	-46.0	-41.7	-48.5
NFCorpus	-16.2	-16.8	-29.7	-19.1	-22.7	-43.2	-24.9	-18.9	-22.9	-38.4	-18.5	-24.7	-25.5
NQ	-3.9	-8.0	-3.8	-4.3	-12.4	-10.2	-8.5	-7.9	-14.7	-16.6	-8.7	-9.0	-9.5
HotpotQA	21.0	-0.5	-1.8	-5.7	0.2	-1.6	-2.9	1.4	-1.8	14.1	4.9	2.5	0.6
FiQA-2018	-3.6	-14.0	-5.8	-20.3	-10.7	-28.9	-19.0	-20.0	-15.1	-25.4	-17.2	-16.4	-17.6
Touché-2020	-32.1	-41.4	-40.2	-12.3	-9.7	-46.7	6.4	-18.2	-53.0	-69.9	-53.8	-33.7	-33.9
CQADupStack	17.8	-20.4	-15.7	-0.9	-1.7	-4.1	1.3	-20.7	-4.7	-7.5	3.5	-4.8	-7.1
DBPedia	21.3	-9.9	-14.1	-17.7	-5.1	-5.3	-10.8	-13.1	-10.9	-2.6	10.9	-5.2	-7.9
SCIDOCS	1.3	-1.7	-21.0	-1.6	-1.3	-6.5	-5.7	-10.8	-18.6	-12.7	-23.0	-9.2	-10.3
FEVER	-5.2	0.2	-0.2	-25.8	-4.5	-6.3	-8.0	0.6	-7.0	-0.7	6.2	-4.6	-4.5
Climate-FEVER	10.3	-7.4	-8.1	-38.4	-10.2	-6.9	-9.3	-4.6	-5.0	3.2	-37.9	-10.4	-12.5
SciFact	0.6	-12.5	-2.5	-5.9	-16.4	2.2	-10.3	-6.2	-7.5	1.3	-8.4	-6.0	-6.6
Zero-shot Evaluation (Out-of-Domain Datasets Collected in the LLM Era)													
NQ-UTD	9.0	-13.3	-14.2	-13.9	-9.7	-17.5	-10.7	-18.7	-20.6	-21.8	-9.0	-12.8	-14.9
Averaged Result													
Supervised	69.6	-38.0	-21.1	-9.4	-25.5	-11.4	-12.1	-28.0	-28.3	-14.9	-12.8	-12.0	-20.2
Zero-shot	3.5	-15.1	-15.2	-15.4	-13.6	-18.2	-13.1	-12.5	-15.4	-18.7	-15.2	-13.5	-15.2
All	15.9	-19.4	-16.3	-14.2	-15.9	-16.9	-12.9	-15.4	-17.8	-18.0	-14.7	-13.2	-16.2

Table 14: Overall source bias evaluation w.r.t. Relative Δ (NDCG@3) across all benchmarked datasets in Cocktail. The **numbers** (i.e., Relative $\Delta > 0$) suggest that retrieval models generally prefer human-written content while the **numbers** (i.e., Relative $\Delta \leq 0$) indicate retrieval models prefer LLM-generated content.

Model (→)	Lexical BM25	Neural Retrievers								Neural Re-rankers		Average	
		BERT	RoBERTa	ANCE	TAS-B	Contriever	coCondenser	RetroMAE	DRAGON	CE	monoT5		
PLM	-	BERT	RoBERTa	RoBERTa	DistilBERT	BERT	BERT	BERT	BERT	MiniLM	T5	Average	
# Paras	-	110M	125M	125M	66M	110M	110M	110M	110M	66M	220M	All	Neural
Supervised Evaluation (In-Domain Datasets Collected in Pre-LLM Era)													
MS MARCO	43.5	58.0	58.1	59.0	61.7	63.0	62.0	62.6	66.6	<u>62.7</u>	61.7	59.9	61.5
DL19	49.9	<u>75.3</u>	73.4	69.5	74.3	71.8	75.0	73.9	76.7	76.7	74.6	71.9	74.1
DL20	48.3	74.9	73.9	71.6	73.7	70.7	75.0	<u>76.0</u>	77.9	74.6	72.0	71.7	74.0
Zero-shot Evaluation (Out-of-Domain Datasets Collected in Pre-LLM Era)													
TREC-COVID	61.3	63.2	57.7	68.1	64.4	62.0	70.7	<u>74.4</u>	68.4	70.9	80.1	67.4	68.0
NFCorpus	41.5	34.3	30.0	30.1	37.5	39.7	38.2	37.3	39.7	<u>45.4</u>	45.6	38.1	37.8
NQ	49.3	64.8	62.3	62.8	68.9	<u>73.5</u>	68.7	70.4	73.7	70.2	70.9	66.9	68.6
HotpotQA	67.8	56.6	46.8	53.7	67.4	71.4	64.0	71.6	72.0	79.0	<u>77.8</u>	66.2	66.0
FiQA-2018	21.6	21.1	22.1	25.5	25.7	29.2	24.4	27.8	<u>32.3</u>	30.5	33.7	26.7	27.2
Touché-2020	<u>50.4</u>	45.6	40.9	48.7	42.9	43.7	32.6	44.4	50.1	50.2	52.0	45.6	45.1
CQADupStack	27.3	25.4	25.7	30.1	24.5	33.5	32.6	30.2	35.7	33.1	<u>34.8</u>	30.3	30.6
DBPedia	33.3	49.2	44.4	43.5	51.8	55.5	52.6	52.6	54.0	<u>55.1</u>	33.3	47.8	49.2
SCIDOCS	14.1	10.7	9.5	11.4	13.3	14.2	12.4	13.6	14.8	<u>15.8</u>	16.5	13.3	13.2
FEVER	71.4	82.5	78.1	82.4	85.8	88.1	78.5	89.2	<u>89.5</u>	89.9	48.1	80.3	81.2
Climate-FEVER	22.1	23.8	23.9	24.0	28.6	28.0	23.7	27.9	29.0	<u>31.3</u>	31.5	26.7	27.2
SciFact	<u>62.0</u>	41.0	43.0	42.9	56.4	59.6	53.0	57.1	59.9	61.2	68.2	54.9	54.2
Zero-shot Evaluation (Out-of-Domain Datasets Collected in the LLM Era)													
NQ-UTD	68.7	71.3	63.9	72.4	76.5	74.2	69.4	74.3	74.6	<u>83.9</u>	84.3	74.0	74.5
Averaged Result													
Supervised	47.2	69.4	68.5	66.7	69.9	68.5	70.7	70.8	73.7	<u>71.3</u>	69.4	67.8	69.9
Zero-shot	45.4	45.3	42.2	45.8	49.5	51.7	47.8	51.6	<u>53.4</u>	55.1	52.1	49.1	49.4
All	45.8	49.9	47.1	49.7	53.3	54.9	52.1	55.2	<u>57.2</u>	58.2	55.3	52.6	53.3

Table 15: Overall ranking performance (NDCG@5) across all benchmarked datasets in Cocktail. The second-to-last column is the average result across all models, while the last column is the average for all neural retrieval models. The best performance result for each dataset is marked in bold, and the second best is underlined.

Model (→)	Lexical BM25	Neural Retrievers								Neural Re-rankers		Average	
		BERT	RoBERTa	ANCE	TAS-B	Contriever	coCondenser	RetroMAE	DRAGON	CE	monoT5		
PLM	-	BERT	RoBERTa	RoBERTa	DistilBERT	BERT	BERT	BERT	BERT	MiniLM	T5	Average	
# Paras	-	110M	125M	125M	66M	110M	110M	110M	110M	66M	220M	All	Neural
Supervised Evaluation (In-Domain Datasets Collected in Pre-LLM Era)													
MS MARCO	30.5	-4.0	-6.6	1.0	-7.8	-1.0	2.2	-4.7	-2.6	-2.5	2.0	0.6	-2.4
DL19	59.5	-37.9	-27.7	-17.0	-39.3	-20.7	-16.5	-38.6	-48.1	-21.7	-23.0	-21.0	-29.0
DL20	74.1	-31.8	-34.1	-5.6	-13.7	-10.6	-14.0	-16.8	-27.8	-28.9	-14.3	-11.2	-19.8
Zero-shot Evaluation (Out-of-Domain Datasets Collected in Pre-LLM Era)													
TREC-COVID	16.6	-42.4	-37.1	-33.4	-62.7	-58.7	-56.9	-28.5	-23.4	-66.3	-32.7	-38.7	-44.2
NFCorpus	-15.8	-15.1	-19.4	-10.4	-13.6	-34.9	-13.8	-17.6	-16.4	-26.1	-12.3	-17.8	-18.0
NQ	-2.3	-6.7	-3.2	-2.2	-9.7	-8.8	-7.0	-5.8	-11.3	-14.8	-7.5	-7.2	-7.7
HotpotQA	16.5	-0.4	-1.4	-4.5	0.8	-0.7	-1.4	1.8	-1.1	10.2	4.6	2.2	0.8
FiQA-2018	-5.2	-14.1	-5.1	-12.2	-9.2	-18.8	-14.5	-18.2	-12.2	-17.8	-14.1	-12.9	-13.6
Touché-2020	-32.2	-35.3	-42.9	-19.8	-32.0	-34.1	-8.5	-19.1	-45.2	-55.1	-55.0	-34.5	-34.7
CQADupStack	15.5	-15.8	-11.0	-0.4	0.5	-3.8	3.6	-17.9	-2.5	-4.7	3.3	-3.0	-4.9
DBPedia	18.3	-11.8	-13.1	-17.7	-4.4	-2.2	-11.1	-9.0	-9.4	-3.9	13.9	-4.6	-6.9
SCIDOCS	1.4	-1.8	-16.3	-3.4	-4.4	-8.2	-7.9	-10.1	-13.2	-12.3	-17.8	-8.5	-9.5
FEVER	-4.2	0.3	0.2	-22.3	-3.9	-5.0	-6.1	0.4	-6.1	-0.8	4.9	-3.9	-3.8
Climate-FEVER	7.2	-4.8	-5.3	-37.2	-10.1	-4.6	-7.2	-4.6	-3.5	2.3	-28.3	-8.7	-10.3
SciFact	0.4	-10.9	-3.2	-4.0	-11.6	1.2	-8.7	-5.2	-7.0	-0.6	-6.4	-5.1	-5.6
Zero-shot Evaluation (Out-of-Domain Datasets Collected in the LLM Era)													
NQ-UTD	4.4	-12.2	-5.0	-7.7	-8.1	-9.0	-5.7	-16.4	-13.0	-16.3	-4.1	-8.5	-9.7
Averaged Result													
Supervised	54.7	-24.6	-22.8	-7.2	-20.3	-10.8	-9.4	-20.0	-26.2	-17.7	-11.8	-10.5	-17.1
Zero-shot	1.6	-13.2	-12.5	-13.5	-13.0	-14.4	-11.2	-11.6	-12.6	-15.9	-11.7	-11.6	-12.9
All	11.5	-15.3	-14.4	-12.3	-14.3	-13.7	-10.8	-13.1	-15.2	-16.2	-11.7	-11.4	-13.7

Table 16: Overall source bias evaluation w.r.t. Relative Δ (NDCG@5) across all benchmarked datasets in Cocktail. The **numbers** (i.e., Relative $\Delta > 0$) suggest that retrieval models generally prefer human-written content while the **numbers** (i.e., Relative $\Delta \leq 0$) indicate retrieval models prefer LLM-generated content.