ExcluIR: Exclusionary Neural Information Retrieval

Wenhao Zhang[1], **Mengqi Zhang**[1*], **Shiguang Wu**[1], **Jiahuan Pei**[2], **Zhaochun Ren**[3], **Maarten de Rijke**[4], **Zhumin Chen**[1*], **Pengjie Ren**[1]

[1] Shandong University, Qingdao, China
[2] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[3] Leiden University, Leiden, The Netherlands
[4] University of Amsterdam, Amsterdam, The Netherlands
{zhangwenhao,shiguang.wu}@mail.sdu.edu.cn,
{mengqi.zhang,chenzhumin,renpengjie}@sdu.edu.cn,
jiahuan.pei@cwi.nl, z.ren@liacs.leidenuniv.nl, m.derijke@uva.nl

Abstract

Exclusion is an important and universal linguistic skill that humans use to express what they do not want. There is little research on exclusionary retrieval, where users express what they do not want to be part of the results produced for their queries. We investigate the scenario of exclusionary retrieval in document retrieval for the first time. We present ExcluIR, a set of resources for exclusionary retrieval, consisting of an evaluation benchmark and a training set for helping retrieval models to comprehend exclusionary queries. The evaluation benchmark includes 3,452 high-quality exclusionary queries, each of which has been manually annotated. The training set contains 70,293 exclusionary queries, each paired with a positive document and a negative document. We conduct detailed experiments and analyses, obtaining three main observations: (i) existing retrieval models with different architectures struggle to comprehend exclusionary queries effectively; (ii) although integrating our training data can improve the performance of retrieval models on exclusionary retrieval, there still exists a gap compared to human performance; and (iii) generative retrieval models have a natural advantage in handling exclusionary queries.

1 Introduction

Selective attention (Treisman 1964; LaBerge 1990; Cherry 2020), defined as the ability to focus on relevant information while disregarding irrelevant information, plays a crucial role in shaping user's search behaviors. This principle, originating from cognitive psychology, not only shapes human perception of the environment but also extends its influence to interactions with information retrieval systems. When searching for information, users can express a desire to exclude certain information. We refer to this phenomenon as *exclusionary retrieval*, where users explicitly indicate their preference to exclude specific information.

Exclusionary retrieval emphasizes a crucial need for precision and relevance in information retrieval. It shows how users use their knowledge and expectations to find information that meets their specific needs. Therefore, the failure to understand exclusionary queries can present a potentially serious problem. For example, as shown in Figure 1, imagine a user searching for movies in the retrieval system. He poses a query like "What other sci-fi movies (besides Avengers: Endgame) were released in 2019?" If the retrieval system cannot correctly address this exclusionary requirement, it may return results containing content irrelevant to the user's interests (e.g., the movie "Avengers: Endgame"), thus reducing user satisfaction.

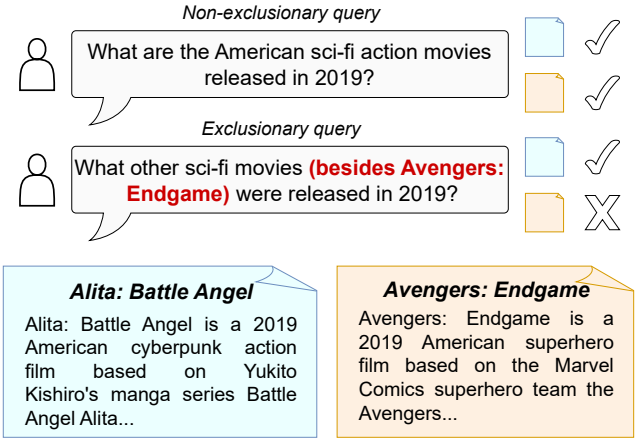

Figure 1: A comparison between non-exclusionary and exclusionary queries. Exclusionary queries often specify content to be excluded (e.g., "Avengers: Endgame") to express the user's requirements for omitting certain information. In this case, if the retrieval system fails to comprehend the exclusionary nature of a query (e.g., one containing the term "besides,") it will produce retrieval results that users do not desire.

Research on exclusionary retrieval remains relatively rare. Early studies mainly focus on keyword-based methods (Nakkouzi and Eastman 1990; McQuire and Eastman 1998; Harvey et al. 2003). These approaches rely on constructing boolean queries that include negation terms, which is essentially a post-processing strategy. However, these

*Corresponding author.

methods have limitations due to their reliance on structured queries, making them unsuitable for more diverse and complex natural language queries. Moreover, post-retrieval methods, such as rule-based filtering, are impractical in real-world applications, because they are difficult to optimize end-to-end with other models and can introduce potential side effects and instability to the final results. Although recent work has explored the impact of negation in modern retrieval models (Rokach, Romano, and Maimon 2008; Koopman et al. 2010; Weller, Lawrie, and Van Durme 2024), their focus is on comprehending the negation semantics within documents rather than the exclusionary nature of queries.

At present, there is no evaluation dataset to assess the capability of retrieval models in exclusionary retrieval. To address this gap, our first contribution in this paper is the introduction of the resources for exclusionary retrieval, namely ExcluIR. ExcluIR contains an evaluation benchmark to assess the capability of retrieval models in exclusionary retrieval, while also providing a training dataset that includes exclusionary queries. The dataset is built based on HotpotQA (Yang et al. 2018). We first use ChatGPT[1] to generate an exclusionary query for two given relevant documents, requiring that only one document contains the answer while explicitly rejecting information from the other document. Subsequently, we employ 17 workers to check each data instance in the benchmark to ensure data quality. The training set comprises 70,293 exclusionary queries, while the benchmark includes 3,452 human-annotated exclusionary queries. This dataset can evaluate whether retrieval models can correctly retrieve documents when dealing with exclusionary queries, providing a new perspective for evaluating retrieval models.

Our second contribution is to analyze the performance of existing retrieval methods with different architectures on exclusionary retrieval, including sparse retrieval (Robertson and Zaragoza 2009; Nogueira, Lin, and Epistemic 2019), dense retrieval (Karpukhin et al. 2020; Ni et al. 2022a), and generative retrieval methods (Bevilacqua et al. 2022; Wang et al. 2022a). We conduct extensive experiments and arrive at three main observations: (i) Existing retrieval models cannot fully understand the real intent of exclusionary queries; (ii) Generative retrieval models possess unique advantages in exclusionary retrieval, while late interaction models like ColBERT have obvious limitations in handling such queries; (iii) Fine-tuning the retrieval models with the training set of ExcluIR can improve the performance on exclusionary retrieval, but the results are still far from satisfactory. We provide in-depth analyses of these observations. These conclusions contribute valuable insights for future research on addressing the challenges of exclusionary retrieval. We share the benchmark and evaluation scripts on https://github.com/zwh-sdu/ExcluIR.

2 Dataset Construction

As depicted in Figure 2, the construction of the ExcluIR dataset involves the following steps: (i) we first extract document pairs from HotpotQA (Yang et al. 2018), where each data instance consisting of two interrelated documents; (ii) for each document pair, we employ ChatGPT to generate an exclusionary query. (iii) to enhance the diversity of the synthetic queries, we further use ChatGPT to rephrase them; and (iv) finally, to ensure a high quality of the dataset, we establish annotation guidelines and hire workers for manual correction.

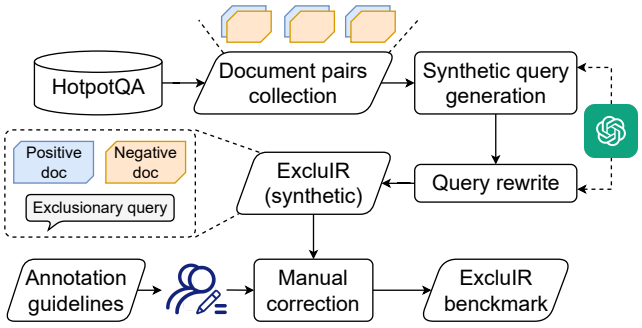

Figure 2: Overview of ExcluIR dataset construction process.

2.1 Collecting documents pairs

We begin the construction process by collecting documents from the HotpotQA (Yang et al. 2018) dataset, which is designed for multi-hop reasoning in question-answering task. Each data instance includes two supporting documents that are related. The model needs to comprehend the association between them and extract information from them to answer the question. We extract two related documents from each data instance to form our document pairs. In total, we collected 74,293 document pairs. After merging and removing duplicates, we obtained a document collection containing 90,406 documents.

2.2 Generating exclusionary queries

To efficiently construct our dataset, we design a prompt carefully to guide ChatGPT in generating exclusionary queries for each pair of documents. To ensure that the generated queries cover both positive and negative documents, we design a two-step construction strategy. Specifically, we first instruct ChatGPT to generate a query relevant to both documents, and then guide ChatGPT to revise this query by adding a constraint to include the semantics of refusal to information from the negative document.

2.3 Rewriting synthetic queries

Although the prompt has been carefully adjusted, the generated queries often express the exclusionary phrases in a limited manner, such as "excluding any information about," "except for any information," and "without referencing any information about." These expressions lack naturalness and deviate from real-world queries. To increase the diversity and naturalness of the queries, we further instruct ChatGPT to rephrase them. Then, we partition the ExcluIR dataset obtained in this step into training and test sets. The test set is further manually corrected to construct the benchmark, which we will describe next.

[1]https://platform.openai.com/docs/models/gpt-3-5

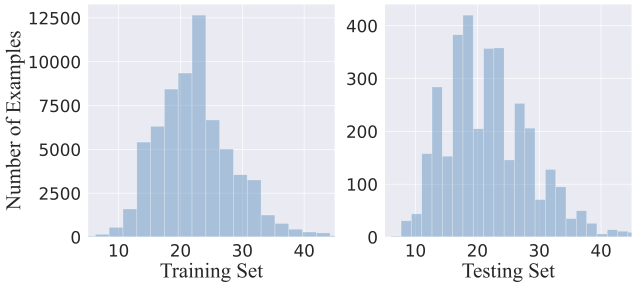

Figure 3: Distribution of the lengths of exclusionary queries in ExcluIR.

2.4 Manually correcting data

To build a reliable ExcluIR benchmark, we hire 17 workers for manual data correction. We first sample 4,000 instances from the 74,293 exclusionary queries obtained in the previous step. Each instance contains two documents along with a synthetic query generated by ChatGPT. We ask workers to check the synthetic exclusionary query to ensure its naturalness and correctness and they are encouraged to express the exclusionary nature of queries using diverse expressions. To facilitate the correction process, we construct an online correction system. In the system, we define three operations for workers to correct each data instance:

1. *Criteria Met*. If the synthetic query already meets the criteria, no further modifications are necessary.
2. *Query Modification*. If the synthetic query fails to meet the criteria, modify or rewrite the query to align with the requirements.
3. *Discard Data*. If it is difficult to write a query that meets the criteria based on these two documents, the workers can choose to discard the data.

2.5 Quality assurance

We take several measures to ensure data quality: (i) we provide detailed documentation guidelines, including task definition, correction process, and specific criteria for exclusionary queries; (ii) we present multiple examples of exclusionary queries to help workers understand the task and requirements; (iii) we record a video to demonstrate the entire correction process and emphasize the key considerations that need special attention; (iv) we adopt a real-time feedback mechanism to allow workers to share the issues they encounter during the correction process; we discuss these issues and provide solutions accordingly; and (v) we randomly sample 10% of the data of each worker for quality inspection. If there are errors in the sampled data, we will ask the worker to correct the data again.

2.6 Dataset statistics

Following the dataset construction process described above, we obtained 3,452 human-annotated entries for the benchmark and 70,293 exclusionary queries for the training set. The average word counts for exclusionary queries in the training set and benchmark are 22.37 and 21.64, respectively. To further investigate the diversity of data, we visualize the distribution of the lengths of exclusionary queries in Figure 3. We show that the lengths of exclusionary queries are diverse, reflecting varying levels of complexity and details.

3 Experimental Setup

Methods for comparison. To evaluate the performance of various retrieval models on exclusionary retrieval, we select three types of retrieval models with different architectures: sparse retrieval, dense retrieval, and generative retrieval.

Sparse retrieval methods calculate the relevance score of documents using term matching metrics such as TF-IDF (Robertson and Walker 1997).

- **BM25** (Robertson and Zaragoza 2009) is a classical probabilistic retrieval method based on the normalization of the frequency of the term and the length of the document.
- **DocT5Query** (Nogueira, Lin, and Epistemic 2019) expands documents by generating pseudo queries using a fine-tuned T5 model before building the BM25 index (Raffel et al. 2020).

Dense retrieval uses pre-trained language models (PLMs) as the backbones to represent queries and documents as dense vectors for computing relevance scores.

- **DPR** (Karpukhin et al. 2020) is a dense retrieval model based on dual-encoder architecture, which uses the representation of the [CLS] token of BERT (Devlin et al. 2019).
- **Sentence-T5** (Ni et al. 2022a) uses a fine-tuned T5 encoder model to encode queries and documents into dense vectors.
- **GTR** (Ni et al. 2022b) has the same architecture as Sentence-T5 and has been pretrained on two billion question-answer pairs collected from the Web.
- **ColBERT** (Khattab and Zaharia 2020) is a late interaction model that learns embeddings for each token in queries and documents, and then uses a MaxSim operator to calculate the relevance score.

Generative retrieval is an end-to-end retrieval paradigm.

- **GENRE** (De Cao et al. 2020) retrieves entities by generating their names through a seq-to-seq model, it can be applied to document retrieval by directly generating document titles. The original GENRE is trained based on BART as the backbone, and we reproduce it using T5.
- **SEAL** (Bevilacqua et al. 2022) retrieves documents by generating n-grams within them.
- **NCI** (Wang et al. 2022a) proposes a prefix-aware weight-adaptive decoder architecture, leveraging semantic document identifiers and various data augmentation strategies like query generation.

Evaluation metrics. For the original test queries, we report the commonly used metrics: Recall at rank N (R@N, $N = 1, 5, 10$) and Mean Reciprocal Rank at rank N (MRR@N,

Type	Model	HotpotQA				ExcluIR				
		R@2	R@5	R@10	MRR	R@1	MRR	ΔR@1	ΔMRR	RR
Sparse Retrieval	BM25	67.16	76.65	80.98	92.47	49.68	65.17	7.82	4.66	53.48
	DocT5Query	69.19	77.88	81.65	94.10	50.98	67.50	7.85	3.81	53.85
Dense Retrieval	DPR	55.53	67.44	73.49	81.73	49.63	65.79	7.34	5.01	54.02
	Sentence-T5	57.63	68.45	74.29	82.48	51.04	66.27	10.11	7.01	55.41
	GTR	61.82	73.57	79.42	85.50	54.87	70.88	14.40	8.79	57.42
	ColBERT	73.58	83.73	87.95	94.44	54.00	71.24	10.72	6.42	55.57
Generative Retrieval	GENRE	48.87	51.67	53.24	75.25	48.03	63.22	4.35	0.13	52.10
	SEAL	60.78	72.26	78.20	85.76	51.33	67.88	11.64	7.71	55.52
	NCI	47.60	58.14	64.37	74.59	37.22	51.37	1.97	2.29	50.93

Table 1: Performance of models trained on HotpotQA and tested on HotpotQA and ExcluIR. For the evaluation on HotpotQA, we report Recall@2 rather than Recall@1, since each query in HotpotQA has two supporting documents.

Type	Method	NQ320k				ExcluIR				
		R@1	R@5	R@10	MRR	R@1	MRR	ΔR@1	ΔMRR	RR
Sparse Retrieval	BM25	37.96	61.24	68.86	47.86	49.68	65.17	7.82	4.66	53.48
	DocT5Query	42.63	66.18	73.38	52.69	50.98	67.50	7.85	3.81	53.85
Dense Retrieval	DPR	54.81	79.50	85.52	65.39	48.55	60.50	16.45	13.49	58.76
	Sentence-T5	59.63	82.78	87.42	69.57	57.76	66.34	32.90	27.96	67.83
	GTR	62.35	84.67	89.17	71.90	59.79	69.00	34.85	28.12	68.31
	ColBERT	60.08	84.19	89.41	70.50	57.01	70.88	20.02	15.26	59.97
Generative Retrieval	GENRE	56.25	71.21	74.00	62.80	31.63	37.63	11.44	10.15	58.65
	SEAL	55.24	75.13	80.97	63.86	43.54	55.17	16.11	15.27	60.02
	NCI	60.41	76.10	80.19	67.18	31.46	38.95	15.87	16.81	56.84

Table 2: Performance of models trained on NQ320k and tested on NQ320k and ExcluIR.

$N = 10$). Recall measures the proportion of relevant documents that are retrieved in the top N results. MRR is the mean of the reciprocal of the rank of the first relevant document.

In ExcluIR, each exclusionary query q has a positive document d^+ and a negative document d^-. Thus, the difference between the rank of d^+ and the rank of d^- can reflect the retrieval model's capability of comprehending the exclusionary query. So we report $\Delta\text{R@}N$ and $\Delta\text{MRR@}N$, which can be formulated as:

$$\begin{aligned} \Delta\text{R@}N &= \text{R@}N(d^+) - \text{R@}N(d^-), \\ \Delta\text{MRR@}N &= \text{MRR@}N(d^+) - \text{MRR@}N(d^-). \end{aligned} \tag{1}$$

In addition, we report Right Rank (RR), which is the proportion of results where d^+ is ranked higher than d^-. The expected value of RR is 50% with random ranking.

4 Results and Analyses

In this section, we present four groups of experimental results and analyses to study: (i) the performance of the existing retrieval models on ExcluIR (Section 4.1), (ii) the strategy to improve the performance on ExcluIR, including incorporating our dataset into the training data (Section 4.2), and scaling up the model size (Section 4.3), (iii) the explanation for the superiority of generative retrieval in ExcluIR (Section 4.4).

4.1 How well do existing methods perform on ExcluIR?

To evaluate the performance of various retrieval models trained on existing datasets in ExcluIR, we conduct our experiments on two well-known standard retrieval datasets: Natural Questions (NQ) (Kwiatkowski et al. 2019) and HotpotQA (Yang et al. 2018). NQ is a large-scale dataset for document retrieval and question answering. The version we use is NQ320k, which consists of 320k query-document pairs. HotpotQA is a question-answering dataset that focuses on multi-hop reasoning. We split the original HotpotQA in the same way as our ExcluIR dataset, resulting in a 70k training set and a 3.5k test set.

The main performance of retrieval methods on the ExcluIR benchmark and other test data are presented in Table 1 and 2. We have the following observations from the results.

First, although these methods achieve good performance on the standard test data including HotpotQA and NQ320k, their performance on the ExcluIR benchmark is unsatisfactory. Nearly all models score less than 10% higher than random ranking on the RR metric. Despite the fact that the Sentence-T5 and GTR models trained on NQ320k achieve the highest ΔR@1/ΔMRR/RR scores, they are far

Model	Training Data	NQ320k				ExcluIR				
		R@1	R@5	R@10	MRR	R@1	MRR	ΔR@1	ΔMRR	RR
DPR	NQ320k	54.81	**79.50**	**85.52**	65.39	48.55	60.50	16.45	13.49	58.76
	N. w/ ExcluIR	**55.08**	79.31	85.49	**65.58**	**55.04**†	**67.89**†	**21.52**†	**16.38**†	**61.00**†
Sentence-T5	NQ320k	59.63	**82.78**	**87.42**	**69.57**	57.76	66.34	32.90	**27.96**	67.83
	N. w/ ExcluIR	**59.80**	81.58	87.13	69.36	**63.09**†	**74.57**†	**34.47**†	26.19	**68.00**†
GTR	NQ320k	**62.35**	**84.67**	**89.17**	**71.90**	59.79	69.00	34.85	28.12	68.31
	N. w/ ExcluIR	61.44	83.82	88.34	71.01	**65.64**†	**76.98**†	**39.05**†	**28.46**	**69.98**†
ColBERT	NQ320k	60.08	**84.19**	**89.41**	**70.50**	57.01	70.88	**20.02**	**15.26**	**59.97**
	N. w/ ExcluIR	**60.20**	83.59	88.60	70.29	**57.91**	**73.52**†	19.30	13.05	59.71
GENRE	NQ320k	**56.25**	**71.21**	**74.00**	**62.80**	31.63	37.63	11.44	10.15	58.65
	N. w/ ExcluIR	55.15	70.00	72.85	61.55	**65.67**†	**73.01**†	**41.19**†	**20.31**†	**70.48**†
SEAL	NQ320k	**55.24**	**75.13**	**80.97**	**63.86**	43.54	55.17	16.11	15.27	60.02
	N. w/ ExcluIR	53.86	74.84	80.34	62.78	**70.39**†	**78.40**†	**52.14**†	**43.25**†	**78.02**†
NCI	NQ320k	60.41	76.10	80.19	67.18	31.46	38.95	15.87	16.81	56.84
	N. w/ ExcluIR	**60.61**	**76.53**	**80.55**	**67.46**	**56.92**†	**64.67**†	**41.13**†	**39.92**†	**72.97**†

Table 3: The results of the impact of augmenting NQ320k with the ExcluIR training set. † indicates significant improvements with p-value < 0.05.

from achieving ideal performance. This is attributed to the fact that negative documents are erroneously retrieved and ranked high, indicating that these models fail to comprehend the exclusionary nature of queries.

Second, the diversity of training data impacts the model's ability to comprehend exclusionary queries. As can be seen from Table 1 and 2, the models trained on NQ320k exhibit better performance on ExcluIR than those trained on HotpotQA. We consider that this is because the queries in NQ320k are more diverse and contain more exclusionary queries. Therefore, increasing the domain and diversity of training data can be beneficial for exclusionary retrieval. To further investigate how expanding the training data influences performance, we conducted additional experimental analyses. We have conducted further experimental analysis in Section 4.2.

Additionally, as expected, sparse retrieval methods demonstrate a significant limitation in comprehending the exclusionary nature of queries, so they have almost no ability to handle ExcluIR. As shown in Table 2, the RR scores of BM25 and DocT5Query are only 53.48% and 53.85%, which are only slightly higher than random. Their ΔR@1 and ΔMRR scores are lower than most neural retrieval models trained on NQ320k. This is an expected result, because these methods are based on a lexical match between queries and documents. This limitation prevents them from focusing on the exclusionary phrases in the query, instead leading to a high relevance score for negative documents.

Furthermore, we also evaluate the performance of additional models trained on different datasets in ExcluIR. Due to space limitations, these results are presented in appendix[2].

[2]Appendix is available at https://arxiv.org/abs/2404.17288

4.2 How does incorporating our dataset into training data affect the performance?

Previous experiments have demonstrated that models trained on HotpotQA and NQ320k perform unsatisfactorily on ExcluIR. We believe that this is partly due to a lack of exclusionary queries in the training data. Therefore, in this section, we incorporate the ExcluIR training set into the training data to assess its impact on performance. The results of augmenting NQ320k with the ExcluIR training set are presented in Table 3. Due to space limitations, the results of augmenting HotpotQA are included in appendix. For ease of analysis, we have summarized the results from both tables in Figure 4. From the results, we have three main observations.

First, merging the ExcluIR training set into the training data can enhance most models' ability to comprehend exclusionary queries. Additionally, the performance of all generative retrieval models on ExcluIR has significantly improved. For instance, with NQ320k as the original dataset, SEAL achieves 18% improvement (60.02% vs. 78.02%) in RR by integrating the ExcluIR training set, with only a small (1.08%) decrease (63.86% vs. 62.78%) in performance on the original test data. This is because the ExcluIR training set contains a large number of exclusionary queries, which can help the retrieval model to comprehend the exclusionary nature of queries better.

Second, when training data contain exclusionary queries, generative retrieval models are better at handling exclusionary retrieval task compared to dense retrieval models. As shown in Figure 4, although dense retrieval models trained on two original datasets perform better on ExcluIR, augmenting with the ExcluIR training set leads to a greater improvement in generative retrieval models, ultimately surpassing dense retrieval methods overall. On average, generative retrieval models achieve a 17.75% improvement, in con-

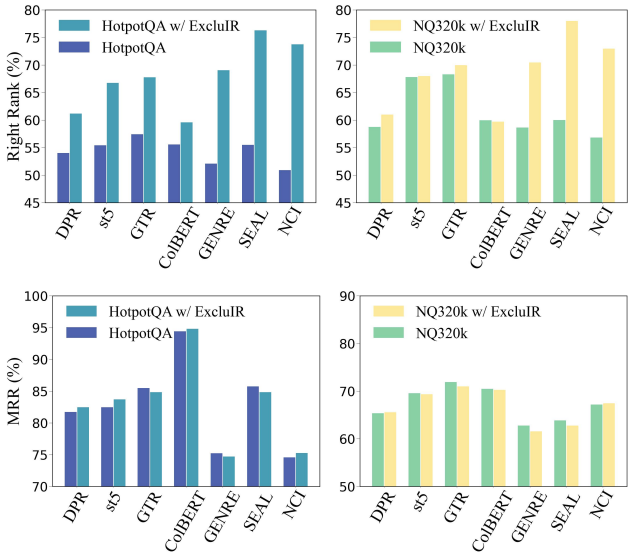

Figure 4: Performance of models under different training data settings. The upper figures show the RR score of various models on the ExcluIR benchmark, and the lower figures show the performance of these models on HotpotQA and NQ320k. The different colors of the bars represent different training data.

Training set	Model	Base	Large
HotpotQA	DPR	54.02	54.25 ↑
	Sentence-T5	55.41	53.78 ↓
	GENRE	52.10	49.01 ↓
	NCI	50.93	50.64 ↓
HotpotQA w/ ExcluIR	DPR	61.19	62.63 ↑
	Sentence-T5	66.75	69.01 ↑
	GENRE	69.07	70.96 ↑
	NCI	73.75	73.61 ↓
NQ320k	DPR	58.76	61.62 ↑
	Sentence-T5	67.83	69.02 ↑
	GENRE	58.65	55.82 ↓
	NCI	56.84	62.54 ↑
NQ320k w/ ExcluIR	DPR	61.00	63.47 ↑
	Sentence-T5	68.00	69.65 ↑
	GENRE	70.48	72.86 ↑
	NCI	72.97	74.45 ↑

Table 4: RR scores with different model sizes on ExcluIR. ↑ indicates that an increase in model size improves performance, while ↓ indicates the opposite.

trast to the average 4.77% improvement observed in dense retrieval models. This is because the generative retrieval model is more suitable for capturing the complex relationships between queries and documents in terms of model architecture. We present a more detailed analysis in Section 4.4.

Third, ColBERT fails to achieve satisfactory performance, even after fine-tuning on ExcluIR. Among the models trained with the ExcluIR training set, ColBERT exhibits the lowest performance. This is because ColBERT calculates the document relevance score based on token-level matching, leading it to overlook exclusionary phrases in queries, which is crucial for exclusionary retrieval. We have visualized the relevance calculation of ColBERT to further understand its performance in appendix.

Moreover, we consider that a model trained only on our dataset would perform well on ExcluIR but poorly on HotpotQA and NQ320k. This is because the diversity of training data is crucial for training a powerful retrieval model. We have conducted preliminary experiments on Sentence-T5 to confirm this, due to space limitations, the results are presented in appendix. The results indicate that the model trained only on our dataset struggles to perform well on standard retrieval datasets due to the lack of general training data, so we didn't conduct more experiments in this setting.

4.3 How does model size affect performance?

To analyze the impact of model size on the performance of ExcluIR, we increase model sizes of DPR, sentence-t5, GENRE, and NCI, and train them on different datasets. Specifically, for DPR, we use two variants: bert-base-uncased and bert-large-uncased. For sentence-t5, GENRE, and NCI, we adopt t5-base and t5-large.

The results are presented in Table 4. We note that increasing the model size generally improves performance on ExcluIR when the training data includes exclusionary queries. This is consistent with observations by Ravichander, Gardner, and Marasović (2022), who show that larger models are better at understanding the implications of negated statements in documents.

However, when training on original datasets, increasing the model size does not always lead to improved performance on ExcluIR. We conducted additional experiments on more models. The results indicated that the performance of stsb-roberta-large decreases compared to stsb-roberta-base. This indicates that simply increasing model size cannot solve the challenges of exclusionary retrieval, we should investigate building more training data and proposing new training strategies.

4.4 Why are generative retrieval models superior in ExcluIR?

Generative retrieval models have inherent advantages in comprehending exclusionary queries. We try to analyze and explain the reason based on the architecture of generative models.

First, as a comparison, we show that bi-encoder models have a representation bottleneck for exclusionary queries. When two documents are similar but have some differences that the user would like to distinguish, it is difficult to ensure that the vector representation of the query remains distant from the negative document while closely aligning with the positive document. This representation bottleneck prevents the model from correctly comprehending the true intent of the query. We present this proof in appendix.

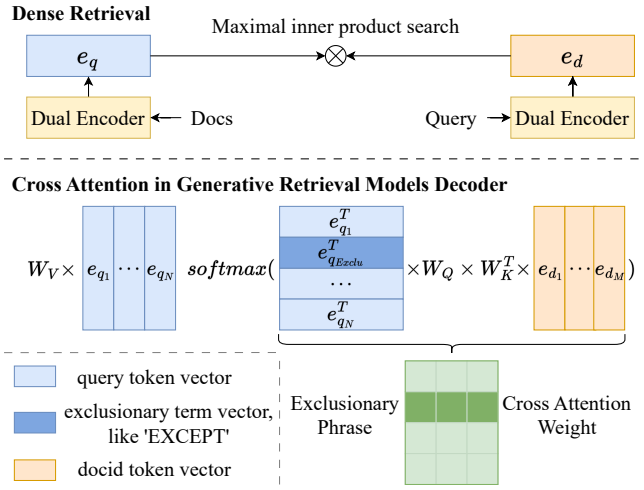

Figure 5: Summary of the analysis that shows the differences between dense retrieval and generative retrieval models in handling ExcluIR.

Generative retrieval models adopt a sequence-to-sequence framework, such as T5 or BART, which estimates the probability of generating the document IDs given the query using a conditional probability model: $P(d|q)$. When generating document IDs, multiple cross-attention layers in the decoder can capture the token-level semantic information in the query, a phenomenon also explored by Wu et al. (2024). Assuming the decoder consists of L layers, for the l-th layer ($0 \leq l < L$), the cross-attention layer is given by:

$$S^{(l+1)} = \text{softmax}\left(\frac{Q^{(l)}K^{(l)T}}{\sqrt{d_k}}\right)V^{(l)}, \quad (2)$$

where $Q^{(l)} = W_q^{(l)}S^{(l)}$, $K^{(l)} = W_k^{(l)}H_q^{(l)}$, $V^{(l)} = W_v^{(l)}H_q^{(l)}$, and $H_q^{(l)} = [e_{q_1}, \cdots, e_{q_N}]$ are query token vectors generated by encoder, $S^{(l)} = [e_{d_1}, \cdots, e_{d_M}]$ are generated embedding vectors for docid tokens at l-th layer, $W_q^{(l)}$, $W_k^{(l)}$ and $W_v^{(l)}$ are learnable cross-attention weight matrices. We visualize the cross-attention architecture in generative models to summarize our analysis. As shown in Figure 5, the multi-level cross-attention mechanism allows the model to strongly focus on key terms in the query, including exclusionary phrases (highlighted in dark green). Thus, when faced with queries with complex semantics, generative retrieval models are capable of effectively capturing the query intent.

Notably, this architectural advantage is also present in cross-encoders, such as the classic BERT re-ranker. We have evaluated the performance of cross-encoder models on ExcluIR, the results are presented in appendix. It can be seen that, within the zero-shot setting, a strong cross-encoder model outperforms both dense retrieval and generative retrieval models on ExcluIR. This result is expected, as the cross-encoder calculates the similarity between a query and a document individually, allowing it to better understand the relation between the query and the document. However, employing such models for retrieval from the entire corpus is time-prohibitive, which is why we excluded them from the main experiments.

5 Related Work

Early studies in exclusionary retrieval primarily focus on keyword-based methods. These approaches typically treat user queries as logical expressions of boolean operations (Nakkouzi and Eastman 1990; Strzalkowski 1995; McQuire and Eastman 1998; Harvey et al. 2003). However, these methods depend on explicit and deterministic rules, lack the flexibility to handle subtle exclusions, and are not suitable for more realistic retrieval scenarios.

In addition, there is a task related to exclusionary retrieval, known as argument retrieval (Wachsmuth, Syed, and Stein 2018), which aims to retrieve the best counterargument for a given argument on any controversial topic. While argument retrieval implicitly requires the model to find the counterargument to the query, the intention of exclusion is not explicitly expressed in the query. Wang et al. (2022b) first investigate exclusionary retrieval in Text-to-Video Retrieval (T2VR). They demonstrate that existing video retrieval models performed poorly when dealing with queries like "find shots of kids sitting on the floor and not playing with the dog." To the best of our knowledge, there has been no research on exclusionary retrieval in document retrieval.

(Weller, Lawrie, and Van Durme 2024) introduce NevIR, a benchmark designed to assess the ability of neural information retrieval systems to handle negation. NevIR requires retrieval models to rank two documents that differ only in negation, where both documents remain consistent in all other aspects except the key negation. Similarly, Rokach, Romano, and Maimon (2008); Koopman et al. (2010) investigate the impact of negation contexts within documents on retrieval performance. For example, a search for "headache" might retrieve patient records containing "the patient has no symptoms of headache." Our work is different as we focus on exclusionary retrieval, studying whether the retrieval model can comprehend the intent of exclusionary queries.

6 Conclusion

In this work, we focus on a common yet understudied retrieval scenario called exclusionary retrieval, where users explicitly express which information they do not want to obtain. We have provided the community with a new benchmark, named ExcluIR, which focuses on exclusionary queries that explicitly express the information users do not want to obtain. We have conducted extensive experiments that demonstrate that existing retrieval methods with different architectures perform poorly on ExcluIR. Notably, ExcluIR cannot be solved by simply adding training data domains or increasing model sizes. Additionally, our analyses indicate that generative retrieval models inherently excel at comprehending exclusionary queries compared with sparse and dense retrieval models. We hope that this work can inspire future research on ExcluIR.

Acknowledgements

This work was supported by the Key R&D Program of Shandong Province with grant 2024CXGC010108, the Natural Science Foundation of China (62472261, 62102234, 62372275, 62272274, 62202271, T2293773, 62072279), the National Key R&D Program of China with grant No.2022YFC3303004, the Natural Science Foundation of Shandong Province (ZR2021QF129), and by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union's Horizon Europe program under grant agreement No 101070212. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Bevilacqua, M.; Ottaviano, G.; Lewis, P.; Yih, S.; Riedel, S.; and Petroni, F. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems*, 35: 31668–31683.

Cherry, K. 2020. How we use selective attention to filter information and focus. Verywell Mind.

De Cao, N.; Izacard, G.; Riedel, S.; and Petroni, F. 2020. Autoregressive Entity Retrieval. In *ICLR 2021-9th International Conference on Learning Representations*, volume 2021. ICLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Harvey, V. J.; Baugh, J. M.; Johnston, B. A.; Ruzich, C. M.; Grant, A. J.; et al. 2003. The challenge of negation in searches and queries. *Review of Business Information Systems (RBIS)*, 7(4): 63–76.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781.

Khattab, O.; and Zaharia, M. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 39–48.

Koopman, B.; Bruza, P.; Sitbon, L.; and Lawley, M. 2010. Analysis of the effect of negation on information retrieval of medical data. In *Proceedings of 15th Australasian Document Computing Symposium*, 89–92. School of Computer Science and IT, RMIT University.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

LaBerge, D. L. 1990. Attention. *Psychological Science*, 1(3): 156–162.

McQuire, A. R.; and Eastman, C. M. 1998. The ambiguity of negation in natural language queries to information retrieval systems. *Journal of the American Society for Information Science*, 49(8): 686–692.

Nakkouzi, Z. S.; and Eastman, C. M. 1990. Query formulation for handling negation in information retrieval systems. *Journal of the American Society for Information Science*, 41(3): 171–182.

Ni, J.; Abrego, G. H.; Constant, N.; Ma, J.; Hall, K.; Cer, D.; and Yang, Y. 2022a. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, 1864–1874.

Ni, J.; Qu, C.; Lu, J.; Dai, Z.; Abrego, G. H.; Ma, J.; Zhao, V.; Luan, Y.; Hall, K.; Chang, M.-W.; et al. 2022b. Large Dual Encoders Are Generalizable Retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9844–9855.

Nogueira, R.; Lin, J.; and Epistemic, A. 2019. From doc2query to docTTTTTquery. *Online preprint*, 6: 2.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Ravichander, A.; Gardner, M.; and Marasović, A. 2022. CONDAQA: A Contrastive Reading Comprehension Dataset for Reasoning about Negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8729–8755.

Robertson, S.; and Zaragoza, H. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333–389.

Robertson, S. E.; and Walker, S. 1997. On relevance weights with little relevance information. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, 16–24.

Rokach, L.; Romano, R.; and Maimon, O. 2008. Negation recognition in medical narrative reports. *Information Retrieval*, 11: 499–538.

Strzalkowski, T. 1995. Natural language information retrieval. *Information Processing & Management*, 31(3): 397–417.

Treisman, A. M. 1964. Selective attention in man. *British Medical Bulletin*, 20(1): 12–16.

Wachsmuth, H.; Syed, S.; and Stein, B. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 241–251.

Wang, Y.; Hou, Y.; Wang, H.; Miao, Z.; Wu, S.; Chen, Q.; Xia, Y.; Chi, C.; Zhao, G.; Liu, Z.; et al. 2022a. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems*, 35: 25600–25614.

Wang, Z.; Chen, A.; Hu, F.; and Li, X. 2022b. Learn to understand negation in video retrieval. In *Proceedings of the*

30th ACM International Conference on Multimedia, 434–443.

Weller, O.; Lawrie, D.; and Van Durme, B. 2024. NevIR: Negation in Neural Information Retrieval. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2274–2287. St. Julian's, Malta: Association for Computational Linguistics.

Wu, S.; Wei, W.; Zhang, M.; Chen, Z.; Ma, J.; Ren, Z.; de Rijke, M.; and Ren, P. 2024. Generative Retrieval as Multi-Vector Dense Retrieval. *arXiv preprint arXiv:2404.00684*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380.