Leveraging Large Language Models for Adversarial Attacks on Information Retrieval Systems

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated exceptional proficiency in generating responses to diverse user queries and prompts. Recent studies have shown that synthetic test collections generated by LLMs are at least as effective at training and evaluating ranking models as existing collections like MS MARCO, which are based on text and relevance judgments from humans. In this paper, we harness the capabilities of LLMs to generate adversarial attacks against information retrieval systems by introducing counterfactual documents into corpora. We prompt LLMs to generate these counterfactual documents, which we call "`evil-twin`" documents, from a combination of queries and factually correct documents that are known to be relevant to these queries. The `evil-twin` documents deliberately contain disinformation that mirrors and refutes information contained in their associated "`good-twin`" documents. To evaluate our approach we employ various neural ranking models to re-rank `good-twin` and `evil-twin` documents, demonstrating that `evil-twin` documents can achieve higher positions in rankings, thereby increasing the likelihood that a searcher will be exposed to the disinformation they contain. Because we use a variety of factually correct documents as mirror images for the `evil-twin` documents, their content is more diverse than disinformation generated by LLMs prompted with queries alone.

1 Introduction

Ensuring the integrity and accuracy of the results presented to searchers by information retrieval systems is crucial, particularly in sensitive domains like health and politics. Despite recent advances in Neural Ranking Models (NRMs), studies have shown that these methods still suffer from a lack of robustness and are vulnerable to adversarial attacks (Raval and Verma, 2020; Wu et al., 2023; Thorne and Vlachos, 2019; Liu et al., 2023a). These attacks, commonly known as black-hat Search Engine Optimization (SEO) or web spamming are designed to find human-imperceptible perturbations to maliciously manipulate target documents to deceive the ranking algorithm to rank targeted document in a higher ranking position, increasing the probability that searchers will be exposed to the malicious content they contain (Morahan-Martin and Anderson, 2000; Fernández-Pichel et al., 2022). To ensure integrity and accuracy, information retrieval systems must be robust to these attacks.

In the past, adversarial attacks might take the form of term spamming, which involves the intentional insertion of a cluster of query-related keywords into a targeted document through term repetition, with the hope of deceiving a retrieval system to rank the target document in a higher/better ranking position (Imam and Vassilakis, 2019; Castillo et al., 2011; Sasaki and Shinnou, 2005). While these methods can deceive ranking models, spam detection tools can generally detect and filter term spamming and other simplistic attacks, protecting searchers from exposure to them. While humans can also be involved in the creation of spam documents and disinformation, cost provides a natural limit to its scale (Spirin and Han, 2012; Lau et al., 2012).

Recently large language models (LLMs) have begun to replace humans over a range of tasks, including information retrieval related tasks such as query expansion (Wang et al., 2023), document expansion (Askari et al., 2023b), relevance assessment (Faggioli et al., 2023; Thomas et al., 2023), and test collection generation (Askari et al., 2023a; Alaofi et al., 2023). For example, Arabzadeh et al. (2024) demonstrate that texts generated by different LLMs exhibit a high level of comparability with gold standard datasets of test collections in terms of both quality and accuracy. In a separate study, Askari et al. (2023a) explore the efficacy

of documents generated by LLMs in training neural ranking models. The outcomes of their experiments reveal that neural ranking models trained on content generated by LLMs tend to outperform those trained on human-generated content in out-of-domain ranking scenarios.

Motivated by this prior research on utilizing LLMs for document generation in response to queries, this paper aims to leverage their language generation capabilities for crafting adversarial attacks on information retrieval systems. Unlike previous studies that focus on attacking already existing documents, our approach involves introducing counterfactual documents into corpora, presenting convincing and seemingly credible disinformation in response to a given target search query. Our objective is to create counterfactual documents containing misleading information that rank higher than authentic and factually accurate ones, increasing the probability that a searcher entering a target query will be exposed to disinformation.

One approach is simply to prompt the LLM to generate a counterfactual document for the target query. However, we may want to generate more than one counterfactual document for each target query. In the extreme, we might want to flood the corpus with counterfactual documents, each providing a distinct and convincing argument promulgating the desired disinformation. Promoting diversity in our counterfactual documents may also reduce any detectability associated with redundancy. Even if one counterfactual document is labeled as disinformation by a filter, the other documents might be sufficiently different from it to escape detection.

To generate a diverse range of counterfactual documents, we prompt the LLM to generate a document that follows the same format and structure as a relevant, credible document, but presenting an opposing view. We refer to these counterfactual documents as "`evil-twin`" documents since they provide a "radically inverted" counterpart[1] to an authentic "`good-twin`". By generating counter-arguments and advocating opposing views based on a given target query and an authentic document, this method avoids repetition and potentially enhances the subtlety of the false content. Compared to humans, employing LLMs for this purpose is cost-effective and scalable. Ideally, the resulting counterfactual documents would not only adhere to the format and structure of their factual counterparts, but also exhibit a level of diversity, fluency, and consistency that renders them challenging to detect as spam by conventional spam filtering methods, as well as identifying them as disinformation documents by an LLM.

In order to validate our approach, we report experiments using TREC 2020 and TREC 2021 Health Misinformation Tracks. These tracks comprise queries and their corresponding gold standard documents, labeled by human annotators in terms of relevance, correctness, and credibility. Our results demonstrate that on average `evil-twin` documents rank higher than their corresponding `good-twin` documents. In addition, we show that these counterfactual documents are more diverse and more difficult to detect by conventional spam filtering methods when compared to counterfactual documents generated by the target queries alone.

All our data and code is publicly available at `https://anonymous.4open.science/r/adv-ir-llm-2C83/`.

2 Related Work

The investigation of the robustness of deep neural network models has been extensively explored across various domains, including natural language processing (Song et al., 2020; Zou et al., 2019; Ebrahimi et al., 2017; Maimon and Rokach, 2022), computer vision (Akhtar et al., 2021; Siddhant et al., 2019), and recommender systems (Fan et al., 2022; Zhang et al., 2020; Chen et al., 2022).

With the advancement of neural ranking models in information retrieval systems and their remarkable performance, there has been a significant shift from traditional term-frequency-based methods to neural ranking models. Recently, there has been a growing attention towards assessing the robustness of these models against black-hat SEO and web spamming attacks (Patil Swati et al., 2013; Gyöngyi et al., 2005). These adversarial attacks aim to manipulate a target document to deceive the model into ranking the perturbed document higher and thereby increase its exposure to user search queries (Castillo et al., 2011). Adversarial attacks can be classified into traditional term spamming attacks, word-level attacks (Liu et al., 2023b; Raval and Verma, 2020; Wu et al., 2023; Wang et al., 2022; Liu et al., 2024), sentence-level attacks (Chen et al., 2023; Liu et al., 2024), and trigger generation attacks (Liu et al., 2022, 2023b). All of these attacking strategies are applied on a

[1] `https://en.wikipedia.org/wiki/Evil_twin`

<|System|> You are an expert assistant in the field of information retrieval.
<|User|> Given the query, its description, and the stance provided below, please generate a document with an opposite stance.

Query: {query}
Description: {description}
Stance: {stance}
Document:

Figure 1: Prompt for generating `liar` documents.

set of already existed malicious target documents and are not used to generate disinformation or misleading content. Our approach diverges by crafting counterfactual documents to poison the corpus with newly introduced disinformation. Notably, these rank attack techniques can also be applied to these counterfactual documents to enhance their ranking position further.

In a similar study, Zhong et al. (2023) aim to poison the corpus by proposing an optimization-based method that generates adversarial passages that maximize similarity to a set of queries, employing a gradient-based technique similar to HotFlip (Ebrahimi et al., 2017) for token replacement. Although the generated passages rank in a high ranking position due to containing terms similar to user search queries, they do not contain any misleading information or malicious content. In addition, the generated passages might be easily detected by document filtering methods due to lack of consistency and coherency between document terms. Since we generate counterfactual documents based on factual documents using LLMs, the generated documents contain disinformation with respect to a target query, and they may be less perceptible to humans and machines due to the capabilities of LLMs in generating human-like content.

3 Experimental Setup

In this section, we provide overview on the datasets used for conducting the experiments as well as the the process of generating counterfactual documents using a large language model. In addition, we provide details about the ranking models and the large language model used for pairwise ranking.

3.1 Dataset

We conduct our experiments with TREC 2020 and TREC 2021 Health Misinformation Track (Clarke et al., 2020, 2021) test collections, which are designed to evaluate the performance of information retrieval systems on health queries, where the goal is to provide correct and credible information while avoiding disinformation. These test collections are particularly suitable for our experiments since they contain explicitly "helpful" documents that are judged by human assessors as correct, credible, and useful with respect to various health-related topics. These helpful documents make ideal `good-twin` documents.

The TREC 2020 test collection comprises 46 coronavirus pandemic (COVID-19) related topics each asking questions about COVID-19 treatments ("Can vitamin D cure COVID-19?"); the corpus for this collection consists of news documents from the Common Crawl dataset[2] that covered the first four months of 2020. The TREC 2021 test collection comprises 35 topics each proposing a treatment for a general medical condition ("Is the Hoxsey treatment a good cure for cancer?"); the corpus for this collection consists of the "`noclean`" version of the C4 dataset[3]. In both TREC test collections, a topic includes both a keyword query, which might typed into a traditional search engine, and a longer description field containing a natural-language question. Each topic also includes a binary stance indicating whether the proposed treatment helps the medical condition or not.

Each topic has an associated set of assessed documents, labeled according to their correctness, credibility, and usefulness in answering the associated question. In the TREC 2020 dataset, documents are assigned preference codes ranging from -2 to 4, while in TREC 20201, documents receive preference codes ranging from -3 to 12. These preference codes combine individual labels indicating correctness, credibility, and usefulness into a single code for evaluation purposes. Larger codes indicate more helpful documents, with negative codes indicating disinformation ("unhelpful documents").

3.2 Adversarial Document Generation

We employ two distinct strategies to generate adversarial counterfactual documents, both use the GPT-

[2] https://commoncrawl.org/2016/10/news-dataset-available/

[3] https://paperswithcode.com/dataset/c4

4 LLM (`gpt-4-1106-preview`) from OpenAI accessed through its API. The first strategy generates what we term "`liar`" documents by prompting the LLM as shown in Figure 1 to generate documents for topics based on their query, description, and stance in a way that the generated document espouses the opposite stance. For example, if the topic indicates that the treatment is helpful, the LLM is prompted to generate a document arguing against the efficacy of the treatment.

To generate the `evil-twin` documents, we provide not only the query, description, and stance but also an example of a correct, credible, and useful document (Figure 2). These `good-twin` documents all support the stance, are labeled as relevant or highly relevant, and are labeled as having excellent or good credibility, representing the best retrievable documents for the topics. As `good-twin` documents we selected up to 10 of the most helpful documents for each topic, based on their preference codes, as described below. For each topic, we generated the same number of `liar` documents and `evil-twin` documents as `good-twin` documents.

For TREC 2020, we selected as `good-twin` documents only documents with preference code 4, since only these documents are relevant, correct, and credible. Out of 46 topics, only 36 topics had at least one labeled document with a preference code of 4, and topics lacking any potential `good-twin` documents were excluded from our experiments. When there were more than 10 documents with code 4 for a topic, we randomly selected 10 of those documents for our experiments. If there are less than 10 documents, we used all available documents.

For TREC 2021, documents with scores between 9 and 12 are correct, credible, and useful, with differing levels of credibility and usefulness. If there were 10 or more documents with code 12, we randomly selected 10 of those those documents. If there are less than 10 documents with code 12, we randomly selected additional documents from those with code 11, and so on, until we had 10 documents. For topics that had less than 10 documents with a preference score of 9 or above, we used all available documents. Consequently, of the 35 topics, two lacked any documents with a preference score of 9 or above, and these two queries were excluded from our experiments.

<|System|> You are an expert assistant in the field of information retrieval.
<|User|> Given the query, its description, and the arguments and viewpoints presented in its relevant document provided below, please generate a document that presents a counter-argument and advocates for an opposing view. Emphasize key points of disagreement and provide reasons supporting the alternative perspective. The generated document should be coherent, persuasive, and more importantly follow the same writing structure as the original document. Do not say any word about being a counter-arguemnt document.

Query: {query}
Description: {description}
Document: {document}
Counter-Argument Document:

Figure 2: Prompt for generating `evil-twin` documents.

3.3 Re-Ranking Methods

To compare `evil-twin`, `good-twin`, and `liar` documents in an information retrieval context, we assume they were retrieved by a first-stage model and measure how five re-ranking methods would rank them. Two methods are established supervised re-ranking methods: `monoBERT` (Nogueira and Cho, 2019) and `monoT5` (Nogueira et al., 2020). Two are zero-shot ranking methods built on OpenAI embeddings: `text-ada-v2` and `text-3-embedding-small`[4]. Finally, guided by prior work (Sun et al., 2023; Qin et al., 2023), we directly employ GPT-4 (`gpt-4-1106-preview`) as a re-ranker through Pairwise Ranking Prompting (PRP) by prompting it to make a comparative ranking ("Which passage is more relevant?"). For re-ranking purposes, we represent the target query by combining the text of topic's query and description.

4 Results and Findings

In order to evaluate the potential impact of adversarial documents on information retrieval systems from different perspectives, we consider four research questions (RQ) as follows:

[4]For the sake of space in the figures, `text-3-embedding-small` is labeled as `text-3-small`

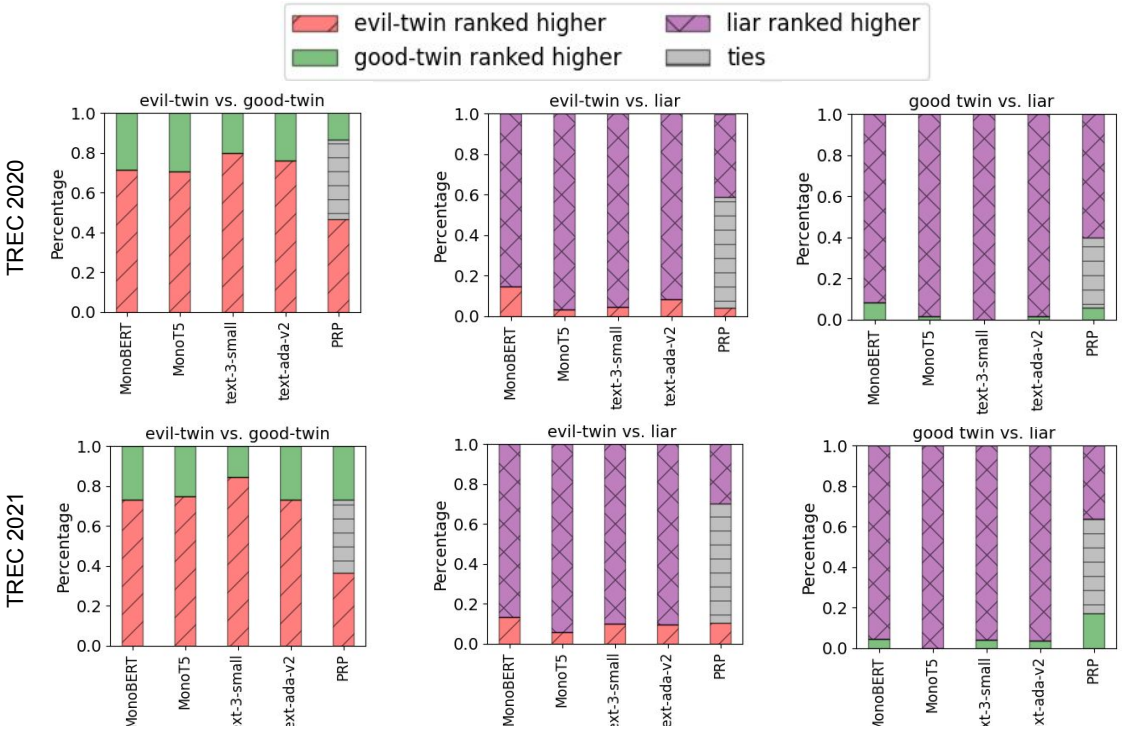

Figure 3: Percentage of pairings where one type of document was ranked higher than the other under five different rankers.

1. How do different document categories, particularly `good-twin` verses `evil-twin`, compare in terms of their relative *ranking* positions within search results?

2. What level of *diversity* do different types of adversarial documents exhibit relative to each other and to their `good-twin` counterparts?

3. How effective are current *spam detection* mechanisms in identifying and filtering the different types of adversarial content, thereby preserving the integrity of search results?

4. To what extent can a LLM effectively *detect disinformation* in `evil-twin` and `liar` documents?

4.1 Ranking (R1)

We investigate how the five rankers, introduced in Section 3.3, rank the three categories of documents detailed in Section 3.2. Our goal is to evaluate the vulnerability of ranking models to adversarial documents. For each topic we create three sets of pairings. One set pairs each `evil-twin` with each `good-twin`, one set pairs each `evil-twin` with each `liar`, and one set pairs each `good-twin` with each `liar`. For each of the five rankers we compute the percentage of pairings where a document of one category is ranked higher than the other.

Figure 3 presents the results of this experiment over TREC 2020 and 2021 test collections. Given the large document sizes in the Common Crawl news collection and the C4 collection, for the first four ranking methods we divide documents into chunks of 512 tokens with a stride of 256 tokens. We determine the relevance score of the topic-document pair used in the re-ranking process by considering the maximum similarity score between the topic vector representation and each chunk. The fifth method (PRP) can produce ties between documents (Qin et al., 2023) that are shown in gray.

We make several observations that are consists across both TREC 2020 and 2021 test collections. First, `evil-twin` documents generally rank higher than their corresponding `good-twin` documents across all ranking methods. We attribute this outcome to the design of `evil-twin` documents, which are generated from both the target query and an `good-twin`, enhancing their relevance to the target query relative to the `good-twin`. Second, all re-ranking methods show a strong preference for `liar` documents. Since the `liar` documents are generated from the target query alone, they should

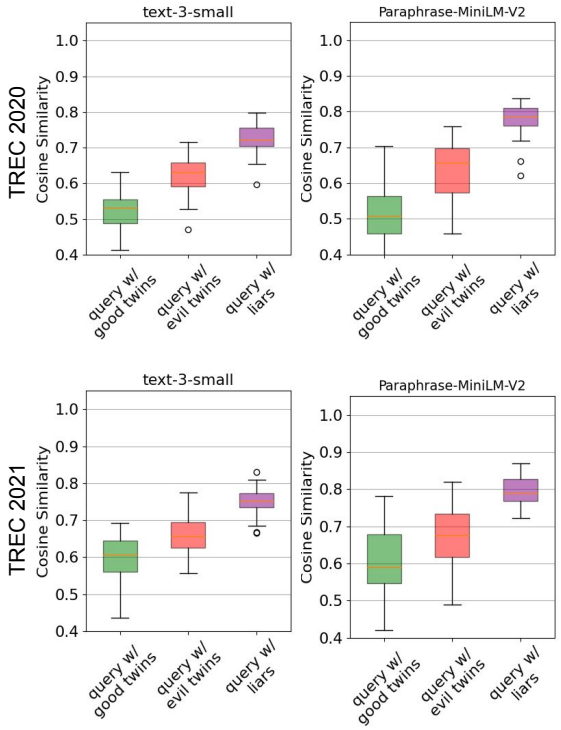

Figure 4: Distribution of pairwise cosine similarities between queries and different document categories using two different language models.

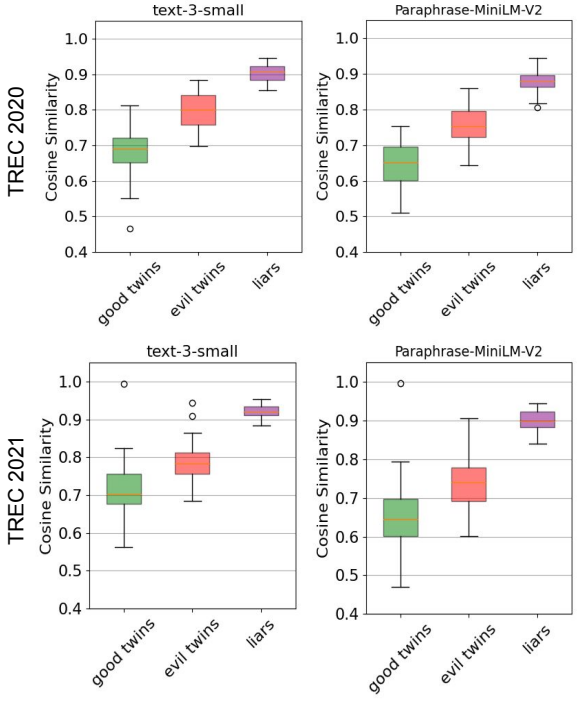

Figure 5: Distribution of pairwise cosine similarities within document categories using two different language models.

be highly relevant to it. Finally, the aggregation of these findings – particularly the superior rankings of `evil-twin` over `good-twin` documents — underscore a critical need for methods to ensure the integrity and accuracy of information retrieval systems.

4.2 Diversity (R2)

We consider the diversity exhibited by different document category through the lens of: 1) the similarity between the query and various document category as depicted in Figure 4, and 2) the inter-document similarity within each category as shown in Figure 5. We measure similarity with two different language models: (1) the `text-3-embedding-small` model, recognized as one of the latest and most efficient models from OpenAI for semantic search and document clustering; and (2) the `paraphrase-MiniLM-L6-v2` model, recognized for its adeptness in converting sentences and paragraphs into dense vector space, with state-of-the-art semantic search capabilities. Using these models, we compute the similarity between queries and the three document categories, as well as the inter-document similarity within each category. The process involves concatenating the query and description of each topic, dividing documents into chunks of 512 tokens with a stride of 256, and computing the maximum similarity between the topic's vector and the chunk's vector. For inter-document similarity, we partition documents into chunks and compute the mean similarity of their chunks, facilitating the comparison of similarity scores across document pairs.

The plots in Figure 4 show query-document similarity across both test collection, with `liar` documents demonstrating the highest similarity to queries, followed by `evil-twin` and `good-twins` documents, which exhibit the lowest similarity. These differences align with the ranking trends observed in Figure 3.

Figure 5 considers the internal diversity within document categories, assessed through pairwise document similarities across both test collection. A higher level of pairwise similarity indicates lower diversity, as documents within the category are more alike. The `liar` documents exhibit the least diversity. These documents are all generated by the same model with the same prompt. Conversely, the `good-twin` documents show the most diversity. They are distinct documents taken from a range of different web sites. The `evil-twin` documents occupy the middle ground. Ideally they

Table 1: Spamicity detection rates (%) of good-twin, evil-twin, and liar documents.

Collection	Category	Spamicity Threshold					
		0.35	**0.30**	**0.25**	**0.20**	**0.15**	**0.10**
TREC 2020	good-twin	0.00%	0.00%	0.33%	1.01%	5.40%	29.05%
	evil-twin	0.00%	0.00%	2.70%	14.86%	44.25%	81.08%
	liar	0.00%	3.71%	17.90%	56.75%	94.93%	99.32%
TREC 2021	good-twin	0.00%	0.69%	3.11%	14.87%	41.52%	71.62%
	evil-twin	0.00%	1.03%	12.45%	41.86%	68.85%	89.61%
	liar	0.69%	16.26%	55.36%	91.69%	99.65%	99.99%

would exhibit the same level of diversity as the good-twin documents, making them more like authentic documents. We also observe outliers among the good-twin documents in Figure 5, which exhibit near-perfect similarity. Upon manual examination, we discovered these outliers are due to near-duplicate documents in the C4 corpus, differing only in their titles.

4.3 Spam Detection (R3)

Following the approach of previous studies (Liu et al., 2022; Wu et al., 2023), we employ a term-frequency-based spam detection filter (Zhou and Pei, 2009) to identify spam documents. Our investigation focuses on discerning whether adversarially generated documents can evade detection as spam. We apply this detection method to good-twin documents, evil-twin documents, and liar documents.

Table 1 presents the results. The table header indicates the detection threshold. Beginning with a threshold of 0.35, where the spam detection ratio of good-twin documents is still 0%, we decrement the threshold by 0.05 until reaching 0.1. Lower thresholds result in stricter detection criteria, leading to increased detection rates across all document categories, albeit with higher false positives. Notably, the spam detection ratios of evil-twin documents closely resemble those of good-twin documents at thresholds of 0.35 and 0.30, making them challenging to identify as spam. Ideally the evil-twin documents would exhibit detection ratios lower than the good-twin documents. Nonetheless, relative to the liar documents, the evil-twin documents are better able to conceal themselves within the corpus. As we decrease the threshold, the liar documents exhibit lower resistance to detection. The evil-twin documents not only exhibit greater diversity but also present a greater challenge to spam filters compared to liar documents.

4.4 Disinformation Detection (R4)

In this section, we consider the efficacy of large language models (LLMs) in identifying indicators of disinformation or malicious content within documents categorized as liar and evil-twin. Our primary objective is to assess whether LLMs can efficiently discern documents that contain malicious content from a corpus on a large scale. To achieve this, we employ a specific prompt strategy, outlined in Figure 7, and prompt the LLM (GPT-4) by providing not only liar and evil-twin documents but also good-twin documents. This enables us to compare the disinformation detection rates of documents containing disinformation with those of correct, credible, and useful documents.

Figure 6 illustrates the disinformation detection rates derived from LLM responses across three document categories across both the TREC 2020 and TREC 2021 test collections. It is evident that evil-twin documents exhibit the lowest detection ratio for disinformation, in comparison to liar and good-twin documents, across both test collections, with detection ratios approximately at 7.8% and 1.4%, respectively. This observation underscores the challenge posed by evil-twin documents, as they manage to evade detection or filtration by LLMs. In TREC 2020, good-twin documents rank second with a detection ratio of almost 15%, while liar documents have the highest detection ratio, reaching 26.7%. This indicates that liar documents are more readily identified as disinformation compared to the other two categories.

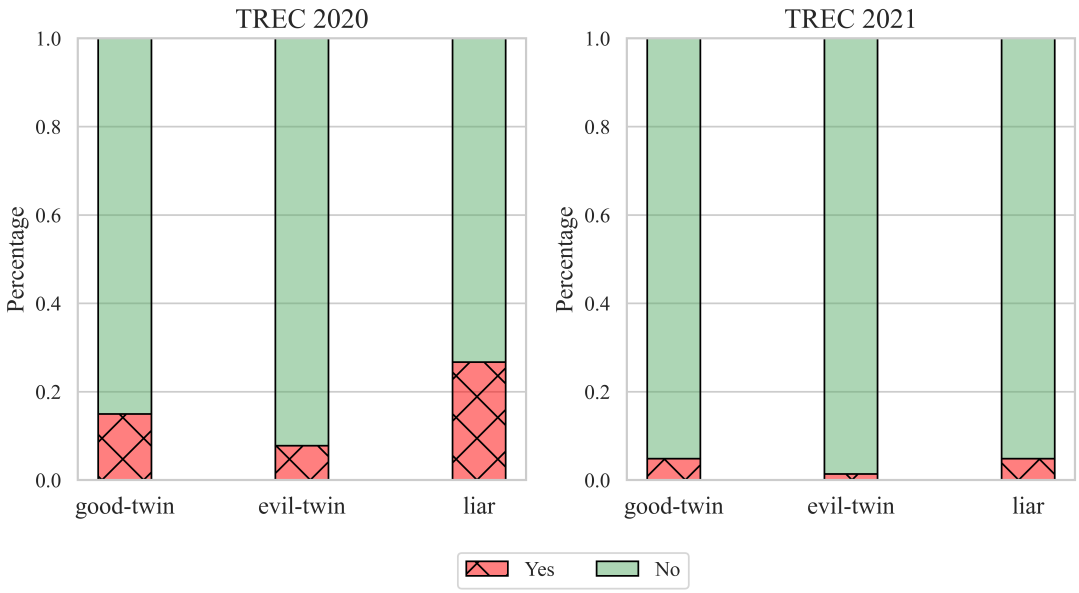

Figure 6: Disinformation detection rates of GPT-4 across various document categories in TREC 2020 and TREC 2021 test collections as determined by LLM responses.

<|System|> You are an expert assistant in the field of information retrieval.
<|User|> Given the document below, can you confirm whether there are any signs of disinformation or malicious content in the document intended to mislead a reader? Respond with either "Yes" or "No" only.

Document: {document}
Response:

Figure 7: Prompt for assessing the presence of disinformation in a document.

In the TREC 2021 evaluation, it is observed that the detection rates for all three categories were notably low, each falling below 5%. Additionally, the detection rates across the categories were relatively similar. Among the document types analyzed, `liar` documents ranked second with a detection rate of nearly 4.8%, while `good-twin` documents exhibited the highest detection rate, approximately 4.9%.

The results of this experiment demonstrate that the LLM performs poorly in detecting disinformation as the detected ratio of disinformation among factual and credible documents is more than counterfactual documents that contain malicious content. In addition, We speculate that the increased disinformation detection ratio among `liar` and `evil-twin` documents in TREC 2020 compared to TREC 2021 may be attributed to a greater volume of LLM training data pertaining to COVID-19. However, an LLM cannot be used for disinformation detection or filtering for documents of the corpus due to considerable false positive rate among `good-twin` documents compared with the other categories.

5 Conclusion and Future Work

We present and evaluate a method for generating synthetic disinformation with large language models. Using the combination of queries and a genuine documents that are factually correct, we are able to generate sets of counterfactual documents that contain disinformation about the query and often rank higher than genuine documents. When compared to disinformation generated by LLMs prompted with queries alone, these counterfactual documents exhibit greater diversity and less spamminess.

Of course, our ultimate goal is not to produce disinformation, but to facilitate research into detecting synthetic disinformation generated by large language models. To this end we provide code and generated documents at OMITTED. Our method is simple, requiring only a commercial LLM available through an API, a target query for disinformation, and examples of genuine documents. Given its simplicity, our method can be viewed as a baseline. More sophisticated methods may produce even "better" disinformation.

Limitations

We use only a single LLM to generate `evil-twin` documents. Our goal is to demonstrate the potential of the approach, and not to determine which of the current models provides the best performance on the task of generating `evil-twin` documents. We expect that the efficacity of the approach to improve as models continue to improve on a range of tasks.

For this study we choose the TREC Health Misinformation test collections because many topics had multiple, curated "helpful" documents that could be used as `good-twin` documents. While the total number of topics may be relatively small, they are genuine targets of misinformation in the associated corpora. Nonetheless, not all topics are genuine targets of adversarial disinformation, particularly the non-COVID topics from TREC 2021. For example, some topics relate to traditional cures ("ice on a burn") that are not supported by science. In addition, our experiments do not consider political topics, which are more frequently the subject of disinformation.

We propose a method for generating disinformation, but we do not consider the challenges of detecting it. Further research is needed to develop robust techniques for identifying and filtering out such adversarial content in real-world information retrieval systems.

References

Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. 2021. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196.

Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can generative llms create query variants for test collections? an exploratory study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1869–1873.

Negar Arabzadeh, Amin Bigdeli, and Charles LA Clarke. 2024. Adapting standard retrieval benchmarks to evaluate generated answers. *arXiv preprint arXiv:2401.04842*.

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. 2023a. A test collection of synthetic documents for training rankers: Chatgpt vs. human experts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5311–5315.

Arian Askari, Mohammad Aliannejadi, Chuan Meng, Evangelos Kanoulas, and Suzan Verberne. 2023b. Expand, highlight, generate: RL-driven document generation for passage reranking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10087–10099, Singapore. Association for Computational Linguistics.

Carlos Castillo, Brian D Davison, et al. 2011. Adversarial web search. *Foundations and trends® in information retrieval*, 4(5):377–486.

Jingfan Chen, Wenqi Fan, Guanghui Zhu, Xiangyu Zhao, Chunfeng Yuan, Qing Li, and Yihua Huang. 2022. Knowledge-enhanced black-box attacks for recommendations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 108–117.

Xuanang Chen, Ben He, Zheng Ye, Le Sun, and Yingfei Sun. 2023. Towards imperceptible document manipulations against neural ranking models. *arXiv preprint arXiv:2305.01860*.

Charles L. A. Clarke, Maria Maistro, and Mark D. Smucker. 2021. Overview of the TREC 2021 health misinformation track. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021*, volume 500-335 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Charles L. A. Clarke, Saira Rizvi, Mark D. Smucker, Maria Maistro, and Guido Zuccon. 2020. Overview of the TREC 2020 health misinformation track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50.

Wenqi Fan, Xiangyu Zhao, Xiao Chen, Jingran Su, Jingtong Gao, Lin Wang, Qidong Liu, Yiqi Wang, Han Xu, Lei Chen, et al. 2022. A comprehensive survey on trustworthy recommender systems. *arXiv preprint arXiv:2209.10117*.

Marcos Fernández-Pichel, David E Losada, and Juan C Pichel. 2022. A multistage retrieval system for health-related misinformation detection. *Engineering Applications of Artificial Intelligence*, 115:105211.

Zoltán Gyöngyi, Hector Garcia-Molina, et al. 2005. Web spam taxonomy. In *AIRWeb*, volume 5, pages 39–47. Citeseer.

Niddal H Imam and Vassilios G Vassilakis. 2019. A survey of attacks against twitter spam detectors in an adversarial environment. *Robotics*, 8(3):50.

Raymond YK Lau, SY Liao, Ron Chi-Wai Kwok, Kaiquan Xu, Yunqing Xia, and Yuefeng Li. 2012. Text mining and probabilistic language modeling for online review spam detection. *ACM Transactions on Management Information Systems (TMIS)*, 2(4):1–30.

Jiawei Liu, Yangyang Kang, Di Tang, Kaisong Song, Changlong Sun, Xiaofeng Wang, Wei Lu, and Xiaozhong Liu. 2022. Order-disorder: Imitation adversarial attacks for black-box neural ranking models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2025–2039.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023a. Black-box adversarial attacks against dense retrieval models: A multi-view contrastive learning method. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1647–1656.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023b. Topic-oriented adversarial attacks against black-box neural ranking models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1700–1709.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Multi-granular adversarial attacks against black-box neural ranking models. *arXiv preprint arXiv:2404.01574*.

Gallil Maimon and Lior Rokach. 2022. A universal adversarial policy for text classifiers. *Neural Networks*, 153:282–291.

Janet Morahan-Martin and Colleen D Anderson. 2000. Information and misinformation online: Recommendations for facilitating accurate mental health information retrieval and evaluation. *CyberPsychology & Behavior*, 3(5):731–746.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

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

P Patil Swati, BV Pawar, and S Patil Ajay. 2013. Search engine optimization: A study. *Research Journal of Computer and Information Technology Sciences*, 1(1):10–13.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Nisarg Raval and Manisha Verma. 2020. One word at a time: adversarial attacks on retrieval models. *arXiv preprint arXiv:2008.02197*.

Minoru Sasaki and Hiroyuki Shinnou. 2005. Spam detection using text clustering. In *2005 International Conference on Cyberworlds (CW'05)*, pages 4–pp. IEEE.

Bhambri Siddhant, Muku Sumanyu, Tulasi Avinash, and Buduru Arun Balaji. 2019. A survey of black-box adversarial attacks on computer vision models. *arXiv preprint arXiv:1912.01667*.

Congzheng Song, Alexander M Rush, and Vitaly Shmatikov. 2020. Adversarial semantic collisions. *arXiv preprint arXiv:2011.04743*.

Nikita Spirin and Jiawei Han. 2012. Survey on web spam detection: principles and algorithms. *ACM SIGKDD explorations newsletter*, 13(2):50–64.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621*.

James Thorne and Andreas Vlachos. 2019. Adversarial attacks against fact extraction and verification. *arXiv preprint arXiv:1903.05543*.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Yumeng Wang, Lijun Lyu, and Avishek Anand. 2022. Bert rankers are brittle: a study using adversarial document perturbations. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 115–120.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten De Rijke, Yixing Fan, and Xueqi Cheng. 2023. Prada: practical black-box adversarial attacks against neural ranking models. *ACM Transactions on Information Systems*, 41(4):1–27.

Shijie Zhang, Hongzhi Yin, Tong Chen, Quoc Viet Nguyen Hung, Zi Huang, and Lizhen Cui. 2020. Gcn-based user representation learning for unifying robust recommendation and fraudster detection. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 689–698.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning retrieval corpora by injecting adversarial passages. *arXiv preprint arXiv:2310.19156*.

Bin Zhou and Jian Pei. 2009. Osd: An online web spam detection system. In *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, volume 9.

Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. 2019. A reinforced generation of adversarial examples for neural machine translation. *arXiv preprint arXiv:1911.03677*.