Few-Shot Adversarial Attacks against Neural Ranking Models

Anonymous Author(s)

ABSTRACT

Neural ranking models have become the backbone of modern information retrieval systems, yet they remain vulnerable to adversarial manipulation. This paper introduces Few-Shot Adversarial Prompting (FSAP), a novel framework that leverages large language models (LLMs) to generate harmful, high-ranking adversarial documents without access to model gradients or internal states. Unlike prior attacks that modify existing documents or rely on handcrafted templates, FSAP exploits in-context learning to synthesize realistic adversarial documents conditioned on a small support set of previously seen harmful examples. We propose two variants: $\text{FSAP}_{\text{IntraQ}}$, which uses examples from the same query, and $\text{FSAP}_{\text{InterQ}}$, which transfers adversarial patterns across unrelated topics. Through comprehensive evaluation on the TREC 2020 and TREC 2021 Health Misinformation Tracks and across four neural rankers, we show that FSAP achieves superior attack effectiveness, strong stance fidelity, and high undetectability. Our findings demonstrate that FSAP generalizes across different LLMs, posing a transferable and scalable threat model for neural retrieval systems.

1 INTRODUCTION

Ensuring the integrity and accuracy of the results presented to searchers by Information Retrieval (IR) 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 [3, 7, 19, 20, 38]. These attacks, commonly known as black-hat search engine optimization or web spamming, are designed to find human-imperceptible perturbations to maliciously manipulate existing 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 [23].

In the past, adversarial attacks may have taken 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 [5, 16, 30]. 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. However, documents manipulated by recent state-of-the-art adversarial attack models are now capable of bypassing these spam detection mechanisms and are often imperceptible to both human evaluators and automated systems, thereby undermining the robustness and integrity of modern retrieval systems [3, 7].

With the rise of Large Language Models (LLMs), several recent works have raised concerns about LLMs being adopted to generate misinformation at scale [6, 15, 36, 42]. For example, Chen and Shu [6] demonstrated how both ChatGPT and open-source models like Llama2 can be used to produce misinformation in arbitrary settings (generating content from scratch) or controlled settings (rewriting misleading documents or inverting facts in factual texts). Hu et al. [15] employed GPT-4o-mini and LLaMA-3.1 to generate misinformation news articles and investigate their impact on recommender systems. Their findings showed that the presence of LLM-generated misinformation led to fake news being ranked above real news and as a result distorts recommendation outputs. The findings of these studies underscore the growing threat posed by LLMs, which can produce content that is both persuasive and difficult to be detected with existing safeguards. While human-generated disinformation has traditionally been constrained by the cost and effort of manual creation [18, 33], LLMs enable scalable, low-cost generation of deceptive content which significantly amplifies the risk.

Building on prior research that generate counterfactual text by simply prompting the LLMs, this paper investigates how to instruct LLMs, within a few-shot framework, to autonomously generate adversarial content tailored to deceive NRMs. Unlike previous studies that have primarily focused on altering existing documents for attacking NRMs, we introduce a new threat landscape for NRMs by proposing a Few-Shot Adversarial Prompting (FSAP) framework for harmful document generation. Instead of relying on document-specific editing or query-only generation, in our threat model an adversary instructs LLMs with few previously annotated human generated harmful documents from different queries to generate new counterfactual documents that are structured convincingly and seamlessly to convey adversarial content in response to a new target search query. This design supports generalization across queries and does not require task-specific supervision for the target query. By leveraging LLMs' generalization abilities, we demonstrate that such few-shot adversarial prompts can reliably induce the model to generate harmful counterfactual documents that are capable of ranking higher than authentic and accurate sources (see Figure 1).

Our proposed work is grounded in the observation that LLMs possess strong in-context generalization abilities and can internalize stylistic, structural, and semantic patterns from limited examples without explicit fine-tuning [4, 22]. We hypothesize that an adversary can exploit these properties to generate high-quality adversarial content by prompting the model with a small number of previously observed harmful query-document pairs. This strategy assumes *only black-box access* to a language model and a modest collection of known examples. Based on these assumptions, our FSAP framework constructs structured prompts by concatenating several harmful examples, which serve as implicit behavioral cues for the model. When conditioned on a new search query, the model draws on these few-shot examples to generate a document that is grammatically fluent, topically relevant, and stylistically natural yet subtly embeds misleading or false information. The success of the attack is measured by the model's ability to produce documents that are ranked higher than accurate, credible content by a target neural ranking system. FSAP supports two instantiations: one focused on single-topic attacks using examples from the same query to enhance coherence, referred to as $\text{FSAP}_{\text{IntraQ}}$, and another that enables broader generalization by using examples drawn from unrelated topics, denoted by $\text{FSAP}_{\text{InterQ}}$.

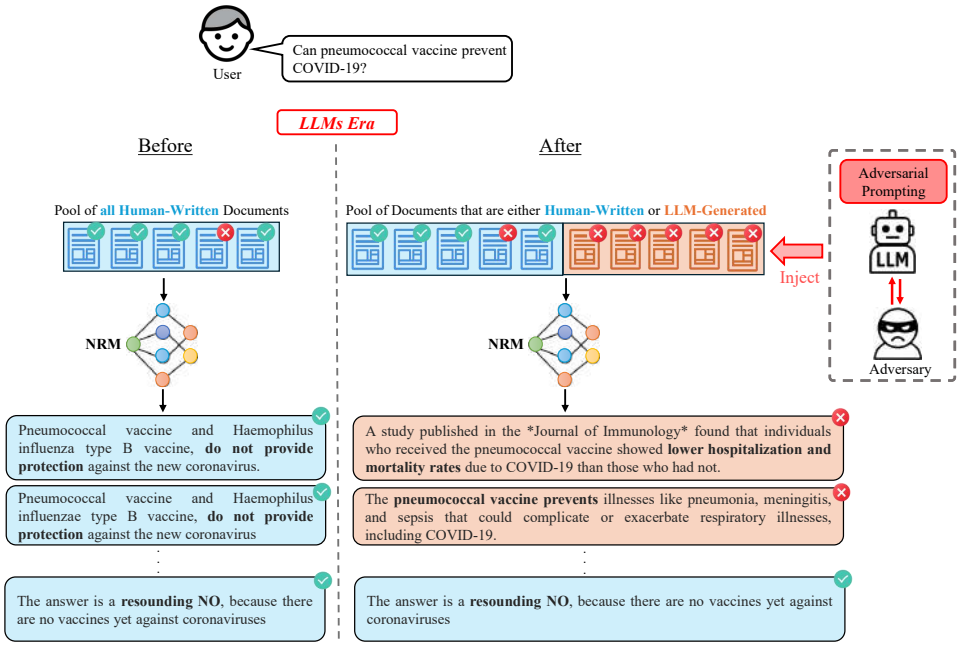

Figure 1: Illustration of LLM-based pool poisoning. As LLMs are prompted with adversarial intent, they generate adversarial harmful documents that poison the document pool. These LLM-generated texts, when added to the pool of documents, can deceive NRMs and appear above credible content, increasing user exposure to adversarial content.

Both instantiations of FSAP do not require any fine-tuning or internal access to the LLM model, making them a realistic and scalable mechanism for adversarial document generation in retrieval settings.

To evaluate our proposed framework, we conduct experiments using the TREC 2020 and TREC 2021 Health Misinformation Tracks, which are the only collections available that provide target queries alongside gold standard helpful and harmful documents labeled for relevance, correctness, and credibility by human annotators. Our experimental results reveal that LLM-generated adversarial documents created by our proposed FSAP framework across different LLM models can deceive state-of-the-art NRMs to rank the adversarially generated content above credible factual documents. Specifically, the $\text{FSAP}_{\text{InterQ}}$ variant of our framework is particularly effective, with the generated harmful documents achieving a Mean Helpful Defeat Rate of 90% on average across various NRMs, compared to the helpful counterparts. In addition, the high undetectability of these adversarial documents compared to baselines highlights the effectiveness of our method in enabling realistic and effective attacks.

More concretely, the contributions of our work in this paper include:

(1) We propose Few-Shot Adversarial Prompting (FSAP), a framework that leverages LLMs to generate adversarial documents. FSAP leverages a small number of harmful examples to generate new adversarial content capable of ranking above credible factual documents. It supports two instantiations: $\textbf{FSAP}_{\textbf{IntraQ}}$, which uses multiple harmful documents from the same query, and $\textbf{FSAP}_{\textbf{InterQ}}$, which transfers adversarial patterns from unrelated queries.

(2) Through extensive experiments on the TREC 2020 and TREC 2021 Health Misinformation Tracks, we show that adversarial documents generated via FSAP can rank above credible, helpful ones when ranked by state-of-the-art neural rankers.

(3) We conduct a detailed analysis of the generated harmful content, evaluating stance alignment, stylistic diversity, and ranking success across multiple synthetic variants. Our findings highlight not only the effectiveness but also the evasiveness of FSAP-generated documents pointing to the need for further NRMs robustness in face of adversarial attacks using LLMs.

2 RELATED WORK

Adversarial Manipulation Attacks in Neural Ranking Models. With the advancement of neural ranking models, 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 [14, 27]. These adversarial attacks aim to manipulate an existing target document to deceive the model into ranking the perturbed document higher and thereby increase its exposure to the users [5]. Adversarial attacks can be classified into traditional term spamming attacks, word-level attacks [29, 37, 38], sentence-level attacks [3, 7], and trigger generation attacks [19, 37].

There are also various studies that apply these attack strategies in different contexts. For instance, Liu et al. [20] developed a framework that uses reinforcement learning to manipulate documents, improving the ranking position of the target document for similar queries by

employing existing strategies [19, 38]. Another study by Liu et al. [21] proposes a framework that integrates various attack methods through reinforcement learning, using GPT-4's fluency as a reward function to manipulate documents. All of these attacking strategies are applied on a set of already existing malicious target documents and are not used to generate adversarial content. Our approach diverges by crafting counterfactual documents to poison the pool of documents with newly introduced adversarial documents.

LLM-Generated Adversarial Text Generation. There have also been an increasing number of papers that investigate the use of LLMs to generate persuasive misleading content using prompting strategies that vary in terms of specificity and content grounding [12, 15, 26, 36]. The rewriting-based approach [12, 15, 36] show that misleading narratives can be rewritten to appear more credible by mimicking the linguistic style of credible sources. Paraphrasing-based strategies [12, 15, 36] aim to rephrase harmful content to increase lexical diversity or evade detection mechanisms, while preserving its deceptive intent. The fact inversion approach prompts LLMs to produce counterfactual claims by corrupting factual assertions within authentic documents [26, 36, 41]. Several studies explore the zero-shot generation of adversarial content by prompting LLMs to hallucinate coherent but false documents based only on a narrative or false stance without reference content [15, 26, 36].

Despite this progress, the impact of such LLM-generated adversarial documents on NRMs remains largely unexplored. Existing studies focus primarily on LLM-generated counterfactual and misleading content, without ever examining its downstream effect on document ranking. This paper proposes a method for adversarial document generation that introduces new documents into the candidate pool and investigates not only the impact of this method but also other baselines on the ranking process.

3 METHODOLOGY

We present FSAP, a novel, input-level adversarial framework that exploits the in-context learning abilities of Large Language Models (LLMs) to generate syntactically coherent, semantically plausible, yet factually misleading documents. These adversarial documents are constructed to deceive NRMs into ranking them above credible, factually correct documents. Unlike prior adversarial retrieval attacks that rely on token-level perturbations [37, 38], sentence-level manipulation [7], or supervised fine-tuning [19, 37], FSAP is model-agnostic, requires only black-box access to the LLM, and operates entirely via *few-shot prompting*, aligning with the rapidly growing class of black-box LLM exploitation techniques [31, 40].

FSAP draws theoretical grounding from meta-learning and in-context learning literature [4, 13, 22], where LLMs are shown to behave as conditional function approximators capable of performing complex reasoning and pattern reproduction from a small number of exemplars. These capabilities allow FSAP to create context-conditioned document-level attacks that are scalable, few-shot transferable, and difficult to detect via standard content moderation pipelines.

3.1 Problem Setup and Adversarial Objective

Let $\mathcal{Q} = \{q_1, q_2, ..., q_N\}$ denote a set of natural language search queries and $\mathcal{D}$ denote the corpus of documents. For any query $q \in \mathcal{Q}$, let $\mathcal{D}_q^+ = \{d_q^{+(1)}, ..., d_q^{+(n)}\}$ represent a set of helpful (factual and credible) documents, and $\mathcal{D}_q^- = \{d_q^{-(1)}, ..., d_q^{-(m)}\}$ denote a set of known harmful (misleading or false) documents.

We assume access to: (1) A *black-box generative language model* $\mathcal{M}_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by θ, which maps an input sequence $x \in \mathcal{X}$ to a textual output $y \in \mathcal{Y}$; (2) A *target neural ranking model* $\mathcal{R} : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$ that assigns a relevance score to each query-document pair, and (3) A small *support set* $\mathcal{S}^- = \{(q_i, d_{q_i}^-)\}_{i=1}^k$ of query-document pairs, where each $d_{q_i}^-$ is a human-annotated harmful document for query q_i. The attacker's goal is to craft an adversarial document $\tilde{d}_q^-$ for a target query q such that: *(i)* $\tilde{d}_q^-$ is syntactically fluent and stylistically similar to human-authored content; *(ii)* $\tilde{d}_q^-$ is topically coherent with q yet introduces false, biased, or misleading information; and, *(iii)* The relevance score assigned by $\mathcal{R}$ satisfies:

$$\mathcal{R}(q, \tilde{d}_q^-) > \max_{d \in \mathcal{D}_q^+} \mathcal{R}(q, d)$$

This objective is formalized using an expected indicator loss:

$$\mathcal{L}_{adv}(q) = \mathbb{E}_{\mathcal{G}} \left[1 \left\{ \mathcal{R}(q, \tilde{d}_q^-) > \max_{d \in \mathcal{D}_q^+} \mathcal{R}(q, d) \right\} \right]$$

where $\mathcal{G}$ denotes the stochastic generation process governed by the LLM $\mathcal{M}_\theta$ under adversarial prompting.

3.2 Few-Shot Prompt Construction and LLM Conditioning

To instantiate the adversarial generation process, we define a prompt function $\mathcal{P}_{adv}$ that converts the support set $\mathcal{S}^-$ into a structured input sequence interpretable by $\mathcal{M}_\theta$:

$$\mathcal{P}_{adv} = \bigoplus_{i=1}^{k} \texttt{Format}(q_i, d_{q_i}^-)$$

where $\bigoplus$ denotes sequential concatenation and $\texttt{Format}(\cdot)$ encodes query-document pairs using a natural language template (e.g., "Query: ... \n Document: ..."). The LLM is then conditioned on the target query q and adversarial prompt $\mathcal{P}_{adv}$ to produce the candidate adversarial document:

$$\tilde{d}_q^- \sim \mathcal{M}_\theta(\mathcal{P}_{adv}, q)$$

This prompt-conditioning mechanism can be interpreted through a Bayesian lens, where $\mathcal{P}_{adv}$ acts as a contextual prior over the output distribution $p(y \mid q)$ [39]. As a result, $\tilde{d}_q^-$ inherits stylistic and semantic properties of the harmful examples in $\mathcal{S}^-$.

3.3 Intra-Query Prompting ($\texttt{FSAP}_{\texttt{IntraQ}}$)

In this first instantiation, we assume the attacker has access to multiple harmful examples associated with the *same* target query q^*. Therefore, given the repository of query-specific harmful documents $\mathcal{D}_{q^*}^-$ for q^*, its intra-query support set can be defined as:

$$\mathcal{S}_{\text{intra}}^- = \{(q^*, d_{q^*}^{-(1)}), (q^*, d_{q^*}^{-(2)}), \ldots, (q^*, d_{q^*}^{-(k)})\}$$

The few-shot prompt is then constructed from this homogeneous support set as:

$$\mathcal{P}_{\text{intra}} = \bigoplus_{i=1}^{k} \texttt{Format}(q^*, d_{q^*}^{-(i)})$$

The adversarial generation proceeds as:

$$\tilde{d}^{-}_{q^*} = \mathcal{M}_\theta(\mathcal{P}_{\text{intra}}, q^*)$$

This setting aligns with few-shot learning under homogeneous support where examples share task identity shown to improve generation quality and topic fidelity [4, 22]. It encourages the model to replicate not only topic-relevant lexical structures but also specific rhetorical patterns, such as sensationalism or pseudoscientific framing [34]. This strategy benefits from tight semantic control and high in-topic coherence but is bounded by the availability of labeled adversarial samples tied to the target query. This makes $\texttt{FSAP}_{\texttt{IntraQ}}$ ideal for amplification attacks on known adversarial topics [2, 34].

3.4 Inter-Query Prompting ($\texttt{FSAP}_{\texttt{InterQ}}$)

In this more general instantiation, we assume no prior harmful content exists for the target query q^*. Instead, the attacker constructs the prompt from unrelated queries:

$$\mathcal{S}^{-}_{\text{inter}} = \{(q_1, d^{-}_{q_1}), (q_2, d^{-}_{q_2}), \ldots, (q_k, d^{-}_{q_k})\}$$

where $\mathcal{S}^{-}_{\text{inter}}$ is a cross-topic support set consisting of diverse query-document pairs. The few-shot prompt is then constructed as:

$$\mathcal{P}_{\text{inter}} = \bigoplus_{i=1}^{k} \texttt{Format}(q_i, d^{-}_{q_i})$$

The adversarial document for target query q^* is generated via:

$$\tilde{d}^{-}_{q^*} = \mathcal{M}_\theta(\mathcal{P}_{\text{inter}}, q^*)$$

$\texttt{FSAP}_{\texttt{InterQ}}$ relies on the transferability of adversarial structures and rhetorical patterns [4, 34] across semantically diverse topics, and is suited for low-resource, few-shot adversarial scenarios. This formulation primarily relies on the cross-topic generalization capacity of the LLM, relying on the LLM to project latent adversarial priors (e.g., persuasive tone, manipulative structure) to a semantically disjoint query. This mode builds on recent findings in instruction transfer and meta-instruction prompting [17, 35], where LLMs exhibit emergent generalization across heterogeneous prompts. $\texttt{FSAP}_{\texttt{InterQ}}$ requires no prior attack history for a given topic, making it suitable for few-shot poisoning. Though it may introduce slight topic drift, our findings (in the evaluation section) show that it often generates highly persuasive and deceptively aligned outputs, particularly when harmful exemplars share similar stylistic features.

It is important to note that a key property of FSAP is that it requires no gradient access, fine-tuning, or internal instrumentation of the LLM. This black-box interaction model makes FSAP transferable across model families (e.g., GPT, DeepSeek) and applicable to any NRM $\mathcal{R}$ whose scoring function is sensitive to surface fluency and semantic alignment. Theoretically, FSAP can be interpreted as inducing a document-level adversarial distribution $\mathbb{P}\theta^{adv}$ over the LLM's output space, conditioned on $\mathcal{P}adv$ and query q:

$$\tilde{d}^{-}_{q} \sim \mathbb{P}^{adv}_{\theta}(\cdot \mid q, \mathcal{P}_{adv})$$

This adversarial distribution can be used to characterize the decision boundary vulnerabilities of $\mathcal{R}$, similar to adversarial example theory in vision [1, 32] but instantiated at the level of textual content semantics.

4 EXPERIMENTAL SETUP

Our data, prompts, and code are publicly accessible at our repository[1].

4.1 Datasets

Benchmark Datasets. To evaluate the effectiveness of FSAP, we require test collections that go beyond traditional relevance assessments and explicitly distinguish between helpful and harmful content. In adversarial scenarios, it is critical to measure not only whether a document is relevant to a query but also whether it is accurate, credible, and aligned with trustworthy information. However, most large-scale information retrieval benchmarks such as MS MARCO [24] and various TREC Web and Deep Learning tracks [10, 11] are not suitable for this purpose, as their annotations are limited to topical relevance and do not account for factual correctness or the potential presence of misinformation.

To meet the specific requirements of our evaluation, we conduct experiments using the TREC 2020 and TREC 2021 Health Misinformation Tracks [8, 9]. These collections are uniquely designed to assess IR systems in high-stakes domains such as public health, where the goal is not merely to retrieve relevant information, but to prioritize content that is both correct and credible, while penalizing the ranking of misleading or harmful documents. Each document in these test collections is manually labeled by expert annotators with credibility-focused judgments categorized as "helpful," "harmful," or "neutral" based on its factual alignment, trustworthiness, and utility with respect to a health-related query. This makes them particularly well-suited for evaluating adversarial attacks like those generated by FSAP, where the key concern is whether adversarially generated content can outrank verified, helpful sources.

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 have been 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, while negative codes indicate disinformation ("harmful documents").

[1] https://anonymous.4open.science/r/fsap-attack-010F/
[2] https://commoncrawl.org/2016/10/news-dataset-available/
[3] https://paperswithcode.com/dataset/c4

Target Queries and Documents. To conduct our experiments, we selected queries from the TREC 2020 and TREC 2021 Health Misinformation Tracks that have both helpful and harmful documents annotated by human assessors. For each topic with available assessments, we included up to 10 of the most helpful and up to 10 of the most harmful documents, based on their preference codes as described below.

For TREC 2020, we include documents with preference code 4, since only these documents are relevant, correct, and credible. For harmful documents, we selected those with a preference code of -2, indicating they were judged as relevant, incorrect, and credible. Out of the 46 available topics, only 22 had at least one helpful and one harmful document, and topics lacking either were excluded from our experiments. When more than 10 helpful documents (code 4) were available for a topic, we randomly sampled 10. The same procedure was applied to harmful documents. If more than 10 were available, a random subset of 10 was selected. If fewer than 10 documents were available, we used all available documents in that category. For each topic, the human-annotated helpful and harmful documents collectively form the ranking pool used in the re-ranking process.

For TREC 2021, documents with scores between 9 and 12 are correct, credible, and relevant, with differing levels of credibility and relevance. For harmful documents, those with scores -2 and -3 are credible, incorrect, and relevant with various levels of credibility and relevance. If there were 10 or more documents with code 12, we randomly selected 10 of 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. A similar strategy was used for selecting harmful documents, starting with those scored -3 and, if necessary, adding documents with score -2 to reach up to 10 harmful documents. Consequently, of the 35 topics, only 27 had at least one helpful and one harmful documents. Consistent with TREC 2020, for each topic the helpful and harmful documents associated with the query form its pool of ranking for the re-ranking process.

4.2 Models

Large Language Models. We employed two state-of-the-art LLMs with varying parameter sizes to serve as $\mathcal{M}_\theta$ for adversarial document generation. These models include both open-source and API-based systems, allowing us to assess their performance across baselines and our proposed `FSAP` framework. We utilized OpenAI's `GPT-4o`, accessed via the OpenAI API, as a high-performance proprietary model. For open-source alternatives, we included DeepSeek AI's `DeepSeek-R1-claude3.7`, a 14.8-billion-parameter model with the Claude 3.7 Sonnet system prompt known for its efficiency and strong reasoning capabilities despite its smaller size.

Neural Ranking Models (NRMs). To compare LLM-generated harmful documents with their human-written helpful and harmful counterparts within a pool of documents, we leveraged four different NRMs to rank these documents. Two of these NRMs are well-established supervised re-ranking models: `MonoBERT` [25] and `MonoT5` [25]. The other two NRMs are zero-shot ranking models built based on OpenAI embeddings: `text-embedding-ada-002` and `text-3-embedding-small`.

For re-ranking purposes, we represent the target query by combining the text of topic's query and description. Given the large document sizes in the Common Crawl news collection and the C4 collection, 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.

4.3 Evaluation Metrics

To assess the effectiveness of the LLM-generated adversarial documents, we employ a set of evaluation metrics that capture both ranking performance and adversarial success.

Mean Help-Defeat Rate. This is the primary metric used in our experiments, measuring the ability of adversarial documents to outrank helpful documents. Given a set of n helpful documents $\mathcal{D}_q^+$ for a query q, and a set of m adversarial documents $\{\tilde{d}_1^-, \tilde{d}_2^-, \ldots, \tilde{d}_m^-\}$ generated for the same query using a given attack method, we can compute the fraction of helpful documents outranked by each adversarial document $\tilde{d}_j^-$ as:

$$\text{Help-Defeat Rate}(q, \tilde{d}_j^-) = \frac{1}{n} \sum_{i=1}^{n} 1 \left\{ \mathcal{R}(q, \tilde{d}_j^-) > \mathcal{R}(q, d_i^+) \right\} \quad (1)$$

The Mean Help-Defeat Rate is then computed by averaging this rate across all adversarial documents generated by the same method:

$$\text{MHDR}(q) = \frac{1}{m} \sum_{j=1}^{m} \text{Help-Defeat Rate}(q, \tilde{d}_j^-) \quad (2)$$

Intuitively, MHDR calculates how many helpful documents, on average, are outranked by each adversarial document. A higher MHDR indicates a more effective attack, which means that adversarial document generated by the LLM are consistently ranked above factual ones. This metric provides a fine-grained view of the ranking manipulation potential of different adversarial strategies.

Stance Alignment Accuracy. To assess whether the LLM-generated adversarial document contains the intended adversarial stance, we employed `GPT-4o` using a zero-shot prompting approach to assess the stance conveyed in the generated content. The prompt format is available in our public repository.

Adversarial Detection Pass. One of the most important criteria for evaluating adversarial documents is to investigate if they can evade detection mechanisms and remain unflagged to preserve the attack goal. For adversarial detection pass, we prompt `GPT-4o` to determine if an adversarial content gets flagged as adversial or not. Higher detection pass rate indicates a more successful attack because the adversarial document has a higher chance of being exposed. The prompt used for detectability is also available on our repository.

4.4 Baselines

To evaluate the effectiveness of our proposed `FSAP` framework, we compare it against several strong baselines that represent current approaches to LLM-generated adversarial documents. These baselines vary in their access to prior content and in the prompting strategies used to generate harmful documents. The following methods are used as comparative baselines in our experiments: **(1)** `Rewriter`

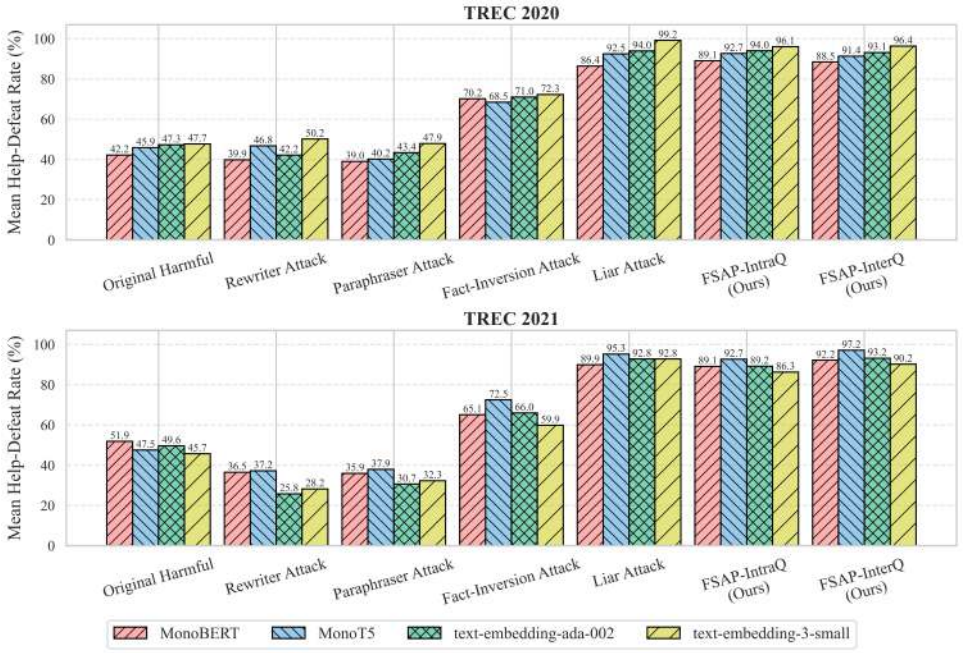

Figure 2: MHDR of original harmful and adversarial documents generated by `GPT-4o` across different attack methods and neural ranking models on the TREC 2020 and TREC 2021 datasets.

`Attack` [12, 15, 36] rewrites a known harmful document associated with the query to enhance deception while preserving its original stance and core claims. **(2)** `Paraphraser Attack` [12, 15, 36] generates stylistic rephrasings of a harmful passage associated with the query to alter its surface form to improve its variability and reduce the likelihood of detection. **(3)** `Fact-Inversion Attack` [26, 36] transforms factual query-related statements in a helpful document into misleading counterfactual claims by reversing or corrupting core factual assertions. **(4)** `Liar Attack` is inspired by works on stance-controlled generation [15, 26, 36]. This baseline provides the LLM only with the target query, its description, and an adversarial stance (e.g., unhelpful or helpful). The model will then generate a document that promotes the specified stance without any supporting examples. This setting tests the LLM's ability to hallucinate adversarial content conditioned solely on query-level metadata.

For each baseline method, we generate one adversarial document for each helpful document based on its associated prompt format. Due to space constraints, full template of baselines prompts is available on our repository. We instantiate both variants of our proposed framework, $\texttt{FSAP}_{\texttt{InterQ}}$ and $\texttt{FSAP}_{\texttt{IntraQ}}$, using few-shot prompting with $k \leq 3$ examples. For $\texttt{FSAP}_{\texttt{InterQ}}$, we fix $k = 3$ and sample disjoint query-document pairs from unrelated topics. For $\texttt{FSAP}_{\texttt{IntraQ}}$, we use up to 3 harmful documents from the same query, if available. Similar to baselines, we generate one document per helpful document using few-shot harmful examples. Prompts are constructed using a standardized natural language format interpretable by the LLM.

5 RESULTS AND FINDINGS

In this section, we evaluate the effectiveness of our proposed `FSAP` framework through investigating four research questions (RQ) as follows:

(1) How does the attack performance of `FSAP` compare to baseline methods in terms of outranking factual helpful documents across various neural ranking models?
(2) How do `FSAP`-generated documents compare to baseline methods in terms of stance alignment with the intended adversarial content and their ability to evade adversarial detection?
(3) How does the size of the few-shot support set affect the effectiveness of `FSAP` in generating high-ranking adversarial documents?
(4) How does the effectiveness of `FSAP`-generated adversarial documents vary across different LLMs used in the prompting process?

5.1 Attack Performance Evaluation (RQ1)

To evaluate the effectiveness of our proposed `FSAP` framework in adversarial ranking, we compare its ability to outrank factual helpful documents against original harmful documents as well as baseline generation methods. For this purpose, we employed `GPT-4o` to generate adversarial documents based on each of the attacking methods prompt style. Figure 2 shows the Mean Help-Defeat Rate (MHDR) results of original harmful documents and adversarial documents generated by each method for the TREC 2020 and TREC 2021 datasets across four neural ranking models. The results demonstrate that both $\texttt{FSAP}_{\texttt{IntraQ}}$ and $\texttt{FSAP}_{\texttt{InterQ}}$ achieve consistently strong performance, often outperforming all baseline attacks and achieving comparable effectiveness with the `Liar Attack`. Notably, $\texttt{FSAP}_{\texttt{InterQ}}$ reaches up to 96.4% MHDR on TREC 2020 using `text-3-embedding-small` and 97.2% on TREC 2021 using `MonoT5`. This demonstrates robust generalization even when support examples are drawn from unrelated query-document pairs.

Among baseline methods, the `Fact-Inversion Attack` shows moderate success, with MHDR values in the 59–72% range. While it

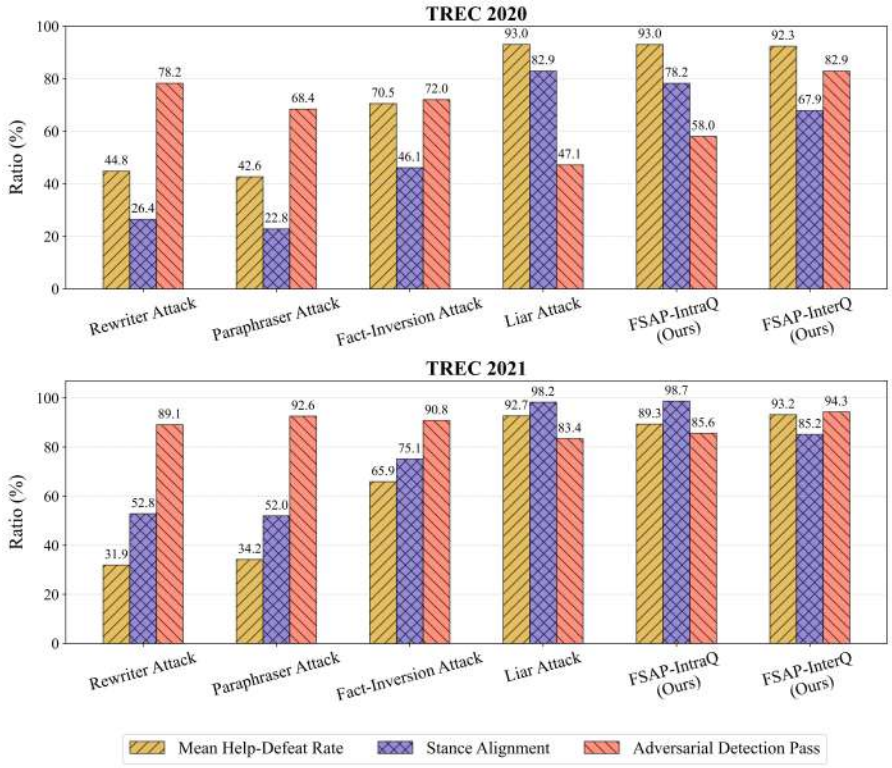

Figure 3: Comparison of stance alignment, detection pass rate, and MHDR across adversarial methods for TREC 2020 and TREC 2021. Our $\texttt{FSAP}_{\texttt{InterQ}}$ method delivers the highest balance of attack performance, undetectability, and stance alignment among all methods.

outperforms shallow surface-level attacks, its performance remains notably lower than both variants of FSAP and the `Liar Attack`, suggesting that inverting factual claims alone may be insufficient to reliably deceive neural rankers. The `Liar Attack` achieves slightly higher MHDR in some cases (e.g., 99.2% in TREC 2020 with `text-embedding-3-small`), its success is due to direct stance conditioning without reference examples. In contrast, FSAP leverages few-shot support examples to produce content that mirrors both the structure and style of human-written harmful documents, which makes it a more realistic and transferable threat model.

The rest of the Baselines, `Rewriter` and `Paraphraser Attacks`, yield substantially lower MHDR values (often below 50%) and in many cases perform even worse than the original human-written harmful documents. This highlights the limitations of surface-level text manipulations and underscores the value of contextualized prompting in crafting persuasive and high-ranking adversarial content.

5.2 Stance Alignment and Detection Evasion (RQ2)

In this section, we assess the ability of different adversarial generation methods to (1) generate documents aligned with an adversarial stance and (2) evade detection by by LLMs. These two aspects are critical for understanding the practical risks posed by adversarial documents in realistic settings.

To evaluate stance alignment and adversarial detection pass, we prompt `GPT-4o` using a standardized zero-shot template designed to assess whether a given document reflects (1) the intended adversarial stance and (2) contains adversarial detectable indicators that would compromise the effectiveness of the attack. This prompt-based detection approach allows for consistent analysis across LLM-generated documents by FSAP and baselines.

Figure 3 presents a comparative analysis of MHDR, stance alignment, and adversarial detection pass rate across FSAP and baselines on TREC 2020 and TREC 2021. For each method, the MHDR value is computed as the average across all four neural ranking models, providing a holistic view of the method's overall attack effectiveness. $\texttt{FSAP}_{\texttt{IntraQ}}$ and $\texttt{FSAP}_{\texttt{InterQ}}$ demonstrate strong stance fidelity by achieving 98.7% and 85.2% respectively on TREC 2021. In terms of adversarial detection pass, $\texttt{FSAP}_{\texttt{InterQ}}$ shows a substantially higher rate (94.3%) compared to the `Liar Attack` (83.4%), while $\texttt{FSAP}_{\texttt{IntraQ}}$ remains similar to the `Liar Attack` (85.6% vs. 83.4%). In addition, although the `Liar Attack` achieves near-perfect stance alignment, it shows detection pass rates below 50% on TREC 2020 and 83.4% on TREC 2021. In contrast, $\texttt{FSAP}_{\texttt{InterQ}}$ achieves the best overall balance, with a 93.2% MHDR and 82.9% and 94.3% detection pass on TREC 2020 and TREC 2021, respectively.

To further analyze trade-off between attack effectiveness and not being detected as adversarial content, Figure 4 plots each method in a two-dimensional space defined by MHDR (x-axis) and detection pass rate (y-axis). The top-right quadrant of each plot is the ideal region that represents methods that are both highly effective and difficult to detect. As shown, $\texttt{FSAP}_{\texttt{InterQ}}$ consistently occupies this optimal region across both datasets, clearly outperforming all baselines in combining high adversarial strength with undetectability. In contrast, the `Liar Attack`, although competitive in terms of MHDR, appears

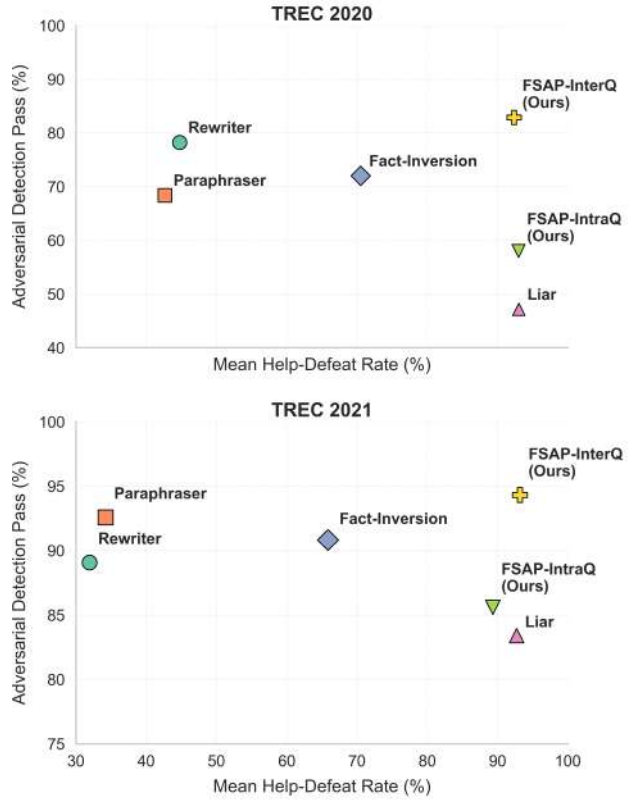

Figure 4: Scatter plot of attack effectiveness (Mean Help-Defeat Rate) vs. adversarial document detection pass rate for various methods on TREC 2020 and TREC 2021.

in the bottom-right quadrant, indicating low adversarial detection pass. Simpler baselines such as `Rewriter` and `Paraphraser Attack` fall in the higher-left region, which shows weak attack capability with modest undetectability.

The results demonstrate the advantage of our proposed few-shot prompting approach in producing adversarial documents that are not only aligned and rhetorically rich, but also evasive to LLM-based adversarial content detection. In particular, $\texttt{FSAP}_{\texttt{InterQ}}$ is the most effective and realistic attack model, achieving high impact while remaining difficult to filter or flag that could preserve attack goals and be exposed to users on top of the ranked list.

5.3 Impact of Support Set Size (RQ3)

In this section, we explore how the number of few-shot support examples k affects the adversarial effectiveness of $\texttt{FSAP}_{\texttt{InterQ}}$. For this purpose, We experimented with different values of support set size $k \in \{1, 3, 5, 7, 9, 10\}$ and computed the MHDR for each setting across the four target NRMs over both datasets. The results of the experiment are presented in Figure 5. As shown in the figure, the support set size leads to improved MHDR, particularly in the lower range (1 to 5 examples). This trend indicates that few-shot prompting quickly enhances the LLM's ability to generate more effective adversarial content. However, the improvements plateau beyond support size 5, with only marginal gains or slight fluctuations observed. For example, on TREC 2021, MHDR rises rapidly as the support set size increases from 1 to 5 and then it stabilizes near 94–98% across all rankers. These findings suggest that a support set of five examples is generally sufficient to reach near-optimal adversarial performance. Note that we could only run this experiment for $\texttt{FSAP}_{\texttt{InterQ}}$, as it was not feasible to run the same evaluation for $\texttt{FSAP}_{\texttt{IntraQ}}$ due to the variable and limited number of harmful documents available per query.

5.4 Impact of Choice of LLM (RQ4)

In RQ4, we are interested in assessing how the choice of LLM used for adversarial generation affects the attack effectiveness (MHDR), stance alignment, and detection pass rate of the adversarial documents. This analysis focuses on the best-performing baseline method (`Liar Attack`) and our most effective approach ($\texttt{FSAP}_{\texttt{InterQ}}$). Figure 6 provides the results for adversarial documents generated by the `DeepSeek-R1-claude3.7` model on both collections. On TREC 2021, adversarial documents generated using $\texttt{FSAP}_{\texttt{InterQ}}$ with DeepSeek achieve high MHDR, ranging from 93.4% to 99.6% across all the four NRMs. This pattern closely match with `GPT-4o`-generated `FSAP` documents. However, stance alignment is noticeably lower for DeepSeek (75.1%) than for `GPT-4o` (85.2%), which indicates a weaker stance fidelity. In contrast, DeepSeek shows superior detection evasion as 96.5% of its `FSAP`-generated documents are not detected as adversarial, compared to `GPT-4o`'s 94.3%. When comparing DeepSeek's `Liar Attack` generation to its `FSAP` counterpart, we can observe higher stance alignment (74.7%) but lower detection pass rate (92.6%). This shows the trade-off between direct stance conditioning and prompt-based contextual generation.

On TREC 2020, DeepSeek again maintains strong ranking effectiveness for $\texttt{FSAP}_{\texttt{InterQ}}$ (MHDRs above 94.9%) but shows substantially lower stance alignment (31.1%) compared to `GPT-4o` (67.9%). However, its detection pass rate is significantly higher at 97.9%, outperforming `GPT-4o` (82.9%). In contrast, DeepSeek's `Liar Attack` yields higher stance fidelity (35.2%) but achieves a lower detection pass rate (87.6%). We observe that while `GPT-4o` remains the most effective LLM for adversarial generation in terms of stance fidelity and MHDR, the DeepSeek model achieves a highly competitive balance by offering near-equivalent attack performance with even greater undetectability. This demonstrates that high-impact adversarial attacks can be launched using smaller or open-access models without significant compromise in effectiveness.

These results of RQ4 confirm that our method, particularly $\texttt{FSAP}_{\texttt{InterQ}}$, generalizes well across LLMs. Regardless of the underlying generator, $\texttt{FSAP}_{\texttt{InterQ}}$ consistently produces high-ranking adversarial documents, with robustness to shifts in stance alignment and and exhibiting a high detection pass rate. In contrast, the most effective baseline attack, namely `Liar Attack`, demonstrates greater sensitivity to the generator choice, as evidenced by variable performance in stance alignment and significantly lower adversarial detection pass. The consistency of $\texttt{FSAP}_{\texttt{InterQ}}$ across two architecturally distinct LLMs demonstrates that our few-shot adversarial prompting strategy is not overly reliant on specific LLM behavior or memorized patterns.

6 CONCLUDING REMARKS

This paper introduced `FSAP`, a few-shot adversarial prompting framework for generating adversarial documents that are able to deceive

Figure 5: Impact of support set size on MHDR for $\texttt{FSAP}_{\texttt{InterQ}}$ generated documents, evaluated across four ranking models on the TREC 2020 and TREC 2021 datasets.

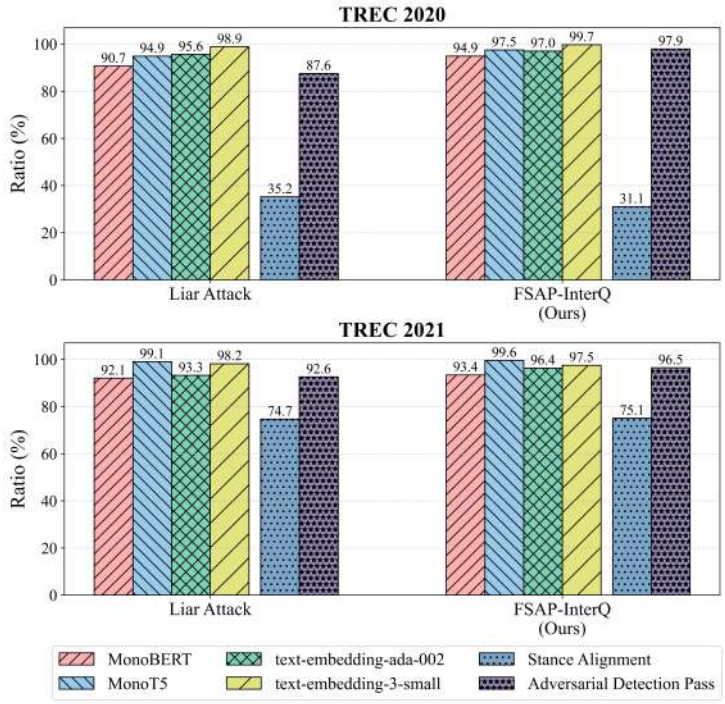

Figure 6: Comparison of MHDR, stance alignment, and adversarial detection pass across NRMs for adversarial documents generated by the `DeepSeek-R1-claude3.7` model on TREC 2020 and 2021.

NRMs. Unlike prior attack strategies focused on token-level perturbation or rewriting, FSAP exploits the in-context learning capabilities of large language models to synthesize realistic, query-conditioned adversarial documents using a small set of harmful exemplars. Through rigorous evaluation on two high-stakes TREC datasets, two distinct LLMs, and diverse NRMs, we showed that $\texttt{FSAP}_{\texttt{InterQ}}$ consistently achieves high attack effectiveness, strong stance alignment, and high undetectability, hence demonstrating its viability as a generalizable and transferable threat model.

As future work, we are interested in expanding the current work in at least two directions:

- *Adversarial generalization theory in in neural ranking models:* We aim to develop a theoretical framework that characterizes the conditions under which adversarial documents generated via few-shot prompting remain effective across diverse queries, ranking models, and LLM architectures. This includes analyzing transferability under distributional shift and deriving generalization guarantees grounded in adversarial risk [28].
- *Game-Theoretic modeling of detection and evasion dynamics:* We plan to formalize the interaction between adversarial generation and detection mechanisms as a game-theoretic problem. By modeling the attacker–defender dynamics, we can study equilibrium strategies that balance attack strength and evasion, enabling design of more robust and anticipatory defenses against few-shot adversarial prompting.

REFERENCES

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

[2] Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications* (2024), 100545.

[3] Amin Bigdeli, Negar Arabzadeh, Ebrahim Bagheri, and Charles LA Clarke. 2024. EMPRA: Embedding Perturbation Rank Attack against Neural Ranking Models. *arXiv preprint arXiv:2412.16382* (2024).

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

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

[6] Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788* (2023).

[7] 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* (2023).

[8] 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 (NIST Special Publication)*, Ian Soboroff and Angela Ellis (Eds.), Vol. 500-335. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec30/papers/Overview-HM.pdf

[9] 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 (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. 1266. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.HM.pdf

[10] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR* abs/2102.07662 (2021). arXiv:2102.07662 https://arxiv.org/abs/2102.07662

[11] 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* (2020).

[12] Rupak Kumar Das and Jonathan Dodge. 2025. Fake News Detection After LLM Laundering: Measurement and Explanation. *arXiv preprint arXiv:2501.18649* (2025).

[13] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).

[14] Zoltán Gyöngyi, Hector Garcia-Molina, et al. 2005. Web Spam Taxonomy.. In *AIRWeb*, Vol. 5. Citeseer, 39–47.

[15] Beizhe Hu, Qiang Sheng, Juan Cao, Yang Li, and Danding Wang. 2025. Llm-generated fake news induces truth decay in news ecosystem: A case study on neural news recommendation. *arXiv preprint arXiv:2504.20013* (2025).

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

[17] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017* (2022).

[18] 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 (2012), 1–30.

[19] 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*. 2025–2039.

[20] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. 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*. 1700–1709.

[21] 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* (2024).

[22] Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[23] 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 (2000), 731–746.

[24] 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. (2016).

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

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

[27] 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 (2013), 10–13.

[28] Muni Sreenivas Pydi and Varun Jog. 2021. The many faces of adversarial risk. *Advances in Neural Information Processing Systems* 34 (2021), 10000–10012.

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

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

[31] Yundi Shi, Piji Li, Changchun Yin, Zhaoyang Han, Lu Zhou, and Zhe Liu. 2022. Promptattack: Prompt-based attack for language models via gradient search. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 682–693.

[32] 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* (2019).

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

[34] Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249* (2024).

[35] Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954* (2024).

[36] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2023. Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838* (2023).

[37] 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*. 115–120.

[38] 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 (2023), 1–27.

[39] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080* (2021).

[40] Jiaqi Xue, Mengxin Zheng, Ting Hua, Yilin Shen, Yepeng Liu, Ladislau Bölöni, and Qian Lou. 2023. Trojllm: A black-box trojan prompt attack on large language models. *Advances in Neural Information Processing Systems* 36 (2023), 65665–65677.

[41] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469* (2023).

[42] Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopal, Katarina Marcincinova, and Matus Mesarcik. 2024. Evaluation of LLM Vulnerabilities to Being Misused for Personalized Disinformation Generation. *arXiv preprint arXiv:2412.13666* (2024).