Datasets for Supervised Adversarial Attacks on Neural Rankers

Amir Khosrojerdi*
amir.khosrojerdi@mail.utoronto.ca
University of Toronto

Amin Bigdeli*
abigdeli@uwaterloo.ca
University of Waterloo

Radin Hamidi Rad
radin.rad@utoronto.ca
University of Toronto

Morteza Zihayat
mzihayat@torontomu.ca
Toronto Metropolitan University

Charles L. A. Clarke
claclark@gmail.com
University of Waterloo

Ebrahim Bagheri
ebrahim.bagheri@utoronto.ca
University of Toronto

ABSTRACT

We introduce a novel resource for *adversarial rank attacks against neural rankers* designed to support systematic research on the robustness of neural retrieval models. Existing adversarial methods for ranking are often unsupervised, rely on surrogate models, and lack ground-truth supervision. Our dataset addresses these limitations by leveraging Retrieval-Augmented Generation (RAG) with a Large Language Model (LLM) to construct high-quality adversarial examples that subtly manipulate document rankings while preserving linguistic coherence and indirect relevance. The dataset is generated through a self-refining LLM–Ranker feedback loop and released in two variants, Gold and Diamond, categorized by attack effectiveness. We provide comprehensive metadata, insertion points, ranking labels, and quality metrics (fluency, acceptability) for each instance. Accompanied by code and LLM prompts, our resource supports both training and evaluation of adversarial models and can serve as a benchmark for ranking robustness. This work offers a reproducible and extensible foundation for the development of robust retrieval systems and adversarial information retrieval methods.

1 INTRODUCTION

Neural Ranking Models (NRMs) have demonstrated state-of-the-art performance in Information Retrieval (IR) tasks by leveraging deep contextual representations and large-scale training corpora [10, 15]. Nevertheless, a growing body of work has identified that these models are susceptible to adversarial cases, which include document perturbations [12] that are imperceptible or semantically marginal yet induce significant changes in the model's output ranking [1, 3, 11, 18, 20]. The manipulation of retrieval rankings via adversarial content has raised substantial concerns not only in academic settings but also in deployed systems, where such vulnerabilities may be exploited for adversarial search engine optimization (SEO), visibility manipulation, or dissemination of low-quality content [2, 5, 8]. These observations motivate a systematic and replicable investigation into the mechanisms and limitations of ranking robustness [4, 12] under adversarial conditions.

From a theoretical standpoint, adversarial vulnerability in NRMs stems from the complex and non-linear relationship between input representations and learned ranking functions [10, 13, 16, 23]. This is especially pronounced in neural architectures that operate in high-dimensional semantic spaces, where small shifts in document embeddings can yield disproportionate changes in similarity scores [7, 15]. In adversarial settings, this phenomenon is often exploited by introducing content that is not directly relevant to the query but contributes latent semantic features aligned with the learned notion of relevance. Crucially, to distinguish adversarial manipulation from legitimate content augmentation, it becomes necessary to formalize a *notion of indirect relevance*, wherein added text increases ranking scores without addressing the query explicitly. This requirement introduces a semantic constraint that prior adversarial IR methods typically ignore, as most rely on syntactic or heuristic perturbations, trigger-based augmentations, or query-injection techniques that trivially improve rank by increasing lexical overlap [11, 18].

While existing work has introduced attack strategies for NRMs, these approaches are often grounded in unsupervised optimization or reinforcement learning with surrogate models [1, 3]. The absence of datasets with supervision over adversarial impact and quality has limited the reproducibility, benchmarking, and theoretical analysis of ranking attacks. More fundamentally, without access to controlled adversarial corpora, it is difficult to analyze the generalization behavior of NRMs under perturbed conditions or to quantify the inductive biases that make them susceptible to particular adversarial patterns.

To address this gap, we present a principled framework for constructing retrieval-conditioned adversarial datasets, grounded in supervised generation and guided by formal constraints on fluency, coherence, and indirect relevance. Our approach leverages Retrieval-Augmented Generation (RAG) [9], wherein a Large Language Model (LLM) is prompted with a query, a target document, and a retrieval context of top-ranked documents. This setting mimics the retrieval condition that occurs in ranking systems, and prompts the LLM to generate content that is topically coherent with the query but does not explicitly answer it. The use of a contextual LLM generator, rather than token-level perturbations, ensures that generated sentences are fluent and plausibly embedded in natural discourse [6, 22].

To formalize our generation criteria, we define three objectives: (1) the adversarial modification must preserve fluency and coherence with the document's style and structure; (2) the added content must exhibit indirect relevance, avoiding direct entailment or paraphrasing of the query; and (3) the modified document must exhibit a measurable improvement in rank under a given NRM. These constraints enable a rigorous definition of adversariality in the ranking context, one that goes beyond lexical overlap or syntactic mutation and aligns with the epistemic and operational semantics of modern retrieval systems.

Furthermore, we introduce a self-refinement mechanism that implements an iterative feedback loop between the LLM and the neural ranker. This loop serves two purposes. First, it increases the likelihood that generated content satisfies the ranking constraint by providing the LLM with reward signals based on actual retrieval outcomes. Second, it enables an implicit approximation of gradient-based optimization over black-box rankers by conditioning future

*Both authors contributed equally to this research.

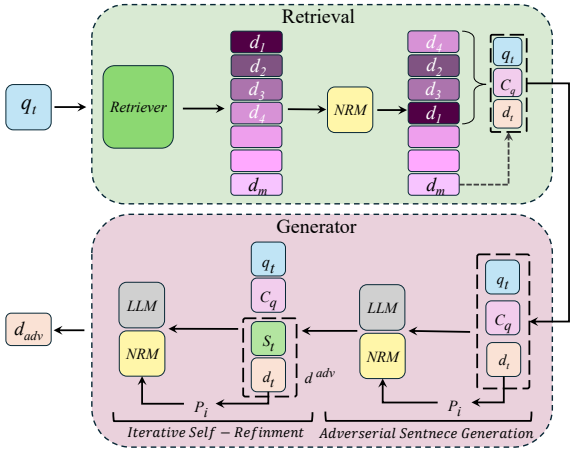

Figure 1: Overview of our proposed workflow.

generations on previously successful perturbations. The iterative refinement structure can be viewed as an application of approximate policy iteration, where the LLM updates its generative policy based on reward signals from the ranking environment.

We instantiate this framework as a dataset and codebase comprising two variants: the *Gold* dataset, which contains the most effective adversarial sentence per document-query pair, and the *Diamond* dataset, which enforces stricter ranking thresholds to isolate high-impact attacks. Each dataset entry is annotated with query-document pairs, insertion position, ranking deltas, fluency scores (via perplexity), and syntactic acceptability metrics, thereby enabling fine-grained evaluation and comparative studies. The datasets are derived from the MS MARCO passage ranking benchmark [14] using a target NRM, and all generation prompts, refinement code, and evaluation scripts are released to ensure transparency and reproducibility.

The primary contributions of this work are threefold. **First**, we introduce the first supervised, retrieval-conditioned dataset for adversarial document ranking that satisfies formal semantic and ranking constraints. **Second**, we provide a complete generation and evaluation pipeline that includes a chain-of-thought refinement mechanism grounded in observable ranking signals. **Third**, we deliver a reusable and extensible resource for the community, supporting downstream research in adversarial IR, robust ranking model evaluation, retrieval-based LLM augmentation, and generative attack modeling.

2 PROPOSED APPROACH

The objective of this resource paper is to introduce a systematic approach for constructing datasets of adversarially modified documents that expose the susceptibility of NRMs to content-level manipulation under realistic retrieval conditions. Each adversarial example is constructed by injecting a sentence into a target document such that the perturbation increases the document's ranking for a given query without introducing obvious semantic artifacts. To ensure the quality and adversarial character of these modifications, we impose three constraints: (O_1) the injected content must preserve fluency and internal coherence with the original document; (O_2) the sentence must exhibit indirect relevance, influencing the ranking without directly satisfying the query intent; and (O_3) the document must achieve a strictly improved ranking under the target NRM after insertion.

2.1 Problem Definition

Let $\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$ denote n user queries submitted to an IR system. For each query $q \in \mathcal{Q}$, the system returns a ranked list of documents $\mathcal{D}_q = \{d_1, d_2, \ldots, d_m\}$, where the ranking is determined by a scoring function $f_{\text{rel}} : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}$, such that:

$$\text{rank}(q, d) = \sum_{d' \in \mathcal{D}_q} 1\left[f_{\text{rel}}(q, d') > f_{\text{rel}}(q, d)\right] \tag{1}$$

Given a query q, a target document $d_t \in \mathcal{D}_q$, and a rank threshold $k \in \mathbb{N}$, the goal is to construct a perturbed version d_t^{adv} of the document such that $\text{rank}(q, d_t^{\text{adv}}) \leq k$, subject to the required objectives (O_1–O_3). The problem is to find such a transformation function $\phi : \mathcal{D} \times \mathcal{Q} \to \mathcal{D}^{\text{adv}}$ that satisfies the above ranking and semantic constraints for a given retrieval model f_{rel}.

2.2 Approach Overview

Figure 1 depicts our multi-stage pipeline that integrates LLM generation and NRM-based evaluation. For each query, we retrieve the top-ranked context documents and prompt the LLM with the query, target document, and retrieval context to generate candidate adversarial sentences (**adversarial sentence generation**). These are inserted at various positions within the target document to create perturbed variants. Each variant is passed through the NRM to compute its rank relative to the original. Only variants satisfying all three constraints are retained in the dataset. The process is iterated through a refinement loop (**self-refinement loop**) in which the LLM conditions future generations on previously successful candidates, enabling more effective and diverse perturbations over time. Each finalized dataset instance includes the original and perturbed documents, the insertion position, ranking delta, and fluency and acceptability metrics, supporting both evaluation and training of adversarial IR models. We introduce this process more formally in the following.

2.3 Adversarial Dataset Construction

Given a query $q \in \mathcal{Q}$, a target document $d_t \in \mathcal{D}_q$, and a set of top-c context documents $C_q = \{d_1, d_2, \ldots, d_c\} \subset \mathcal{D}_q$, we define an adversarial sentence generator $\mathcal{G} : \mathcal{Q} \times \mathcal{D} \times C \to \mathcal{S}$, which outputs a set of candidate adversarial sentences $\mathcal{S} = \{s_1, s_2, \ldots, s_l\}$. Each candidate sentence $s \in \mathcal{S}$ is inserted into the original target document d_t at position $p \in \mathcal{P}(d_t)$, where $\mathcal{P}(d_t)$ denotes the set of allowable insertion points (e.g., sentence boundaries), yielding a perturbed document $d_{t(p,s)}^{\text{adv}} = \text{Insert}(d_t, s, p)$. We include $d_{t(p,s)}^{\text{adv}}$ in the adversarial dataset $\mathcal{D}_{\text{adv}}$ if and only if the following constraints are satisfied: **(a)** *Ranking Improvement*: $\text{rank}(q, d_{t(p,s)}^{\text{adv}}) \leq \text{rank}(q, d)$; **(b)** *Indirect Relevance*: $s \not\models q$, i.e., the sentence does not explicitly or implicitly answer the query; **(c)** *Linguistic Coherence*: $d_{t(p,s)}^{\text{adv}}$ preserves the fluency and semantic consistency of d_t, as judged by a coherence function $\psi(d_{t(p,s)}^{\text{adv}}) > \tau$ for threshold τ. This process yields a dataset $\mathcal{T} = \{(q_i, C_{q_i}, d_t, s, p)\}_{i=1}^{N}$ of adversarial training examples, where each s_i is a validated perturbation at position p that satisfies constraints (a), (b) and (c).

To construct this dataset, we implement a generation pipeline that combines an LLM and a NRM in an iterative loop. The LLM $\mathcal{G}(q, d_t, C_q)$ proposes candidate adversarial sentences based on a query q, a target document d_t, and the set of context documents C_q.

Each sentence is inserted into the target document and evaluated by the NRM $\mathcal{R}(q, d^{\text{adv}}_{t(p,s)})$, which measures its impact on the document's ranking. If a candidate yields a valid rank improvement within the k threshold, while preserving fluency and semantics, it's retained; otherwise, the LLM enters a refinement phase, using ranking feedback to improve its generations. This feedback loop produces a curated dataset of adversarial examples, each validated and quality-filtered.

Adversarial Sentence Generation. Given the query q, our approach first constructs the contextual basis for the LLM generator $\mathcal{G}$ by selecting the top-c ranked documents as the context set $C_q = \{d_1, d_2, \ldots, d_c\}$, where c represents the number of documents incorporated as context for adversarial sentence generation. In order to generate adversarial sentences, we prompt the LLM to produce an initial set $\mathcal{S}_t = \mathcal{G}(q, d_t, C_q)$. Each generated sentence $s \in \mathcal{S}_t$ is then inserted at position p, creating an adversarially modified document.

Next, the NRM model $\mathcal{R}$ evaluates the ranking impact of the adversarial sentence being injected into the document as follows:

$$\text{rank}(q, d^{\text{adv}}_{t(p,s)}) = \mathcal{R}(q, d^{\text{adv}}_{t(p,s)}) \tag{2}$$

If the modified document achieves rank $\text{rank}(q, d^{\text{adv}}_{t(p,s)})$, the process terminates; otherwise, the system enters the feedback loop.

Iterative Self-Refinement. If no adversarial sentence yields the desired ranking improvement, the system refines the sentences iteratively. For this purpose, the NRM component selects the most effective adversarial sentences:

$$\mathcal{S}^{p,\text{high}}_t = \underset{s \in \mathcal{S}^p_t}{\arg\min}\, \text{rank}(q, d^{\text{adv}}_{t(p,s)}) \tag{3}$$

The most effective adversarial sentences from the previous step, $\mathcal{S}^{p,\text{high}}_t$, are used to guide the language model in generating refined candidates. These are then inserted into the original document d_t at position p to produce new adversarial versions. Each updated document is evaluated by the neural ranking model (NRM) to assess its impact on the query-document ranking. This process repeats until the adversarial document reaches a top-k rank or the maximum number of iterations is reached.

3 IMPLEMENTATION SETUP

All dataset variations along with the code and LLM prompts are publicly available in our GitHub repository[1].

Model Details. To implement our RAG-based adversarial dataset construction, we use Qwen3 32B [21], selected due to its strong language understanding, context-aware generation, and fluency preservation. The model is prompted with Target Query, Target Document, and Top-ranked Documents retrieved via a NRM. We select `msmarco-MiniLM-L-12-v2` cross encoder [17] for ranking evaluation. We set the maximum number of refinement iterations to 5, selected the top-5 ranked documents as the context, and stopped refinement once an adversarial document reached the top-10 (k=10).

Establishing Benchmarks. We evaluate our datasets against strong baselines across four attack types: word-level, trigger-based, sentence-level, and prompt-based. **(1)** `PRADA` [20] replaces up to 20 key tokens with synonyms using a surrogate ranker. **(2)** `Brittle-BERT` [18] prepends up to 12 trigger tokens to the document. **(3)** `PAT` [11] inserts trigger words at the beginning based on surrogate model cues. **(4)** `IDEM` [3] injects up to 500 BART-generated connector sentences. **(5)** `EMPRA` [1] perturbs sentence embeddings to align with query-relevant semantics while preserving meaning.

Evaluation Metrics. Following prior work [1, 3, 20], we assess attack effectiveness and text quality based on several metrics, namely *Attack Success Rate (ASR)* is the percentage of documents with rank improvement. *Boosted Top-10* and *Boosted Top-50* report the proportion of documents reaching rank $\leq$ 10 or 50, respectively. *Boost* captures average rank improvement. *Perplexity (PPL)* (via `GPT-2`) evaluates fluency; lower is better. *Acceptability Score (AcS)* uses a neural classifier [19] to score the acceptability of text.

Source Data. Consistent with prior studies [1, 3, 11, 18, 20], we establish the benchmarks using the MS MARCO passage V1 collection [14]. Building the foundation of our proposed datasets on this large-scale, and well-annotated corpus, we deliver an adversarial attack dataset suited for evaluating IR system robustness. Following the approach of [1, 3], we randomly sampled subset of 1,000 queries from the MS MARCO dev small. For each query, we randomly select two distinct types of documents, Easy-5 and Hard-5, selected from the victim NRM's re-ranking of the top-1 K BM25 candidates. *Easy-5* documents consists of five documents initially ranked between positions 51 and 100, sampled at equal intervals (e.g., ranks 51, 63, 76, 84, 91) to assess mid-rank improvements. In contrast, *Hard-5* group comprises the five lowest-ranked documents, representing the most challenging cases for rank boosting. For each query–document pair, we release a RAG-based adversarially augmented dataset.

Our Dataset Variations. To capture a broad spectrum of augmentation scenarios, we categorize our dataset into two variations based on the level of rank improvement: *(1) Gold Dataset.* For every unique query and target document pair, we select the best-performing sentence among positions. This dataset variation isolates the most optimal augmentation instances for boosting the ranking of the target document. *(2) Diamond Dataset.* This dataset includes only instances where the augmented document achieves a final rank improvement beyond a predefined threshold. Specifically, we include documents that reach a final rank of $\leq$10 for Easy-5 documents and $\leq$50 for both Easy-5 and Hard-5 documents. This strict criteria ensures that the Diamond dataset captures the most impactful augmentation cases.

4 DATASET EVALUATION

We evaluate our datasets to demonstrate their utility as practical and rigorous benchmarks for adversarial IR. The Gold and Diamond variants are constructed to reflect various realistic adversarial scenarios, enabling researchers to study the impact of document perturbations across varying levels of ranking difficulty and insertion strategies.

Gold Dataset Performance. The Gold dataset offers a curated collection of adversarial augmentations that satisfy all three constraints—fluency, indirect relevance, and rank improvement—for each query–document pair. As shown in Table 1, it achieves a nearly perfect Attack Success Rate (ASR) across both Easy-5 and Hard-5 scenarios, demonstrating that the RAG-based generation method consistently produces effective, query-conditioned perturbations. In Easy-5 group, 62.2% of documents are boosted into the top-10 (%r $\leq$ 10) and 93.1% into the top-50 (%r $\leq$ 50), with an average boost of 59.2 rank positions. In Hard-5, we can see that 29.1% of adversarial documents reach the top-10 and 49.4% the top-50, with an average boost of 781.1 ranks. This demonstrates that our Gold

[1] https://github.com/KhosrojerdiA/adversarial-datasets

Table 1: Gold Dataset statistics over Easy-5 and Hard-5.

Group	Type	SentPos	Counts	ASR	%r ≤ 10	%r ≤ 50	Boost	PPL	AcS
Easy-5	Orig.	–	5,000	–	–	–	–	37.3	0.78
	Adv.	all	5,000	99.7	62.2	93.1	59.2	42.4	0.77
		v=0	3,134	100.0	72.9	96.6	63.2	42.2	0.77
		v=1	961	99.7	54.4	92.5	57.7	42.7	0.77
		v=2	371	99.7	44.5	87.9	51.6	40.0	0.76
		v=3	163	99.4	33.1	85.9	50.6	40.4	0.73
		v≥4	371	98.1	22.4	72.5	40.5	46.5	0.77
Hard-5	Orig.	–	5,000	–	–	–	–	51.4	0.72
	Adv.	all	5,000	99.9	29.1	49.4	781.1	67.4	0.72
		v=0	3,281	99.9	34.5	55.0	802.5	66.9	0.73
		v=1	854	100.0	25.3	44.7	773.3	67.2	0.69
		v=2	341	100.0	14.1	36.1	726.9	66.1	0.70
		v=3	142	100.0	7.7	28.9	701.6	64.6	0.63
		v≥4	382	100.0	12.8	31.2	692.3	74.0	0.72

Table 2: Diamond Dataset statistics over Easy-5 and Hard-5.

Group	Type	SentPos	Counts	%r ≤ 10	Boost	PPL	AcS
Easy-5	Orig.	—	3,110	–	–	37.6	0.78
	Adv.	all	3,110	100.0	69.7	41.9	0.77
		v=0	2,285	100.0	69.7	41.3	0.78
		v=1	523	100.0	69.7	44.8	0.76
		v=2	165	100.0	70.5	39.3	0.74
		v=3	54	100.0	67.3	38.4	0.75
		v≥4	83	100.0	70.2	48.7	0.74
Hard-5	Orig.	—	2,470	–	–	55.8	0.71
	Adv.	all	2,470	58.9	983.1	73.1	0.71
		v=0	1,805	62.7	984.0	70.8	0.72
		v=1	382	56.5	982.3	79.9	0.66
		v=2	123	39.0	979.3	74.8	0.71
		v=3	41	26.8	974.9	76.2	0.63
		v≥4	119	41.2	979.3	82.0	0.69

dataset can consistently improve the ranking of both Easy-5 and Hard-5 documents to higher positions.

We further explore the impact of Sentence Position (SentPos) of adversarial insertions. For Easy-5, injecting adversaries at the first position yields a perfect 100% ASR, with 72.9% of documents boosted into the top-10 and a rank boost of 63.2. The attack effectiveness reduces slightly at later positions. Documents modified at positions four or beyond still achieve 98.1% ASR but deliver a smaller rank improvement (40.5). Hard-5 exhibits a similar pattern as the earliest insertion point produces the greatest effectiveness, while deeper insertions remain effective but with gradually reduced impact.

In addition to significant rank improvements, the linguistic quality of the adversarial augmentations remains high. Comparing adversarial (Adv.) documents with their original counterparts (Orig.), we can see that perplexity increases slightly from 37.3 to 42.4 on Easy-5 and from 51.4 to 67.4 on Hard-5, indicating slightly higher uncertainty but still within acceptable ranges. Also, acceptability score (AcS) remain at a comparable level. Easy-5 experience AcS drop only from 0.78 to 0.77, and Hard-5 holds at 0.72, indicating that adversarial sentences continue to be perceived as fluent and well-formed text.

Diamond Dataset Performance. The Diamond dataset offers an even more selective view of adversarial effectiveness by including only those perturbations that exceed strict rank-improvement thresholds. As detailed in Table 2, the dataset isolates cases where adversarial content leads to substantial ranking jumps, making it a strong resource for stress-testing ranking models and examining worst-case vulnerabilities. Overall, there are 3,110 query–document pairs in the Easy-5 group, for which augmentations achieve 100% boosted top 10 with a mean boost of 69.7 ranks. Hard-5 remains more challenging as only 2,470 of augmentations achieve 100% boosted top 50 with 58.9% boosted top 10, and an average lift of 983.1 ranks. Exploring the insertion positions shows that for Easy-5, inserts at the beginning achieve the strongest results (69.7 rank boost), with only minor variation at deeper positions. In Hard-5, the earliest insertion delivers the best outcome (boosted top-10 of 62.7% and a 984.0 rank boost), while deeper positions see a gradual decline in performance.

Table 3: Performance comparison across baselines and ours.

Model	Easy-5					Hard-5				
	ASR	%r ≤ 10	%r ≤ 50	Boost	PPL	ASR	%r ≤ 10	%r ≤ 50	Boost	PPL
PRADA	76.9	2.6	44.7	22.1	96.8	75.7	0.0	2.2	98.5	202.8
Brittle-BERT	94.1	61.6	90.7	56.8	118.1	100.0	29.6	72.7	933.1	173.2
PAT	51.2	3.5	25.0	2.9	47.0	80.0	0.0	1.1	119.2	70.2
IDEM	99.3	82.4	97.4	67.0	38.8	99.9	45.5	80.2	926.3	61.8
EMPRA	99.7	87.9	99.2	69.3	37.3	99.4	51.3	80.9	904.8	55.0
Ours	100.0	100.0	100.0	69.7	41.9	100.0	58.9	100.0	983.1	73.1

Despite the remarkable effectivness, linguistic quality remains high. Easy-5 adversarial PPL increases modestly from 37.6 to 41.9, and Hard-5 PPL rises from 55.8 to 73.1, indicating slightly higher uncertainty by a fluency model yet still fluent text. Acceptability scores remain stable for Easy-5, decreasing slightly from 0.78 to 0.77, and for Hard-5 remaining at 0.71, demonstrating that even the most aggressive augmentations pass acceptability checks.

Comparative Evaluation. To demonstrate the utility of our proposed datasets, we conduct a comparative analysis using state-of-the-art adversarial attack methods evaluated on our Diamond dataset. The results, presented in Table 3, illustrate how our dataset enables robust and consistent evaluation across a diverse set of adversarial strategies. Overall, our Diamond dataset augmentations can achieve the highest attack performance metrics across both Easy-5 and Hard-5. Specifically, it achieves average boost of 69.7 for Easy-5 and 983.1 for Hard-5. While the strongest baseline, EMPRA, records a comparable Easy-5 boost of 69.3 and a slightly lower Hard-5 boost of 904.8, our Diamond's augmentations still achieves broader coverage by boosting every document into the top-10 in Easy-5 and into top-50 in both groups. In addition to strong effectiveness, our proposed dataset maintains competitive fluency by having the perplexity of 41.9 and 73.1 for Easy-5 and Hard-5, respectively. Although EMPRA and IDEM achieve lower perplexities, our dataset offers a well-balanced trade-off between attack effectiveness and fluency. This balance of high attack performance and low linguistic distortion underlines the strength of our retrieval-augmented generation framework for generating stealthy and effective adversarial texts.

5 CONCLUDING REMARKS

We present a structured and supervised resource for studying adversarial vulnerabilities in NRMs. Using adversarial sentence generation with iterative refinement and rank-based evaluation, our dataset enables fine-grained benchmarking, controlled training, and reproducible experimentation. The released data, codebase, and generation pipeline provide a practical foundation for advancing robustness research in information retrieval, supporting both attack modeling and defense evaluation. We anticipate this resource will facilitate new lines of inquiry into ranking resilience, adversarial generalization, and the interplay between generative models and retrieval systems.

6 GEN-AI USAGE DISCLOSURE

In compliance with the ACM authorship policy and CIKM 2025 guidelines, we hereby disclose our use of generative AI tools in the research process. Specifically, we employed the open-source generative language model Qwen3-32B [21] to generate adversarial sentences as part of our retrieval-augmented generation (RAG) based dataset construction pipeline. We also utilized this model to refine and optimize prompts used during the data generation process, ensuring more effective adversarial examples. Importantly, no generative AI was used to write the text of this manuscript; all writing and editing of the paper were done exclusively by the human authors without AI assistance. We affirm that our use of GenAI tools adheres to ACM's authorship and GenAI usage disclosure policies. The above details fully disclose how generative AI was utilized in our work, as required. We confirm that these practices are in line with the official guidelines and that the integrity of authorship for this paper remains uncompromised by our limited use of AI tools.

REFERENCES

[1] 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).

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

[3] Xuanang Chen, Ben He, Zheng Ye, Le Sun, and Yingfei Sun. 2023. Towards Imperceptible Document Manipulations against Neural Ranking Models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 6648–6664. https://doi.org/10.18653/V1/2023.FINDINGS-ACL.416

[4] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2018. Ranking robustness under adversarial document manipulations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 395–404.

[5] Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Web Spam Taxonomy. In *AIRWeb 2005, First International Workshop on Adversarial Information Retrieval on the Web, co-located with the WWW conference, Chiba, Japan, May 2005*. 39–47. http://airweb.cse.lehigh.edu/2005/gyongyi.pdf

[6] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=rygGQyrFvH

[7] Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. 2021. How Does BERT Rerank Passages? An Attribution Analysis with Information Bottlenecks. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 496–509. https://doi.org/10.18653/v1/2021.blackboxnlp-1.39

[8] Alex Goh Kwang Leng, P Ravi Kumar, Ashutosh Kumar Singh, and Anand Mohan. 2012. Link-based spam algorithms in adversarial information retrieval. *Cybernetics and Systems* 43, 6 (2012), 459–475.

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[10] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2022. *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature.

[11] 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, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi (Eds.). ACM, 2025–2039. https://doi.org/10.1145/3548606.3560683

[12] Yu-An Liu, Ruqing Zhang, Mingkun Zhang, Wei Chen, Maarten de Rijke, Jiafeng Guo, and Xueqi Cheng. 2024. Perturbation-Invariant Adversarial Training for Neural Ranking Models: Improving the Effectiveness-Robustness Trade-Off. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 8832–8840. https://doi.org/10.1609/AAAI.V38I8.28730

[13] Bhaskar Mitra and Nick Craswell. 2017. Neural models for information retrieval. *arXiv preprint arXiv:1705.01509* (2017).

[14] 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. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[15] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[16] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/V1/D19-1410

[18] Yumeng Wang, Lijun Lyu, and Avishek Anand. 2022. BERT Rankers are Brittle: A Study using Adversarial Document Perturbations. In *ICTIR '22: The 2022 ACM SIGIR International Conference on the Theory of Information Retrieval, Madrid, Spain, July 11 - 12, 2022*, Fabio Crestani, Gabriella Pasi, and Éric Gaussier (Eds.). ACM, 115–120. https://doi.org/10.1145/3539813.3545122

[19] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7 (2019), 625–641.

[20] 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 Trans. Inf. Syst.* 41, 4 (2023), 89:1–89:27. https://doi.org/10.1145/3576923

[21] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[22] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2024. A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models. *ACM Comput. Surv.* 56, 3 (2024), 64:1–64:37. https://doi.org/10.1145/3617680

[23] Lixin Zou, Shengqiang Zhang, Hengyi Cai, Dehong Ma, Suqi Cheng, Shuaiqiang Wang, Daiting Shi, Zhicong Cheng, and Dawei Yin. 2021. Pre-trained language model based ranking in Baidu search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4014–4022.