Towards Imperceptible Document Manipulations against Neural Ranking Models

Xuanang Chen[1,2] **Ben He**[1,2*] **Zheng Ye**[3*] **Le Sun**[2*] **Yingfei Sun**[1*]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]South-Central University for Nationalities, Wuhan, China
chenxuanang19@mails.ucas.ac.cn, benhe@ucas.ac.cn
yezheng@scuec.edu.cn, sunle@iscas.ac.cn, yfsun@ucas.ac.cn

Abstract

Adversarial attacks have gained traction in order to identify vulnerabilities in neural ranking models (NRMs), but current attack methods often introduce noticeable errors. Moreover, current methods rely heavily on using a well-imitated surrogate NRM to guarantee the attack effect, making them difficult to use in practice. This paper proposes a framework called Imperceptible DocumEnt Manipulation (IDEM) to produce adversarial documents that are less noticeable to both algorithms and humans. IDEM instructs a well-established generative language model like BART to generate error-free connection sentences, and employs a separate position-wise merging strategy to balance between relevance and coherence of the perturbed text. Evaluation results on the MS MARCO benchmark demonstrate that IDEM outperforms strong baselines while preserving fluency and correctness of the target documents. Furthermore, the separation of adversarial text generation from the surrogate NRM makes IDEM more robust and less affected by the quality of the surrogate NRM.

1 Introduction

Adversarial Information Retrieval (AIR) has been a topic of significant attention from the research community (Davison et al., 2006; Leng et al., 2012; Farooq, 2019). It refers to the scenario in which a portion of the collection can be maliciously manipulated, with Adversarial Web Search (Castillo and Davison, 2010) being a typical example. In the context of the Web, this type of manipulation is commonly referred to as black-hat search engine optimization (SEO) or Web spamming, whose goal is to deceive ranking algorithms by artificially inflating the relevance of targeted Web pages, resulting in undeservedly high rankings for those pages (Gyöngyi and Garcia-Molina, 2005). Meanwhile, in the last few years, neural ranking models (NRMs), particularly those that utilize pre-trained language models (PLMs), have demonstrated remarkable performance across a diverse range of text ranking tasks (Lin et al., 2021). Furthermore, these NRMs have also been implemented in various industrial applications (Lin et al., 2021), such as Web search engines (Zou et al., 2021), to improve the accuracy and relevance of search results.

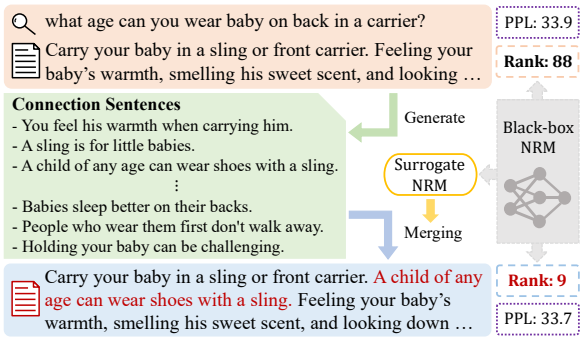

Figure 1: In IDEM, a generative language model is instructed to generate a group of connection sentences between the query and the target document, then a position-wise merging strategy is applied to select and position an optimal connection sentence within the target document, in order to promote the ranking (from 88th to 9th) but maintain the fluency measured by perplexity (PPL).

However, recent studies have revealed the vulnerabilities of NRMs to adversarial document manipulations (Wu et al., 2022; Liu et al., 2022; Wang et al., 2022; Song et al., 2022), that is, small deliberate perturbations in the input documents can cause a catastrophic ranking disorder in the outcome of NRMs. This highlights the need to investigate and identify the potential weaknesses of NRMs before their deployment, in order to ensure their robustness and prevent potential risks. Several manipulation techniques have been proposed to maliciously boost the ranking of low-ranked documents, such as PRADA (Wu et al., 2022) and PAT (Liu et al., 2022). Although these adversarial attack methods have shown the ability to fool NRMs by replacing crucial words, or appending query text or other ad-

Findings of the Association for Computational Linguistics: ACL 2023, pages 6648–6664
July 9-14, 2023 ©2023 Association for Computational Linguistics

versarial tokens, they are still subject to two major limitations. Firstly, existing attack methods tend to introduce grammatical errors, impossible expressions, or incoherent text snippets into the original document, making the attack easy to mitigate, such as by perplexity filters (Song et al., 2020) or grammar checkers (Liu et al., 2022). Secondly, existing attack methods heavily rely on a well-imitated surrogate NRM to produce adversarial text, but this requires a lot of in-domain training data collected by querying the victim NRM, which can be infeasible or even unavailable in real-world situations.

To this end, we propose IDEM, an imperceptible document manipulation framework that aims to produce adversarial documents that are less perceptible to both humans and algorithms. As depicted in Figure 1, a well-established generative language model (GLM), such as BART, is first engineered to generate a series of grammatically correct connection sentences between the query and the target document, a position-wise merging mechanism is then employed to select an optimal connection sentence to be appropriately positioned within the target document. During the generation of connection sentences, we take advantage of the language modeling objective in the GLM without adding any ranking-incentivized objective. This not only helps to produce more natural and fluent text without introducing extra errors that are easy to detect, but also separates the surrogate NRM from the generation process to substantially reduce the dependence of attack performance on the surrogate NRM. Extensive experiments carried out on the widely-used MS MARCO passage ranking dataset indicate that IDEM is able to achieve better attack performance against black-box NRMs than recent baselines, regardless of whether the surrogate NRM is similar to or far from the victim NRM. According to both automatic and human evaluations, the adversarial documents produced by IDEM can maintain semantic fluency and text quality.

Our contributions are three-fold: 1) We propose an imperceptible document manipulation framework, IDEM[1], that employs contextualized blank-infilling and coherent merging for the ranking attack. 2) Extensive attack experiments under three types of surrogate NRMs show that IDEM is able to robustly promote the rankings of the target documents. 3) Automatic and human evaluations of text quality indicate that IDEM is capable of generating more natural and fluent adversarial documents.

2 Problem Statement

Task description. Typically, given a query q and candidates $\mathcal{D} = \{d_1, d_2, \cdots, d_{|\mathcal{D}|}\}$ collected by a retrieval model (e.g., BM25; Robertson et al., 1995) from the whole corpus, a NRM $\mathcal{M}_V$ (i.e., the victim NRM) produces relevance scores $s(q, d_i)$ for all query-document pairs, outputting a re-ranking list $\mathcal{L} = [d_1, d_2, \cdots, d_{|\mathcal{D}|}]$, wherein $s(q, d_1) > s(q, d_2) > \cdots > s(q, d_{|\mathcal{D}|})$. In the adversarial ranking attack, an adversarial document is constructed by applying a deliberate perturbation p_i to the target document $d_i \in \mathcal{L}$ such that it can be ranked higher by the victim NRM. Existing attack methods search and find perturbations on different levels of granularity, such as replacing important words with synonyms (e.g., PRADA) or adding an extra text piece (e.g., PAT). Ideally, adversarial documents should display semantic consistency and fluency to avoid detection. However, there is still room for improvement regarding the imperceptibility of adversarial examples produced by existing methods, such examples can be found in Appendix A.5.

Task setting. For consistency with real-world situations (e.g., black-hat SEO), akin to recent studies (Wu et al., 2022; Liu et al., 2022), this work focuses on the decision-based black-box attack setting, where attackers can only obtain a limited number of queries $\{q_i\}_{i=1}^m$ and their corresponding ranking lists $\{\mathcal{L}_i = [d_{i,1}, d_{i,2}, \cdots, d_{i,|\mathcal{D}|}]\}_{i=1}^m$ with rank positions by querying the victim NRM $\mathcal{M}_V$, but have no access to the exact relevance scores, as well as the architecture, parameters, gradients, or training data of the victim NRM. In this setting, a weak-supervised training data $\mathcal{T}_S = \{(q_i, d_{i,j}, d_{i,k})\}_{i=1}^m$ is usually collected on basis of $\{(q_i, \mathcal{L}_i)\}_{i=1}^m$, wherein $1 \leq j < k \leq |\mathcal{D}|$ so that $d_{i,j}$ and $d_{i,k}$ are considered as relatively positive and negative documents, respectively. A surrogate NRM $\mathcal{M}_S$ is trained on this $\mathcal{T}_S$ to imitate the victim NRM $\mathcal{M}_V$ using a pairwise loss, such as hinge loss (Wu et al., 2022). Besides, existing methods require this surrogate NRM to be functionally similar to the victim NRM, as it determines the direction of adversarial attack (Liu et al., 2022). This means the attack performance of existing methods heavily depends on the surrogate NRM, but it needs lots of in-domain training samples (i.e., m is large) as well as the associated cost of querying the victim NRM, while our proposed

[1]We make our code and data publicly available at https://github.com/cxa-unique/IDEM.

IDEM can achieve reliable attack effect even with an out-of-domain (OOD) surrogate NRM.

3 Method

As outlined in Figure 1 and Algorithm 1, IDEM contains two stages, the first is to obtain a series of connection sentences that can bridge the semantic gap between the query and the target document. After that, these connection sentences are considered at all positions in the inter-sentences of the target document to trade-off the semantic fluency and attacked relevance, and the best one is output as the final adversarial document.

3.1 Connection Sentences Generation

To obtain more natural and fluent connection sentences, we choose to draw support from the well-established GLMs, such as BART in our default setting. The BART model (Lewis et al., 2020) has been pre-trained using a text infilling loss function, which enables it to fill in blanks within the context in a more flexible way. Hence, in this section, we use the BART model as an example to illustrate how to engineer GLMs in a blank-infilling pattern to generate connection sentences that include information about both the query and the document.

Specifically, given a query q and a target document $d_i \in \mathcal{L}$, which is simplified as d from here, we concatenate them with a blank in between as seen in the template as follows.

$$q \oplus \text{It is known that } \_\_\_\_ \oplus d \qquad (1)$$

wherein query q ends with appropriate punctuation, and the blank is replaced by a single special token defined and used in the GLM, such as `[MASK]` token in the BART model. Given both query q and document d as the context, the prefix text "It is known that" serves as a prompt (Liu et al., 2021) to instruct the BART model to output more informative text that is related to both q and d. The choice of prompt words can have an impact on the final attack effect, which is analyzed in Section 4.3.

Afterward, the BART model takes Eq. 1 as the input and fills a variable number of tokens (i.e., sentences with varying lengths) into the blank position. Herein, we do not incorporate extra ranking-incentivized objectives into the BART model in order to produce grammatically correct connection sentences well suited to the surrounding text. Besides, we employ the Top-k sampling (Fan et al., 2018) strategy to ensure the diversity of the generated sentences. As depicted in Algorithm 1, we sample multiple times (i.e., K), take M candidates each time and save at most N connection sentences, denoted as $\mathcal{S}_c = \{\overline{s}_1, \overline{s}_2, \cdots \overline{s}_N\}$. We also limit the maximum length of the connection sentence to L words to control the interference in the semantic content of the original document. All connection sentences in $\mathcal{S}_c$ are considered in the next stage for the merging with the original document d.

Algorithm 1 IDEM

Input: a query q, a target document d, a surrogate NRM $\mathcal{M}_S$ for a victim NRM $\mathcal{M}_V$
Parameter: the number of sentence in the target document $|d|$, the max number of connection sentence N, sample times K, sample size M, the max word length of connection sentence L
Output: an adversarial document d^{adv}

Stage 1: Connection Sentences Generation
Let $\mathcal{S}_c = \{\}$
for 1 to K **do**
  while $|\mathcal{S}_c| < N$ **do**
    Sample M connection sentence candidates $\{\overline{s}_u\}_{u=1}^{M}$ from the BART model by taking Eq. 1 as the input
    if the length of $\overline{s}_u$ is smaller than L **then**
      Keep $\overline{s}_u$ in $\mathcal{S}_c$
    end if
  end while
end for
Stage 2: Merging with Original Document
for $\overline{s}_u$ in $\mathcal{S}_c$ **do**
  for position index v from 0 to $|d|$ **do**
    Get an adversarial candidate $d^{adv}_{\langle u,v \rangle}$ as in Eq. 2, evaluate its *coherence* score (Eq. 3) and *relevance* score (Eq. 4), and compute the weighted merging score (Eq. 5)
  end for
end for
Find and save the top-1 ranked $d^{adv}_{\langle u,v \rangle}$ as d^{adv}
return d^{adv}

3.2 Merging with Original Document

As Transformer-based NRMs have a positional bias towards the start of document (Jiang et al., 2021; Hofstätter et al., 2021), adding adversarial text pieces at the beginning can usually achieve remarkable attack effect (Wang et al., 2022), but this also has the obvious drawback that the attack can be easily detected (Liu et al., 2022). Thus, to

balance between the attack effect and the fidelity of perturbed document content, we design a position-wise merging strategy to place an appropriate connection sentence at an optimal position within the original target document.

Specifically, given a query q, a target document with multiple sentences $d = s_1, s_2, \cdots, s_{|d|}$ and a set of connection sentences $\mathcal{S}_c$, we evaluate a large number of candidate adversarial documents with respect to the various combinations of d and $\mathcal{S}_c$. For each connection sentence $\overline{s}_u \in \mathcal{S}_c$, we merge $\overline{s}_u$ with the target document d by placing it at position index v (i.e., an integer from 0 to $|d|$), obtaining adversarial candidates in the form of Eq. 2.

$$d^{adv}_{\langle u,v\rangle} = \begin{cases} \overline{s}_u \oplus d & v = 0 \\ d_{1\to v} \oplus \overline{s}_u \oplus d_{(v+1)\to|d|} & 0 < v < |d| \\ d \oplus \overline{s}_u & v = |d| \end{cases} \tag{2}$$

wherein $d_{a\to b}$ means the text piece consists of consecutive sentences from s_a to s_b, and the subscript $\langle u, v\rangle$ means the connection sentence $\overline{s}_u$ is inserted at the index v of the target document d.

Subsequently, we examine all of the candidate adversarial documents $d^{adv}_{\langle u,v\rangle}$ and find out the best one as the final adversarial document d^{adv}. An effective adversarial example should be imperceptible to human judges yet misleading to NRMs (Wu et al., 2022). If the semantics of the added connection sentence differs greatly from the surrounding text in the target document, the resulting content will be incoherent and easily noticed by humans. Thus, with the help of the next sentence prediction (NSP) function in the pre-trained BERT model (Devlin et al., 2019), we define a coherence score on the junction between the connection sentence and the original target document.

Specifically, when the connection sentence $\overline{s}_u$ is placed in the middle of the target document d (i.e., $0 < v < |d|$), both the text pieces before and after the connection sentence are fed into the NSP function to get the *coherence* score as seen in Eq. 3.

$$\begin{aligned} coh(d^{adv}_{\langle u,v\rangle}) = 0.5 \times [&f_{nsp}(d_{1\to v}, \overline{s}_u \oplus d_{(v+1)\to|d|}) \\ &+ f_{nsp}(d_{1\to v} \oplus \overline{s}_u, d_{(v+1)\to|d|})] \end{aligned} \tag{3}$$

wherein $f_{nsp}(\mathrm{A}, \mathrm{B})$ is the NSP score of text A being followed by text B. Similarly, when $\overline{s}_u$ is attached to the beginning or the end of the document d (i.e., $v = 0$ or $v = |d|$), the coherence score is given as $f_{nsp}(\overline{s}_u, d)$ or $f_{nsp}(d, \overline{s}_u)$, respectively.

Apart from the semantic coherence, another important factor that impacts the feasibility of attack is the relevance between the query q and the candidate $d^{adv}_{\langle u,v\rangle}$. Herein, due to the black-box setting as described in Section 2, we need a surrogate NRM $\mathcal{M}_S$ to estimate the *relevance* score as in Eq. 4.

$$rel(q, d^{adv}_{\langle u,v\rangle}) = \mathcal{M}_S(q, d^{adv}_{\langle u,v\rangle}) \tag{4}$$

wherein the surrogate NRM $\mathcal{M}_S$ is not required to be heavily ranking-consistent with the victim NRM $\mathcal{M}_V$ but is only expected to distinguish which connection sentence contains spurious-relevant information (e.g., query keywords). In other words, the surrogate NRM does not participate in the generation of connection sentences, so that the adversarial information is from the BART model rather than the surrogate NRM. Thus, IDEM does not compulsively require a well-imitated surrogate NRM that is trained on a large number of in-domain training samples. As demonstrated later in Section 4.2, even when using an OOD surrogate NRM that is only slightly better than BM25, our proposed IDEM still achieves relatively high attack performance.

Finally, to trade-off the semantic coherence and relevance, as seen in Eq.5, we resort to a weighted sum of them as the merging score for each $d^{adv}_{\langle u,v\rangle}$.

$$\begin{aligned} score_m(q, d^{adv}_{\langle u,v\rangle}) = \alpha \times coh(d^{adv}_{\langle u,v\rangle}) \\ + (1-\alpha) \times rel(q, d^{adv}_{\langle u,v\rangle}) \end{aligned} \tag{5}$$

wherein α is a weight factor, and both scores are transformed to $[0, 1]$ by the min-max normalization before being added together. Out of all the candidates $d^{adv}_{\langle u,v\rangle}$, the one with the highest merging score is chosen as the final adversarial document d^{adv}.

4 Experiments

4.1 Experimental Setup

Dataset. Akin to previous works (Wu et al., 2022; Wang et al., 2022; Liu et al., 2022), we employ popular MS MARCO passage dataset (Nguyen et al., 2016) for our experiments, which contains about 8.8M passages (seen as documents in this paper) as the corpus. The evaluation data contains a Dev set with 6,980 queries and an Eval set with 6,837 queries. In addition, we use Natural Question (NQ) dataset (Kwiatkowski et al., 2019) that has been processed by Karpukhin et al. (2020) as the OOD training source for the surrogate NRM.

Victim NRM. Akin to Liu et al. (2022), the 'ms-marco-MiniLM-L-12-v2' model publicly available

at Sentence-Transformers (Reimers and Gurevych, 2019) is used as the representative victim NRM (i.e., $\mathcal{M}_V$) in our study. This victim NRM adopts MiniLM (Wang et al., 2020) as the backbone, and it is fine-tuned on MS MARCO in a cross-encoder architecture (Nogueira and Cho, 2019). Overall, $\mathcal{M}_V$ achieves highly effective ranking performance on the MS MARCO Dev set as seen in Table 2. The transfer attacks from $\mathcal{M}_V$ to other types of victim NRMs (e.g., MonoT5) are available in Section 4.2.

Surrogate NRMs. To examine the practicality of ranking attack methods in real-world situations, wherein access to the victim ranking system could be limited or unavailable, we experiment with three kinds of surrogate NRMs: $\boldsymbol{\mathcal{M}_{S_1}}$ is trained on all 6,837 Eval queries, $\boldsymbol{\mathcal{M}_{S_2}}$ is trained on the randomly sampled 200 Eval queries, and $\boldsymbol{\mathcal{M}_{S_3}}$ is trained on the OOD NQ queries. More details of surrogate NRMs can be found in Appendix A.1.

Target documents. Following Wu et al. (2022), we evaluate attack methods on randomly sampled 1K Dev queries with two types of target documents, i.e., Easy-5 and Hard-5, which are sampled from the re-ranked results by the victim NRM on the top-1K BM25 candidates. Determining whether a document is "Easy" or "Hard" depends on the difficulty of boosting it into the top-10 or top-50 positions. Specifically, **Hard-5** is the five bottom-ranked documents, and **Easy-5** is the five documents ranked between $[51, 100]$ in which one document is randomly picked out of every 10 documents.

Details of IDEM. By default, pre-trained BART-Base (Lewis et al., 2020) is instructed to generate connection sentences. The impacts of other GLMs, including pre-trained T5 (Raffel et al., 2020) and fine-tuned GPT-2 (Donahue et al., 2020), are also examined in Section 4.3. In the generation stage, k in sampling strategy is set as 50, and we sample 10 or 50 times (50 ones per time) and save at most 100 or 500 connection sentences with a max length of 12 words (i.e., $M=50$, $K=10/50$, $N=100/500$ and $L=12$). In the merging stage, we employ the NSP function in pre-trained BERT-Base to evaluate the coherence, and the α in Eq.5 is set as 0.5 (0.1) for Easy-5 (Hard-5) target documents. The impacts of L and N are available in Section 4.3, and the impact of α is analyzed in Appendix A.4.

Compared methods. We compare the following attack methods against NRMs: **Query+** (Liu et al., 2022) directly adds the query text at the beginning of the target document. **PRADA** (Wu et al., 2022) finds and replaces important words with synonyms in the target document. **Brittle-BERT** (Wang et al., 2022) and **PAT** (Liu et al., 2022) append a few tokens at the beginning of the target document, the former only considers the ranking-oriented objective, while the later further considers semantic and fluency constraints. For fair comparisons, PRADA perturbs at most 20 tokens, Brittle-BERT and PAT add at most 12 tokens. More details of these attack baselines are available in Appendix A.2.

Metrics. As for the re-ranking results, we report official MRR@10/1K metrics on the MS MARCO Dev set. Akin to Liu et al. (2022), we calculate the overlap of top-10 (Inter@10) and the Rank Biased Overlap (Webber et al., 2010) (RBO@1K, p is set as 0.7) to measure the ranking consistency between the surrogate and victim NRMs. As for the attack results, we report the percentage of successfully boosted target documents (i.e., attack successful rate, ASR) following Wu et al. (2022). Akin to Liu et al. (2022), we also report the average boosted ranks (*Boost*) of the target documents, and the percentage of the target documents that are promoted into top-10 ($\%\, r \leq 10$) and top-50 ($\%\, r \leq 50$). In addition, we measure the average perplexity (PPL) calculated using a pre-trained GPT-2 model (Radford et al., 2019), where a *lower* PPL value reflects better fluency of the adversarial documents as suggested by Kann et al. (2018) and Lei et al. (2022).

4.2 Results

IDEM demonstrates the ability to achieve comparable attack performance while not greatly diminishing the semantic fluency. The attack results are summarized in Table 1, where all values are averaged across the 5K target documents for 1K Dev queries. When the query is added to the beginning of the target document (referred to as Query+), a notable increase in ranking can be observed, which is as expected. However, this approach negatively impacts semantic fluency, resulting in a loss of approximately 8 (17) perplexity points on Easy-5 (Hard-5) target documents. Ideally, when the victim NRM is freely accessible, a sufficient amount of surrogate training data is obtainable to train a well-imitated surrogate model, like $\mathcal{M}_{S_1}$ in Table 2. As demonstrated in Table 1, when applied in conjunction with $\mathcal{M}_{S_1}$, the use of word-level synonym substitutions (i.e., PRADA) does not provide superior attack performance, but adding adversarial tokens at the beginning of the target documents

Surrogate NRM	Method	Easy-5 ASR	Easy-5 $\%r \leq 10$	Easy-5 $\%r \leq 50$	Easy-5 *Boost*	Easy-5 PPL↓	Hard-5 ASR	Hard-5 $\%r \leq 10$	Hard-5 $\%r \leq 50$	Hard-5 *Boost*	Hard-5 PPL↓
-	Original	-	-	-	-	37.3	-	-	-	-	50.5
-	Query+	100.0	86.9	99.2	70.3	45.4	100.0	47.8	78.3	955.1	67.5
$\mathcal{M}_{S_1}$	PRADA	77.9	3.52	46.2	23.2	94.4	68.0	0.02	0.10	65.2	154.4
	Brittle-BERT	98.7	81.3	96.7	67.3	107.9	**100.0**	**61.5**	**85.9**	**965.5**	152.5
	PAT	89.6	30.6	73.8	41.9	50.9	98.0	6.24	20.1	589.1	71.4
	IDEM$_{N=100}$	99.3	77.9	97.0	67.0	**36.2**	99.5	41.4	66.9	875.6	**54.6**
	IDEM$_{N=500}$	**99.7**	**87.4**	**99.0**	**70.3**	36.4	99.8	54.3	79.3	933.0	54.9
$\mathcal{M}_{S_2}$	PRADA	69.6	1.36	35.0	17.4	90.7	66.0	0.00	0.10	49.3	152.1
	Brittle-BERT	81.6	33.2	69.7	36.4	131.3	94.8	8.98	25.4	565.1	179.5
	PAT	61.9	8.46	37.3	12.5	49.3	84.4	0.82	3.40	221.7	66.2
	IDEM$_{N=100}$	96.8	65.3	91.8	60.7	**36.5**	97.6	31.7	57.0	822.0	**54.7**
	IDEM$_{N=500}$	**98.7**	**74.8**	**95.4**	**65.1**	37.0	**99.1**	**39.6**	**67.4**	**890.7**	55.4
$\mathcal{M}_{S_3}$	PRADA	71.5	1.86	37.5	19.1	91.5	71.9	0.00	0.08	73.4	168.7
	Brittle-BERT	90.0	43.4	80.1	46.2	117.7	**99.9**	17.7	47.6	845.2	156.8
	PAT	51.1	2.70	22.9	2.01	46.8	79.0	0.00	0.66	92.9	64.2
	IDEM$_{N=100}$	97.2	57.3	89.8	58.1	**36.9**	99.2	23.0	48.8	805.7	**55.2**
	IDEM$_{N=500}$	**98.8**	**65.3**	**93.8**	**61.9**	37.7	99.8	**29.1**	**57.9**	**866.2**	56.0

Table 1: Attack results on two types of target documents under three kinds of surrogate NRMs. Lower values of perplexity (PPL) indicate better performance, while higher values are desirable on other metrics.

Model	MRR@10	MRR@1K	Inter@10	RBO@1K
BM25	18.4	19.5	22.8	31.5
$\mathcal{M}_V$	39.5	40.3	-	-
$\mathcal{M}_{S_1}$	37.0	37.8	73.1	66.2
$\mathcal{M}_{S_2}$	23.0	24.2	41.0	31.2
$\mathcal{M}_{S_3}$	21.0	22.2	27.3	37.6

Table 2: Ranking similarities between the victim and surrogate NRMs when re-ranking BM25 top-1K candidates on the MS MARCO Dev set.

(i.e., Brittle-BERT and PAT) yields better attack performance. However, adversarial documents generated by these prior methods, particularly PRADA and Brittle-BERT, exhibit excessively high perplexity (PPL) values. In contrast, our proposed IDEM not only achieves comparable attack performance, exemplified by its superior performance on Easy-5 target documents, but also mitigates the detrimental effects on text fluency, as evidenced by the lower PPL values on its adversarial documents.

IDEM is more robust and much less affected by the surrogate NRM, even when it is an OOD NRM. When access to the victim NRM (i.e., $\mathcal{M}_V$) is restricted, as shown in Table 2, there is a notable discrepancy between the surrogate $\mathcal{M}_{S_2}$ and the victim $\mathcal{M}_V$, as only a limited number of in-domain training samples can be used. When working with $\mathcal{M}_{S_2}$, the attack performances of all methods, particularly PAT and Brittle-BERT, are greatly diminished as can be observed in Table 1. For example, when $\mathcal{M}_{S_1}$ is changed to $\mathcal{M}_{S_2}$, the attack performance ($\%r \leq 10$) of Brittle-BERT drops dramatically from 81.3 to 33.2 on Easy-5 target documents, and from 61.5 to 8.98 on Hard-5 target documents. Additionally, the semantic fluency of adversarial documents produced by Brittle-BERT also decreases, as evidenced by an increase in PPL from 107.9 to 131.3 on Easy-5 target documents. Furthermore, when access to the victim NRM is not available, only publicly available OOD training samples can be used, yield a more discrepant surrogate NRM, like $\mathcal{M}_{S_3}$ in Table 2. In this situation, as shown in Table 1, the performances of prior attack methods are also not ideal, especially PRADA and PAT. However, when working with both $\mathcal{M}_{S_2}$ and $\mathcal{M}_{S_3}$, IDEM demonstrates the best attack results among all baselines while preserving the semantic fluency of target documents. For more details, including the attack efficiency and adversarial examples for different attack methods, please refer to Appendixes A.3 and A.5, respectively.

IDEM generates more general adversarial documents that can be well transferred across different victim NRMs. To examine the transferability of adversarial documents, considering that the target victim NRM is subject to continuous improvement, updates, or even changes, we conduct experiments with three different NRMs as the attack targets: ELECTRA ('ms-marco-electra-base'; Reimers and Gurevych, 2019), MonoT5 ('monot5-base-msmarco'; Nogueira et al., 2020)

Victim NRM	Method	ASR	%$r \leq 10$	*Boost*
ELECTRA	PRADA	44.6	0.94	14.6
	Brittle-BERT	80.0	23.6	216.4
	PAT	57.9	4.7	49.5
	IDEM$_{N=100}$	94.7	47.4	336.6
	IDEM$_{N=500}$	**97.5**	**55.9**	**371.8**
MonoT5	PRADA	51.6	0.83	9.28
	Brittle-BERT	85.4	17.2	253.0
	PAT	67.8	3.71	113.6
	IDEM$_{N=100}$	96.2	43.5	400.8
	IDEM$_{N=500}$	**98.2**	**51.7**	**434.9**
ColBERT	PRADA	45.8	0.68	22.1
	Brittle-BERT	87.9	17.2	292.8
	PAT	72.2	3.87	114.1
	IDEM$_{N=100}$	96.2	46.3	418.7
	IDEM$_{N=500}$	**98.0**	**54.7**	**453.8**

Table 3: Cross transfer attack results from source victim NRM $\mathcal{M}_V$ (MiniLM) and its surrogate NRM $\mathcal{M}_{S_2}$ (BERT) to other target victim NRMs.

and ColBERT (Khattab and Zaharia, 2020). The adversarial documents were generated by attacking the $\mathcal{M}_V$ ('ms-marco-MiniLM-L-12-v2') NRM and subsequently used to attack other victim NRMs. Due to space constraints, Table 3 primarily focuses on presenting the cross-attack results of all adversarial target documents generated using the surrogate NRM $\mathcal{M}_{S_2}$ in both Easy-5 and Hard-5 sets. As observed in Table 3, regardless of the target victim NRM, the attack performances of different methods exhibit a consistent trend: IDEM > Brittle-BERT > PAT > PRADA. This trend is also in line with the results presented in Table 1, indicating that adversarial documents that display higher aggression towards one NRM are more likely to be effective against other NRMs due to shared characteristics among the models, such as exact matching.

4.3 Analysis

Impact of prompt in IDEM. During the generation of connection sentence, a prompt (in Eq.1) is utilized to guide the BART model to output more informative text. Herein, we examine the impact of prompt on the final attack results of IDEM. As seen in Table 4, four different prompts are analyzed, in which "It is known that" produces better attack results, particularly on Hard-5 target documents, and also generates lower PPL values. In contrast, "The fact is that" and "We know that" perform slightly worse on Hard-5 target documents, and "It is about that" performs even worse. Overall, the choice of prompt has marginal impact on the ASR but mainly affects the degree of promotion in the attack.

Prompt	ASR	%$r \leq 10$	*Boost*	PPL↓
Easy-5				
It is known that	99.3	77.9	67.0	36.2
It is about that	98.3	60.9	60.3	39.1
We know that	99.3	74.4	65.8	36.8
The fact is that	99.6	77.2	66.9	36.8
Hard-5				
It is known that	99.5	41.4	875.6	54.6
It is about that	99.5	20.0	798.9	57.3
We know that	99.4	33.9	851.1	54.9
The fact is that	99.5	37.2	870.0	54.9

Table 4: Attack results of IDEM$_{N=100}$ under surrogate NRM $\mathcal{M}_{S_1}$ with different prompts for instructing BART to generate connection sentences.

NRM	Index				
	%$v=0$	%$v=1$	%$v=2$	%$v=3$	%$v \geq 4$
Easy-5					
$\mathcal{M}_{S_1}$	83.9	12.2	2.62	0.88	0.48
$\mathcal{M}_{S_2}$	74.1	15.0	4.77	1.56	4.59
$\mathcal{M}_{S_3}$	22.3	29.5	22.3	11.7	14.2
Hard-5					
$\mathcal{M}_{S_1}$	83.2	13.5	2.12	0.76	0.40
$\mathcal{M}_{S_2}$	76.3	14.6	4.32	1.12	3.68
$\mathcal{M}_{S_3}$	9.76	37.7	24.8	12.9	14.9

Table 5: Statistics on the positioning of connection sentences in adversarial documents created by IDEM$_{N=100}$.

Position of connection sentence in IDEM. As mentioned in Section 3.2, IDEM can automatically place an adversarial connection sentence within the target documents. As summarized in Table 5, we count the positions (i.e., index v) of connection sentences. When the surrogate NRM is $\mathcal{M}_{S_1}$ or $\mathcal{M}_{S_2}$, it is observed that less than 30% of the connection sentences are positioned in the middle (i.e., $v \geq 1$). Meanwhile, when the surrogate NRM is $\mathcal{M}_{S_3}$, the distribution of insert positions is found to be more uniform, with only 10-20% of the connection sentences being appended to the beginning (i.e., $v = 0$). These findings indicate that IDEM can place the adversarial text at any positions within the target document, making it harder to be detected.

IDEM based on other GLMs. In addition to evaluating IDEM using BART, we also examine how well IDEM performs when other GLMs with blank-filling capabilities are instructed to generate connection sentences, including T5-Base (Raffel et al., 2020) and ILM (a fine-tuned version of GPT-2; Donahue et al., 2020). As presented in Table 6,

GLM	ASR	$\%\, r \leq 10$	$\%\, r \leq 50$	*Boost*	PPL↓
Easy-5					
BART	99.3	77.9	97.0	67.0	36.2
T5	98.7	75.3	95.4	65.2	38.0
ILM	93.9	45.7	81.5	50.6	41.5
Hard-5					
BART	99.5	41.4	66.9	875.6	54.6
T5	97.8	37.1	60.1	820.9	56.8
ILM	98.3	16.0	35.6	693.2	59.8

Table 6: Attack results of $\text{IDEM}_{N=100}$ under $\mathcal{M}_{S_1}$ with other generative language models (GLMs).

Method	#Correctness↓	#Suggestions↓	Quality
Original	1.80	4.23	59
Query+	2.39	6.50	49
PRADA	5.51	11.0	22
Brittle-BERT	3.74	7.12	41
PAT	2.40	6.21	52
$\text{IDEM}_{N=100}$	2.29	5.06	57
$\text{IDEM}_{N=500}$	2.31	5.18	57

Table 7: Automatic evaluation results provided by the online grammar checker (i.e., Grammarly).

our empirical results show that while the BART-based IDEM is found to be the overall best, all three models achieve high attack success rates, indicating the general applicability of the IDEM approach.

Online grammar checking. In order to examine the extra errors introduced by different attack methods, we employ the popular online grammar checker *Grammarly*[2] to evaluate the quality of adversarial documents. Specifically, we collect 100 adversarial documents (with the same id) produced by each attack method for 50 Dev queries. We report three evaluation metrics given by Grammarly, including the average number of issues in correctness (e.g., spelling, grammar, and punctuation) and suggestions (e.g., wordy or unclear sentences, etc.) in each adversarial document, and the quality score of all adversarial documents. As shown in Table 7, IDEM introduces the fewest issues and obtains the highest quality score among all attack methods, indicating that the adversarial documents produced by IDEM are more machine-imperceptible.

Human evaluation. To further prove that the adversarial documents generated by IDEM are more natural to readers, a human-subject evaluation was conducted to assess the imperceptibility of adversarial text. Specifically, 32 adversarial documents were randomly selected for each attack method,

[2] https://app.grammarly.com

Method	Human-Imperceptibility	
	Avg.	Kappa
Original	0.69	0.44
Query+	0.36	0.55
PRADA	0.45	0.56
Brittle-BERT	0.50	0.43
PAT	0.55	0.69
$\text{IDEM}_{N=100}$	0.78	0.11
$\text{IDEM}_{N=500}$	0.64	0.54

Table 8: Human evaluation results on the imperceptibility of adversarial documents. The closer the average score is to 1, the less likely it is to be noticed by humans.

Method	Linguistic Acceptability	Accuracy
Original	0.76	0.87
Query+	0.50	0.52
PRADA	0.47	0.58
Brittle-BERT	0.19	0.97
PAT	0.42	0.70
$\text{IDEM}_{N=100}$	0.66	0.27
$\text{IDEM}_{N=500}$	0.65	0.28

Table 9: Mitigation by the linguistic acceptability (LA). When the LA score of a document is below 0.5, it is classified as adversarial and subsequently filtered out.

and 32 original unaltered documents were also selected. All these documents were then mixed and randomly divided into two groups, with each group being evaluated by two annotators who are computer science graduate students with the necessary knowledge to understand the nature of the ranking attack. The annotators were tasked with determining whether the document content had been attacked (0) or was normal (1). We averaged all annotations on 32 samples from 2 annotators as the final imperceptibility score, and also computed Kappa coefficient for the annotation consistency. As can be seen in Table 8, IDEM receives the highest score for human imperceptibility among all attack methods. Additionally, the Kappa values of almost all attack methods are larger than 0.4 (considered as "moderate agreement"), while $\text{IDEM}_{N=100}$ has the smallest Kappa value (i.e., 0.11), which seems reasonable since it is hard to reach an agreement on the more imperceptibly attacked documents.

Mitigation by linguistic acceptability. In an effort to mitigate the attacks, we make additional attempts using a classification model[3] trained on the CoLA dataset (Warstadt et al., 2019). The same adversarial documents subjected to online grammar

[3] https://huggingface.co/textattack/roberta-base-CoLA

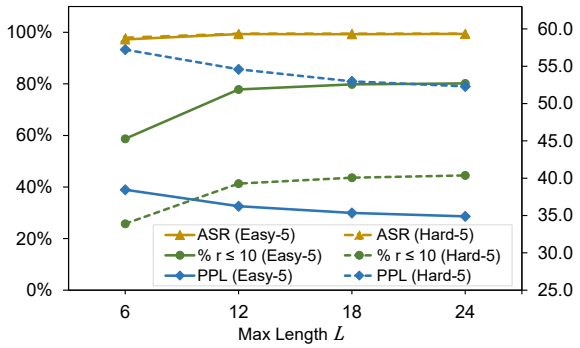

Figure 2: Impact of the length of connection sentences.

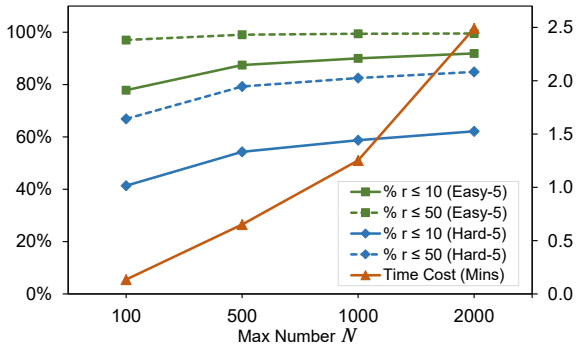

Figure 3: Impact of the number of connection sentences.

checking are evaluated for linguistic acceptability (LA) using this model. In addition to the LA values, the classification accuracy is reported as an indicator of the mitigation effectiveness. As indicated in Table 9, IDEM's adversarial documents exhibit linguistic acceptability closest to that of the original documents, while Brittle-BERT's adversarial documents are deemed the most unacceptable. Moreover, this LA model can successfully identify over 50% of the adversarial documents from Query+, PRADA and PAT, and even 97% of the adversarial documents from Brittle-BERT. In contrast, only 27-28% of IDEM's adversarial documents are filtered out, implying that more than 70% of IDEM's attacks remain effective against this mitigation. Also, the misclassification rate of this LA model on original documents is only 13%.

The impact of hyper-parameters. We evaluate IDEM with two hyper-parameters to study how they affect the attack performance: the max length (L) and max number (N) of connection sentences. For L analysis, we maintain N at 100, while for N analysis, we keep L at 12. In Figure 2, the improvement in IDEM's attack performance becomes less significant as L increases from 12 to 24 compared to that observed from 6 to 12, leading us to select L as 12. Similarly, in Figure 3, IDEM's performance improves as N increases, but the time required to generate one adversarial document increases at a faster rate. Hence, we set N to 500 for a balance between attack effectiveness and efficiency.

5 Related Work

The adversarial IR has been studied for an extended period, such as black-hat SEO, which refers to the intentional manipulation of Web pages with the goal of obtaining an unjustified ranking position, resulting in a decline in the quality of search results and an inundation of irrelevant pages (Gyöngyi and Garcia-Molina, 2005). In this context, mainstream research focuses on studying the adversarial manipulations through various aspects, such as detection (Dalvi et al., 2004; Ntoulas et al., 2006), theoretical and empirical analysis (Raifer et al., 2017), robustness of LTR-based ranking functions (Goren et al., 2018), automatic content modification (Goren et al., 2020), and other research directions (Kurland and Tennenholtz, 2022).

Recently, there has been significant progress in NRMs, particularly those leveraging PLMs, which have shown exceptional performance in text ranking (Lin et al., 2021). Concurrently, an increasing number of studies have shed light on the robustness concerns of NRMs in various scenarios, including the presence of query typos or variations (Zhuang and Zuccon, 2021; Penha et al., 2022; Chen et al., 2022), textual noises (Chen et al., 2023), and adversarial attacks (Raval and Verma, 2020; Song et al., 2022). Although current attack methods like PRADA (Wu et al., 2022), Brittle-BERT (Wang et al., 2022), and PAT (Liu et al., 2022) have shown the ability to deceive NRMs successfully, they introduce additional quality issues and often heavily rely on surrogate NRMs for document manipulation. Instead, our proposed IDEM effectively overcomes the limitations of these existing attack methods and showcases remarkable attack performance.

6 Conclusion

In this study, we introduce a document manipulation framework named IDEM, which is engineered to produce adversarial documents that are not easily detected by both humans and algorithms. Our experiments on the MS MARCO dataset show that IDEM can not only achieve a high level of attack performance, but also generate correct and fluent adversarial documents as evaluated by both automatic and human assessments.

Limitations

In our experiments, as NRMs with cross-encoder are widely used, we focus on evaluating the textual adversarial robustness during the re-ranking stage and do not currently take into account the effect on the retrieval stage. But actually, in a "first retrieval then re-ranking" ranking paradigm, the attack is effective only when the adversarial documents are passed into the top retrieval results. Meanwhile, dense retrieval (DR) models have been widely studied, and they may also inherit adversarial vulnerabilities due to the basics of PLMs. Besides, due to limitations in our computing resources, we only tested adding adversarial text to relatively short documents (i.e., passage-level), but the document content in real-world applications could be much longer. Therefore, further comprehensive investigations on examining the NRMs with different architectures, the effects of attacks on the retrieval models, and the manipulations on longer documents are left for future work. Finally, it is important to note that mitigation and defense methods against adversarial ranking attacks are currently understudied, making it a significant area for future research.

Ethics Statement

In this paper, we investigate the potential vulnerability concerns of the neural information retrieval (IR) systems and propose a document manipulation framework that generates adversarial documents that are not easily detected by both humans and IR systems. We hope that this study could inspire further exploration and design of adversarial ranking defense/detection methods and aid in the development of robust real-world search engines.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants no. U1936207 and 62272439.

References

Carlos Castillo and Brian D. Davison. 2010. Adversarial web search. *Found. Trends Inf. Retr.*, 4(5):377–486.

Xuanang Chen, Ben He, Kai Hui, Le Sun, and Yingfei Sun. 2023. Dealing with textual noise for robust and effective BERT re-ranking. *Inf. Process. Manag.*, 60(1):103135.

Xuanang Chen, Jian Luo, Ben He, Le Sun, and Yingfei Sun. 2022. Towards robust dense retrieval via local ranking alignment. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 1980–1986. ijcai.org.

Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. 2004. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 99–108. ACM.

Brian D. Davison, Marc Najork, and Tim Converse. 2006. Adversarial information retrieval on the web (airweb 2006). *SIGIR Forum*, 40(2):27–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Saad Farooq. 2019. A survey on adversarial information retrieval on the web. *CoRR*, abs/1911.11060.

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, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 395–404. ACM.

Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2020. Ranking-incentivized quality preserving content modification. In *Proceedings of the 43rd International ACM SIGIR conference on*

research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pages 259–268. ACM.

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*, pages 39–47.

Sebastian Hofstätter, Aldo Lipani, Sophia Althammer, Markus Zlabinger, and Allan Hanbury. 2021. Mitigating the position bias of transformer models in passage re-ranking. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 238–253. Springer.

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*, pages 496–509, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Oren Kurland and Moshe Tennenholtz. 2022. Competitive search. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2838–2849. ACM.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Yibin Lei, Yu Cao, Dianqi Li, Tianyi Zhou, Meng Fang, and Mykola Pechenizkiy. 2022. Phrase-level textual adversarial attack with label preservation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1095–1112, Seattle, United States. Association for Computational Linguistics.

Alex Goh Kwang Leng, Ravi Kumar Patchmuthu, Ashutosh Kumar Singh, and Anand Mohan. 2012. Link-based spam algorithms in adversarial information retrieval. *Cybern. Syst.*, 43(6):459–475.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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*, pages 2025–2039. ACM.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics:*

EMNLP 2020, pages 708–718, Online. Association for Computational Linguistics.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Alexandros Ntoulas, Marc Najork, Mark S. Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, pages 83–92. ACM.

Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, pages 397–412. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nimrod Raifer, Fiana Raiber, Moshe Tennenholtz, and Oren Kurland. 2017. Information retrieval meets game theory: The ranking competition between documents? authors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 465–474. ACM.

Nisarg Raval and Manisha Verma. 2020. One word at a time: adversarial attacks on retrieval models. *CoRR*, abs/2008.02197.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Mike Gatford, and A. Payne. 1995. Okapi at TREC-4. In *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995*, volume 500-236 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Congzheng Song, Alexander Rush, and Vitaly Shmatikov. 2020. Adversarial semantic collisions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4198–4210, Online. Association for Computational Linguistics.

Junshuai Song, Jiangshan Zhang, Jifeng Zhu, Mengyun Tang, and Yong Yang. 2022. TRAttack: Text rewriting attack against text retrieval. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 191–203, Dublin, Ireland. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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*, pages 115–120. ACM.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Trans. Assoc. Comput. Linguistics*, 7:625–641.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2022. PRADA: practical black-box adversarial attacks against neural ranking models. *CoRR*, abs/2204.01321.

Jincheng Xu and Qingfeng Du. 2020. Texttricker: Loss-based and gradient-based adversarial attacks on text classification models. *Eng. Appl. Artif. Intell.*, 92:103641.

Shengyao Zhuang and Guido Zuccon. 2021. Dealing with typos for BERT-based passage retrieval and ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2836–2842, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 4014–4022. ACM.

A Appendix

A.1 Details of Surrogate NRMs

As mentioned in Section 4.1, we conduct black-box attacks against the victim neural ranking model (NRM) $\mathcal{M}_V$ using three types of surrogate NRMs. Since the training data of the victim NRM is typically unavailable, we use Eval queries in the MS MARCO dataset rather than train queries to construct the surrogate training data $\mathcal{T}_S$. As we move from $\mathcal{M}_{S_1}$ to $\mathcal{M}_{S_2}$, the number of in-domain Eval queries used for ranking imitation decreases from 6,837 to 200, it means the frequency of querying the victim NRM is greatly reduced. For each Eval query q_i, we collect all 406 document pairs $(d_{i,j}, d_{i,k})$ $(1 \leq j < k \leq 29)$ from the top-29 of the re-ranking list produced by the victim NRM over the top-1K BM25 candidates, and use this to construct the surrogate training data $\mathcal{T}_S$ as described in Section 2. As a result, the in-domain surrogate training data for $\mathcal{M}_{S_1}$ and $\mathcal{M}_{S_2}$ contains 81.2 thousand and 2.77 million training triples, respectively. Additionally, $\mathcal{M}_{S_3}$ implies the scenario where no in-domain data from the victim NRM is available, and we collect 3.72 million training triples from the NQ dataset as the out-of-domain (OOD) surrogate training data. The surrogate NRMs are based on the pre-trained BERT-Base model and fine-tuned for two epochs with a learning rate of 3e-6 and batch size of 16 for $\mathcal{M}_{S_1}$ and $\mathcal{M}_{S_2}$, and one epoch for $\mathcal{M}_{S_3}$ with the same learning rate and batch size. It is important to note that the goal of this work is not to develop a new training method for the surrogate NRMs, and thus we directly adopt the hinge loss with a margin of 1 for ranking imitation as per previous work (Wu et al., 2022).

A.2 Details of Attack Baselines

In our adversarial attack experiments, we examine the following baseline methods:

Query+ (Liu et al., 2022) is an intuitive baseline that directly appends the query text to the beginning of the target document. Although the query text can be placed at any position in the target document, or even determined by our proposed position-aware mechanism, appending to the beginning usually produces greater attack results due to the positional bias in Transformer-based NRMs (Jiang et al., 2021; Hofstätter et al., 2021). Thus, Query+ acts as a baseline method that does not take into account the invisibility aspect of the attack.

PRADA (Wu et al., 2022) is a Word Substitution Ranking Attack (WSRA) method, it first finds important words (i.e., sub-word tokens) in the target document according to the gradient magnitude (Xu and Du, 2020), and then greedily replaces them with the synonyms found in a perturbed word embedding space via PGD (Madry et al., 2018). Based on our observations, when attacking different random target samples, PRADA is able to attain attack performance (ASR) that is close to the results reported in its original publication. This is particularly true when the victim NRM has relatively poor ranking performance, due to the obvious fact that it is much easier to attack a weaker victim NRM. Despite this variability, the overall conclusions drawn from our experimentation remain unchanged.

Brittle-BERT (Wang et al., 2022) studies both local (i.e., a particular query-document instance) and global (i.e., an entire workload of queries) ranking attack to cause a large rank demotion or promotion by adding/replacing a small number of tokens. In our work, we only adopt the local setting and add tokens to the beginning of the target document as it usually produces better attack results. Specifically, Brittle-BERT first initializes a few placeholder tokens at the beginning of the target document and then employs HotFlip (Ebrahimi et al., 2018) algorithm to update them as being more adversarial.

PAT (Liu et al., 2022) generates and adds several trigger tokens at the beginning of the target document. In addition to the ranking-incentivized objective, the search objective of PAT is equipped with semantic and fluency constraints using the pre-trained BERT model. The surrogate NRMs trained with hinge loss have a one-class prediction layer, but PAT needs a surrogate NRM with a two-class prediction layer, namely, 'Pairwise BERT' as denoted in PAT (Liu et al., 2022), so we use the same surrogate training data to obtain 'Pairwise BERT' using the default imitation loss in PAT.

In order to evaluate these baselines, we utilize their publicly available implementations and ensure that all settings are consistent with those described in their respective official publications.

A.3 Time Cost of Attack

In previous PRADA, Brittle-BERT and PAT, the replacement, selection, and search of tokens are carried out one by one using the surrogate NRM to produce an adversarial document, so it needs a large amount of time to complete the attack process. However, in our proposed IDEM framework,

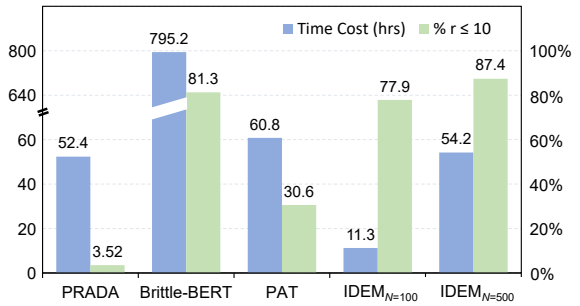

Figure 4: Total time cost of generating 5K adversarial documents for different attack methods, along with the attack performance ($\%\, r \leq 10$) on Easy-5 target documents under surrogate $\mathcal{M}_{S_1}$ (in Table 1).

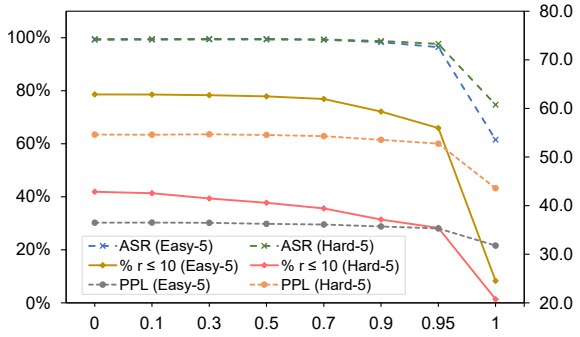

Figure 5: Impact of α in the merging score (Eq. 5) on the attack results of IDEM$_{N=100}$.

the adversarial text is first generated via a GLM (e.g., BART), and then combined in the sentence level, which is more efficient than the token level. As seen in Figure 4, we summarize the total time cost of different attack methods in producing 5,000 adversarial documents using one Titan RTX GPU. When at most 100 candidate connection sentences are generated for the position-wise combination, IDEM$_{N=100}$ only takes 11.3 hours to achieve great attack performance. By comparison, PRADA and PAT consume more time but still perform worse in attack, Brittle-BERT takes a huge time cost (near 800 hours) even though its attack performance is considerable. Additionally, compared with Brittle-BERT, IDEM$_{N=500}$ can produce comparable attack performance but take much less time cost. Therefore, our proposed IDEM method lays equal stress on attack efficiency and performance.

A.4 The Impact of α in the Merging Score

In our IDEM framework, after generating a series of connection sentences between the query and the target document, a position-aware merging mechanism is employed to decide the final adversarial document, wherein a coherence score and a relevance score are added together using α as seen in Eq. 5. As shown in Figure 5, we can observe that α in a wide range (from 0 to 0.95) does not affect the attack success rate (ASR) too much on both Easy-5 and Hard-5 target document sets, and the $\%\, r \leq 10$ metric starts to decrease at $\alpha = 0.5$ and $\alpha = 0.1$ on Easy-5 and Hard-5 target document sets, respectively. As for the perplexity (PPL) metric (smaller is better), when α increases (more attention on the coherence), PPL value does not change a lot until α reaches about 0.9 to 1. Meanwhile, it can produce adversarial documents with lower PPL than original ones, e.g., when α is 1, the average PPL points on Easy-5 and Hard-5 target document sets are only 31.8 and 43.6, respectively.

A.5 Adversarial Examples

To better understand the workings of various attack methods, we show adversarial examples produced by them under three types of surrogate NRMs in Table 10. We can observe that Query+, and Brittle-BERT, PAT, IDEM under $\mathcal{M}_{S_1}$ promote the target document ranked at 88th into top-10. However, adding query text (i.e., Query+) and unnatural token sequence (i.e., Brittle-BERT and PAT) at the beginning make adversarial documents distinguishable, while the inserted adversarial text by IDEM is more semantically consistent with the original surrounding content. When the surrogate NRM degrades from $\mathcal{M}_{S_1}$ to $\mathcal{M}_{S_2}$ as not enough in-domain training samples are available, we can see that the ranking of the adversarial document by PRADA decreases from 27th to 65th, and the ranking of the adversarial document by Brittle-BERT also decreases from 2nd to 14th. Furthermore, when an OOD surrogate NRM (i.e., $\mathcal{M}_{S_3}$) is used due to the forbidden access to the victim NRM, we can find out that the attack effects of PRADA, Brittle-BERT and PAT are greatly suppressed. For example, although PAT under $\mathcal{M}_{S_2}$ promotes the target document to the ranking of 2nd, PAT under $\mathcal{M}_{S_3}$ even demotes the ranking of the target document from 88th to 107th. In contrast, under both $\mathcal{M}_{S_2}$ and $\mathcal{M}_{S_3}$, IDEM robustly promotes the target document into top-10 (i.e., 9th), and the fluency and correctness of adversarial documents are still within an acceptable range. From these adversarial cases, it is evident that IDEM is less dependent on the surrogate NRM and can perform attacks more robustly than previous attack methods, indicating a flexible use condition of IDEM in real-world situations.

Method	Original or Adversarial Text	Rank↓	PPL↓
Original	Carry your baby in a sling or front carrier. Feeling your baby's warmth, smelling his sweet scent, and looking down often to make eye contact with him can help you bond. Spend plenty of close-up face time with your baby. Smile at him, and return the smile when he smiles first.	88	33.9
Query+	what age can you wear baby on back in a carrier? Carry your baby in a sling or front carrier. Feeling your baby's warmth, smelling his sweet scent, and looking ...	2	40.4
Surrogate NRM $\mathcal{M}_{S_1}$			
PRADA	wear your baby in a sling ord front carrier. Feeling your baby's warm, smelling his sweet scent, and looking down often to make eye contact with him can help you bondage. Spend plenty of close-up face time with your baby. Smile at him, and ...	27	100.8
Brittle-BERT	pendingerabidheartedivating aged aged 292,oning worn wear Carry your baby in a sling or front carrier. Feeling your baby's warmth, smelling his sweet scent, and ...	2	125.0
PAT	about 30 year old babies can carry baby carriers back to age Carry your baby in a sling or front carrier. Feeling your baby's warmth, smelling his sweet scent, and ...	2	52.0
IDEM$_{N=100}$	Carry your baby in a sling or front carrier. A child of any age can wear shoes with a sling. Feeling your baby's warmth, smelling his sweet scent, and looking ...	9	33.7
IDEM$_{N=500}$	Carry your baby in a sling or front carrier. Most parents wear infant carriers around age 3, 4, and 5. Feeling your baby's warmth, smelling his sweet scent, and ...	1	36.5
Surrogate NRM $\mathcal{M}_{S_2}$			
PRADA	Carry your baby in a slingshot ord front carrier. Feeling your baby's warm, smelling his sweet scent, and looking down often to make eye contact with him can ...	65	82.2
Brittle-BERT	modernism age hms× chestnut beyonce rappers commercially whilst wearing md respectively Carry your baby in a sling or front carrier. Feeling your baby's ...	14	175.2
PAT	be a year old and can wear around Carry your baby in a sling or front carrier. Feeling your baby's warmth, smelling his sweet scent, and looking down often to ...	2	48.3
IDEM$_{N=100}$	Carry your baby in a sling or front carrier. Wearing your baby on back is always a good idea. Feeling your baby's warmth, smelling his sweet scent, and looking ...	9	29.4
IDEM$_{N=500}$	Carry your baby in a sling or front carrier. You can and should be wearing baby on back in a carrier. Feeling your baby's warmth, smelling his sweet scent, and ...	9	36.9
Surrogate NRM $\mathcal{M}_{S_3}$			
PRADA	wear your baby in a slingshot ord front carrier. Feeling your baby's warmth, smelling his sweet scent, and looking down often to make eye contact with him can ...	28	67.2
Brittle-BERT	offspring coherent examples declined toys widespread adulthood noun whether buckled off wear Carry your baby in a sling or front carrier. Feeling your baby's ...	58	143.8
PAT	for example, may carry twenty cents Carry your baby in a sling or front carrier. Feeling your baby's warmth, smelling his sweet scent, and looking down often to ...	107	49.4
IDEM$_{N=100}$	A child of any age can wear shoes with a sling. Carry your baby in a sling or front carrier. Feeling your baby's warmth, smelling his sweet scent, and looking down ...	9	39.1
IDEM$_{N=500}$	Carry your baby in a sling or front carrier. You can and should be wearing baby on back in a carrier. Feeling your baby's warmth, smelling his sweet scent, and ...	9	36.9

Table 10: Adversarial documents generated by various attack methods under three kinds of surrogate NRMs on the same related but irreverent document for the query "what age can you wear baby on back in a carrier?" from the MS MARCO Dev set. The inserted and perturbed words are marked as Red for easy comparisons.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section: Limitations

☑ A2. Did you discuss any potential risks of your work?
Section: Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1: Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.1: Experimental Setup

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1: Experimental Setup

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The MS MARCO dataset utilized in our research is freely accessible to the public, without any accompanying licenses or intellectual property restrictions.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We employ the MS MARCO dataset exclusively for non-commercial research endeavors, adhering to their intended purpose. Similarly, the data generated within our work is strictly intended for research purposes, aiming to foster progress in the field of information retrieval and related domains.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
In light of the widespread usage of the MS MARCO dataset in information retrieval tasks, we did not specifically examine the textual data contained within it for potential privacy or offensive information.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The MS MARCO dataset utilized in our research is openly accessible, and comprehensive information about it can be found on its official page.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1: Experimental Setup; Appendix A.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 4: Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.3

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1: Experimental Setup

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
The outcomes presented in our work are based on a single execution due to constraints in our computing resources.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4: Experiments; Appendix

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4.3: Analysis

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Our human annotation process does not involve such a question.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4.3: Analysis

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 4.3: Analysis

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Our human annotation process does not involve such a question.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Our human annotation process does not involve such a question.