Perturbation-Invariant Adversarial Training for Neural Ranking Models: Improving the Effectiveness-Robustness Trade-Off

Yu-An Liu[1,2], Ruqing Zhang[1,2], Mingkun Zhang[1,2], Wei Chen[1,2], Maarten de Rijke[3], Jiafeng Guo[1,2], Xueqi Cheng[1,2]

[1]CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]University of Amsterdam, Amsterdam, The Netherlands
{liuyuan21b, zhangruqing, zhangmingkun20z, chenwei2022, guojiafeng, cxq}@ict.ac.cn, m.derijke@uva.nl

Abstract

Neural ranking models (NRMs) have shown great success in information retrieval (IR). But their predictions can easily be manipulated using adversarial examples, which are crafted by adding imperceptible perturbations to legitimate documents. This vulnerability raises significant concerns about their reliability and hinders the widespread deployment of NRMs. By incorporating adversarial examples into training data, adversarial training has become the de facto defense approach to adversarial attacks against NRMs. However, this defense mechanism is subject to a trade-off between effectiveness and adversarial robustness. In this study, we establish theoretical guarantees regarding the effectiveness-robustness trade-off in NRMs. We decompose the robust ranking error into two components, i.e., a *natural ranking error* for effectiveness evaluation and a *boundary ranking error* for assessing adversarial robustness. Then, we define the *perturbation invariance* of a ranking model and prove it to be a differentiable upper bound on the boundary ranking error for attainable computation. Informed by our theoretical analysis, we design a novel *perturbation-invariant adversarial training* (PIAT) method for ranking models to achieve a better effectiveness-robustness trade-off. We design a regularized surrogate loss, in which one term encourages the effectiveness to be maximized while the regularization term encourages the output to be smooth, so as to improve adversarial robustness. Experimental results on several ranking models demonstrate the superiority of PITA compared to existing adversarial defenses.

Introduction

Ranking is a fundamental problem in information retrieval (IR). With advances in deep learning (LeCun, Bengio, and Hinton 2015), neural ranking models (NRMs) (Guo et al. 2020) have achieved remarkable effectiveness. We have also witnessed substantial uptake of NRMs in practice (Lin, Nogueira, and Yates 2022). Recently, it has been demonstrated that NRMs are vulnerable to adversarial examples that are capable of inducing misbehavior with human-imperceptible perturbations (Wu et al. 2023; Liu et al. 2022; Chen et al. 2023). So far, little attention has been devoted to combating this issue. A representative and successful method for attacking NRMs is the word substitution ranking attack (WSRA), which promotes a target document in rankings by replacing important words with synonyms (Wu et al. 2023). Given the prevalence of black-hat search engine optimization (SEO) (Gyöngyi and Garcia-Molina 2005), enhancing the adversarial robustness of NRMs against such attacks is vital for their use in real-world scenarios.

Among adversarial defense mechanisms proposed to improve model robustness (Jia and Liang 2017; Raghunathan, Steinhardt, and Liang 2018; Madry et al. 2018), *adversarial training* remains the top-performer (Shafahi et al. 2019; Zhu et al. 2019). During adversarial training adversarial examples are fed to a model. However, this causes an undesirable reduction in effectiveness on natural (clean) samples, giving rise to a trade-off dilemma between effectiveness and robustness (Tsipras et al. 2019). This is because effectiveness concerns the overall performance under normal conditions, while adversarial robustness centers on performance under malicious behavior. Several refinements have been suggested for vanilla adversarial training, to mitigate the aforementioned trade-off in text and image classification (Zhang et al. 2019; Wang et al. 2021). However, clear differences exist between classification and ranking scenarios concerning the trade-off, given that the former relies on a single sample, whereas the latter involves a ranked list. So far, the ranking task has not benefited from these advances in bridging the gap between effectiveness and robustness. This naturally raises the first question:

> *What is the trade-off between effectiveness and robustness for ranking problems?*

We contribute a theoretical characterization of this question by decomposing the robust ranking error, i.e., the prediction error for adversarial examples, into two terms: (i) a *natural ranking error*, which focuses on the natural effectiveness of the ranked list predicted by the ranking model on clean data, and (ii) a *boundary ranking error*, which indicates the ranking model's adversarial robustness against adversarial examples, measuring the proximity of input features to the decision boundary. We then introduce the *perturbation invariance* of a ranking model, which says that any adversarial perturbation to candidate documents does not alter the resulting document ranking. We prove that the perturbation invariance is a differentiable upper bound on the boundary ranking error, which is sufficiently tight. Differences in measurements of these two errors, which express distinct optimization objectives, showcase the trade-off between effectiveness and robustness for ranking problems.

Next to the effectiveness-robustness trade-off, the second issue we address is:

> *How to design a defense mechanism against adversarial examples while maintaining competitive effectiveness for NRMs guided by our theoretical characterization?*

We introduce a novel *perturbation-invariant adversarial training method* (PIAT) to achieve this goal. The key idea is to capture the trade-off between natural and boundary ranking error by optimizing a regularized surrogate loss composed of two terms: (i) a *natural ranking loss*, which encourages the optimization of the natural ranking error by minimizing the "difference" between the predicted ranked list and the ground-truth based on supervised data, and (ii) an *adversarial ranking loss*, as the regularization term, which encourages the optimization of boundary ranking error by minimizing the "difference" between the predicted ranked list on natural candidates and on attacked candidates using semi-supervised learning. We propose three ways to implement the regularization term to ensure perturbation invariance. By combining supervised and semi-supervised training, we effectively leverage information from large-scale volumes of unlabeled documents to improve the effectiveness-robustness trade-off for NRMs.

Extensive experiments conducted on the widely-used MS MARCO passage ranking dataset show that PIAT offers superior defense against WSRA while maintaining effectiveness as compared to several empirical defense methods, including data augmentation and vanilla adversarial training. Ablations and visualizations are provided for more insights.

Preliminaries

Our work focuses on adversarial robustness to word substitution ranking attacks for NRMs. We review this type of attack in this section.

Attacks in web search. The web, as a canonical example of a competitive search setting, involves document authors who have incentives to optimize their content for better rankings in search results (Kurland and Tennenholtz 2022). This practice is commonly known as search engine optimization (SEO), which aims at improving the visibility and ranking of a web page in retrieved results when specific queries are entered by users (Gyöngyi and Garcia-Molina 2005). This can lead to a decrease in the overall quality of search results, as many irrelevant or low-quality documents may end up being ranked higher than they deserve, while more valuable and accurate content may get pushed down in the results.

Word substitution ranking attack. Recently, there has been much research on adversarial attacks against NRMs to simulate real-world ranking competitions. A representative study is the word substitution ranking attack (WSRA) (Wu et al. 2023), which demonstrates promising results in terms of the attack success rate. Given a ranking model, WSRA aims to promote a target document in rankings by replacing important words in its text with synonyms in a semantics-preserving way. Our research concentrates on WSRA attacks and develops a corresponding defense strategy.

Typically, in ad-hoc retrieval, given a query $\boldsymbol{q}$ and a set of document candidates $\mathcal{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{N_d}\}$, a neural ranking model f predicts the relevance score $f(\boldsymbol{q}, \boldsymbol{d}_i)$ of each query-document pair for ranking the whole candidate set. For example, f outputs the ranked list $[\boldsymbol{d}_{N_d}, \boldsymbol{d}_{N_d-1}, \ldots, \boldsymbol{d}_1]$ if it determines $f(\boldsymbol{q}, \boldsymbol{d}_{N_d}) > f(\boldsymbol{q}, \boldsymbol{d}_{N_d-1}) > \cdots > f(\boldsymbol{q}, \boldsymbol{d}_1)$. The rank position of document $\boldsymbol{d}_i$ with respect to query $\boldsymbol{q}$ predicted by f is $\pi_f(\boldsymbol{q}, \boldsymbol{d}_i)$. And we use $\pi_y(\boldsymbol{q}, \boldsymbol{d}_i)$ to represent the ground-truth rank position of $\boldsymbol{d}_i$ with respect to $\boldsymbol{q}$.

Given a target document $\boldsymbol{d} = (w_1, w_2, \ldots, w_M) \in \mathcal{D}$, the WSRA task constructs an adversarial example $\boldsymbol{d}' = (w_1', w_2', \ldots, w_M')$ by replacing at most $\epsilon \cdot M$ ($\epsilon \leq 1$) words in $\boldsymbol{d}$ with any of their synonyms w_m'. We denote a candidate set of adversarial examples (neighborhood) of $\boldsymbol{d}$ as $\mathbb{B}(\boldsymbol{d}, \epsilon)$, i.e.,

$$\mathbb{B}(\boldsymbol{d}, \epsilon) := \left\{\boldsymbol{d}' : \left\|\boldsymbol{d}' - \boldsymbol{d}\right\|_0 / \|\boldsymbol{d}\| \leq \epsilon\right\}, \tag{1}$$

where $\|\boldsymbol{d}\|$ represents the number of words in document $\boldsymbol{d}$, $\left\|\boldsymbol{d}' - \boldsymbol{d}\right\|_0 := \sum_{m=1}^{M} \mathbb{I}\{w_m' \neq w_m\}$ is the Hamming distance, with $\mathbb{I}\{\cdot\}$ the indicator function. Ideally, the goal of the attacker is to find $\boldsymbol{d}' \in \mathbb{B}(\boldsymbol{d}, \epsilon)$ such that $f(\boldsymbol{q}, \boldsymbol{d}') > f(\boldsymbol{q}, \boldsymbol{d})$ and $\boldsymbol{d}'$ has the same semantic meaning as $\boldsymbol{d}$.

Theoretical Analysis: The Trade-Off Between Effectiveness and Robustness

Tsipras et al. (2019) have shown that the goals of standard performance and adversarial robustness may be at odds. There can be an inherent trade-off between effectiveness and robustness. Drawing inspiration from the definitions of natural and robust accuracy in (Zhang et al. 2019), we characterize the trade-off in ranking by breaking down the robust ranking error into the sum of the natural ranking error and boundary ranking error. We also provide a differentiable upper bound on the boundary ranking error, to inform the design of the defense method.

Natural Ranking Error

So far, much effort in the field of NRMs has been dedicated to improving the ranking effectiveness, which is about the average performance under normal conditions.

Definition 1 (Natural ranking error) Formally, the *natural error* associated with the effectiveness of a ranking model f on natural (clean) examples is denoted as,

$$\mathcal{R}_{\mathrm{nat}}(f) := \mathbb{E}_{\boldsymbol{d}_i \sim \mathcal{D}} \mathbb{I}\{\pi_f(\boldsymbol{q}, \boldsymbol{d}_i) \neq \pi_y(\boldsymbol{q}, \boldsymbol{d}_i)\}, \tag{2}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function that is 1 if an event happens and 0 otherwise. For simplicity, we consider the $0-1$ loss in our theoretical analysis to evaluate the natural error.

Boundary Ranking Error

Here, we first define the decision boundary of a ranking model, and then introduce the boundary ranking error corresponding to the adversarial robustness of ranking models.

Definition 2 (Ranking decision boundary) For a ranking model f, we define the *ranking decision boundary* as the predicted rank position $\pi_f(\boldsymbol{q}, \boldsymbol{d}_i)$ being higher or lower than it truly deserves. Note that for the topmost and the bottommost ranks, we exclusively consider the situations where the predicted rank is one position lower and higher, respectively. Considering practical attacks aimed at ranking improvement, we denote $\pi_n(\boldsymbol{q}, \boldsymbol{d}_i) = \pi_y(\boldsymbol{q}, \boldsymbol{d}_i) - 1$ as the neighborhood rank of $(\boldsymbol{q}, \boldsymbol{d}_i)$. Recall that low values of rank positions attest to high ranking. In this way, the ranking decision boundary can be formulated as:

$$\mathrm{DB}(f) := \{\boldsymbol{d}_i \sim \mathcal{D} : \pi_f(\boldsymbol{q}, \boldsymbol{d}_i) = \pi_n(\boldsymbol{q}, \boldsymbol{d}_i)\} . \quad (3)$$

We use $\mathbb{B}(\boldsymbol{d}_i, \epsilon)$ to represent a neighborhood of $\boldsymbol{d}_i$ under the WSRA attack. Then, for a ranking model f, we denote the neighborhood of the decision boundary of f as:

$$\begin{aligned}&\mathbb{B}(\mathrm{DB}(f), \epsilon) := \{\boldsymbol{d}_i \sim \mathcal{D} : \ \exists\, \boldsymbol{d}_i' \in \mathbb{B}(\boldsymbol{d}_i, \epsilon) \text{ such that} \\ &[\pi_f(\boldsymbol{q}, \boldsymbol{d}_i) - \pi_n(\boldsymbol{q}, \boldsymbol{d}_i)] \cdot [\pi_f(\boldsymbol{q}, \boldsymbol{d}_i') - \pi_n(\boldsymbol{q}, \boldsymbol{d}_i)] \leq 0\}.\end{aligned} \quad (4)$$

This implies that $\boldsymbol{d}_i$ and $\boldsymbol{d}_i'$ are located on different sides of the decision boundary concerning the query $\boldsymbol{q}$. Therefore, a successful adversarial attack could move the target document to the wrong side of the decision boundary, leading to weak robustness of NRMs.

The above analysis elucidates why a ranking model with high effectiveness might still manifest considerable adversarial vulnerability. This discrepancy arises from the distinction between optimizing based on natural ranking error and acquiring a robust decision boundary for NRMs. Based on experimental findings due to Wu et al. (2023), we can tell: (i) decision boundaries learned based on natural ranking errors enable NRMs to achieve high effectiveness on clean documents, and (ii) such boundaries are susceptible to being breached by adversarial examples, resulting in vulnerabilities to easy attacks. Existing attack methods take advantage of this boundary vulnerability to deceive the NRM. As such, training robust NRMs requires a defining boundary ranking error to tackle this vulnerability effectively.

Definition 3 (Boundary ranking error) We introduce the *boundary ranking error* to assess the existence of adversarial examples near the ranking decision boundary of f, i.e.,

$$\begin{aligned}&\mathcal{R}_{\mathrm{bdy}}(f) := \\ &\mathbb{E}_{\boldsymbol{d}_i \sim \mathcal{D}} \mathbb{I}\left\{\boldsymbol{d}_i \in \mathbb{B}(\mathrm{DB}(f), \epsilon), \pi_f(\boldsymbol{q}, \boldsymbol{d}_i) = \pi_y(\boldsymbol{q}, \boldsymbol{d}_i)\right\} .\end{aligned} \quad (5)$$

Optimizing the boundary ranking error poses a challenge, mainly due to the large volume of unlabeled documents in the datasets and the unavailability of ground-truth rankings. To address this obstacle, we present a solution in the form of an upper bound on the boundary ranking error.

Theorem 1 (Upper bound of boundary ranking error) *According to Eq. 4 and 5, for a ranking model* $f : \boldsymbol{q} \times \mathcal{D} \to \mathbb{R}$ *and ranking mechanism* $r : \mathbb{R} \times \mathbb{R} \to \{\pm 1, 0\}$*, we have:*

$$\mathcal{R}_{\mathrm{bdy}}(f) \leq \mathbb{E}_{\boldsymbol{d}_i \sim \mathcal{D}} \max_{\boldsymbol{d}_i' \in \mathbb{B}(\boldsymbol{d}_i, \epsilon)} \mathbb{I}\left\{\pi_f(\boldsymbol{q}, \boldsymbol{d}_i) \neq \pi_f(\boldsymbol{q}, \boldsymbol{d}_i')\right\} . \quad (6)$$

Theorem 1 states the boundary ranking error can be upper-bounded by the expectation that any adversarial example maintaining its original ranking positions. This emphasizes the perturbation invariance of a robust ranking model, that is, any perturbation to the inputted candidate documents does not change the output ranking. Consequently, restraining the boundary ranking error is attainable by maximizing the outputted perturbation invariance of ranking models.

Nonetheless, if an upper bound is too loose, it may lead to the inadequacy of effectively optimizing the error. Hence, we further prove the upper bound in Theorem 1 is tight enough. The tightness ensures the reduction of the boundary ranking error through the optimization of perturbation invariance. The proof of Theorem 1 and its tightness are provided at https://github.com/ict-bigdatalab/PIAT.

Trade-Off Between Two Ranking Errors

Based on the definitions of natural error and boundary error for a ranking model, we present the robust ranking error for adversarial examples.

Definition 4 (Robust ranking error) To train a robust ranking model, the *robust ranking error* $\mathcal{R}_{\mathrm{rob}}(f)$ under the WSRA scenario, can be decomposed as follows,

$$\mathcal{R}_{\mathrm{rob}}(f) = \mathcal{R}_{\mathrm{nat}}(f) + \mathcal{R}_{\mathrm{bdy}}(f), \quad (7)$$

where $\mathcal{R}_{\mathrm{nat}}(f)$ corresponds to naturally wrongly ranked documents; and $\mathcal{R}_{\mathrm{bdy}}(f)$ corresponds to correctly ranked samples but close to the ϵ-extension of the ranking decision boundary. Consequently, these samples are susceptible to successful boundary-crossing attacks (i.e., ranked higher or lower) by introducing human-imperceptible perturbations.

Algorithmic Design: Perturbation-Invariant Adversarial Training

Inspired by our theoretical analysis, we present a new defense method for NRMs, named perturbation-invariant adversarial training (PIAT), to strike a balance between effectiveness and adversarial robustness.

Motivation

Theorem 1 and Definition 4 emphasize the importance of simultaneously optimizing the natural ranking error and the boundary ranking error, when training a robust ranking model while preserving effectiveness. We introduce a refinement to adversarial training, called PIAT, tailored specifically for ranking problems. This involves the incorporation of a *regularized surrogate loss* aimed at optimizing the robust ranking error, comprising two essential terms, i.e.,

$$\mathcal{L} = \lambda \mathcal{L}_{\mathrm{nat}} + (1 - \lambda)\mathcal{L}_{\mathrm{adv}}, \quad (8)$$

where the first term, i.e., the natural ranking loss $\mathcal{L}_{\mathrm{nat}}$, encourages the natural ranking error to be optimized, by minimizing the "difference" between the predicted and ground-truth ranked list. We achieve this by leveraging a traditional pair-wise loss, which is supervised using the labeled query-document pairs. The regularization term, i.e., the adversarial ranking loss $\mathcal{L}_{\mathrm{adv}}$, encourages the boundary ranking error to be optimized. We propose a perturbation-invariant ranking loss to minimize the "difference" between the prediction of a clean document set and that of an attacked document set, which is semi-supervised by the NRM's outputs. The λ is a trade-off parameter that controls the balance between effectiveness and robustness during training.

Natural Ranking Loss

The standard training of NRMs primarily emphasizes the model's effectiveness on the labeled dataset (Dai et al. 2018a; Dai and Callan 2019). In line with existing research, we adopt a pairwise loss as the natural ranking loss, i.e.,

$$\mathcal{L}_{\text{nat}} = -\frac{1}{|N_q|}\sum_{i=1}^{|N_q|}\log\frac{e^{f\left(\boldsymbol{q}_i,\boldsymbol{d}^+\right)}}{e^{f\left(\boldsymbol{q}_i,\boldsymbol{d}^+\right)}+\sum_{j=1}^{\bar{n}}e^{f\left(\boldsymbol{q}_i,\boldsymbol{d}_j^-\right)}}, \quad (9)$$

where N_q is the number of training queries, $\boldsymbol{d}^+$ is the relevant document and $\boldsymbol{d}^-$ is the irrelevant document. We use the negative examples returned by the retrieval stage as hard negative examples and also incorporate random negative examples for the same purpose (Nogueira and Cho 2019).

Adversarial Ranking Loss

To enhance adversarial robustness, we first use the WSRA attack to generate adversarial examples. Subsequently, we utilize augmented adversarial examples to optimize the proposed perturbation-invariant ranking loss.

Adversarial examples. To execute a WSRA attack in a decision-based black-box setting, Wu et al. (2023) introduce a pseudo-relevance based adversarial ranking attack method to generate adversarial examples. Following this work, for each query $\boldsymbol{q}$, given a candidate document set $\mathcal{D}$, we conduct the attack against a portion of the documents evenly to derive the adversarial examples $\mathcal{D}_{adv}$. Each adversarial example $\boldsymbol{d}_{adv}$ in $\mathcal{D}_{adv}$ is selected from the neighborhood of the original document $\boldsymbol{d}$, based on the most threatening attack effect, i.e.,

$$\boldsymbol{d}_{adv} = \arg\max_{\boldsymbol{d}'\in\mathbb{B}(\boldsymbol{d},\epsilon)}\left(f(\boldsymbol{q},\boldsymbol{d}') - f(\boldsymbol{q},\boldsymbol{d})\right). \quad (10)$$

Thus, we obtain adversarial examples $\mathcal{D}_{adv}$ for each query $\boldsymbol{q}$, which will be used in the following loss.

Perturbation-invariant ranking loss. As the regularization term in Eq. 8, the adversarial ranking loss encourages the model's output to be smooth, effectively constraining the sample instances within adjacent ranking decision boundaries of the model. This is achieved by minimizing the ranking order variance between the prediction of natural documents $\mathcal{D}$ and that of adversarial examples $\mathcal{D}_{adv}$. We design the perturbation-invariant ranking loss between $\mathcal{D}$ and $\mathcal{D}_{adv}$ as the adversarial ranking loss, i.e.,

$$\mathcal{L}_{\text{adv}} = -\frac{1}{|N_q|}\sum_{i=1}^{|N_q|}\psi\left(f\left(\boldsymbol{q}_i,\mathcal{D}\right), f\left(\boldsymbol{q}_i,\mathcal{D}_{adv}\right)\right), \quad (11)$$

where $f(\boldsymbol{q},\mathcal{D})$ is the predicted ranked list by a ranking model f over $\mathcal{D}$; $\psi(\cdot)$ is a differential metric to evaluate the difference in the resulting document rankings between $\mathcal{D}$ and $\mathcal{D}_{adv}$. Here, $\mathcal{D}_{adv}$ comprises N_{adv} perturbed documents and $N_d - N_{adv}$ benign documents.

We consider three ways to compute the difference $\psi(\cdot)$ in the ranked results obtained using $\mathcal{D}$ and $\mathcal{D}_{adv}$.

(1) **KL divergence.** To promote smoothness between $\mathcal{D}$ and $\mathcal{D}_{adv}$ during optimization, our objective is to minimize the KL divergence between the similarity distributions of the ranking model f. As a result, the computation of $\mathcal{L}_{\text{adv}}$ in Eq. 11 using the KL divergence, is as follows:

$$\begin{aligned}\mathcal{L}_{\text{adv}}^{\text{KL}} &= \text{KL}(P(\mathcal{S}\mid\mathcal{Q},\mathcal{D};f)\parallel P(\mathcal{S}\mid\mathcal{Q},\mathcal{D}_{adv};f)) \\ &= \frac{1}{N_q}\sum_{i=1}^{N_q}\sum_{j=1}^{N_d}P(s_i\mid\boldsymbol{q}_i,\boldsymbol{d}_j\sim\mathcal{D};f)\cdot \\ &\quad \log\frac{P(s_i\mid\boldsymbol{q}_i,\boldsymbol{d}_j\sim\mathcal{D};f)}{P(s_i\mid\boldsymbol{q}_i,\boldsymbol{d}'_j\sim\mathcal{D}_{adv};f)},\end{aligned} \quad (12)$$

where

$$P(s_i\mid\boldsymbol{q}_i,\boldsymbol{d}_j\sim\mathcal{D};f) = \frac{\exp(f\left(\boldsymbol{q}_i,\boldsymbol{d}_j\right))}{\sum_{\boldsymbol{d}_k\sim\mathcal{D}}\exp(f\left(\boldsymbol{q}_i,\boldsymbol{d}_k\right))},$$

$$P(s_i\mid\boldsymbol{q}_i,\boldsymbol{d}'_j\sim\mathcal{D}_{adv};f) = \frac{\exp\left(f\left(\boldsymbol{q}_i,\boldsymbol{d}'_j\right)\right)}{\sum_{\boldsymbol{d}'_k\sim\mathcal{D}_{adv}}\exp\left(f\left(\boldsymbol{q}_i,\boldsymbol{d}'_k\right)\right)}.$$

Let us consider a scenario where only one document ranked at the bottom within $\mathcal{D}$ is perturbed and moves to the top-1 position, while the other documents are shifted down one position each. In this case, the distribution of the entire permutation would not undergo significant disordering. However, even though the overall re-ordering might be limited, the situation could have implications for practical search engines. Therefore, using KL divergence as a metric may not impose a sufficiently severe penalty for this attack result.

Next, we present alternatives to tackle this issue. We introduce a listwise loss to model the output ranking both before and after perturbation. By concentrating on the ranked list, our approach strives to prevent the perturbed document from excessively rising to the top position, thereby preserving a natural and gradual change in rankings.

(2) **Listwise function – ListNet.** ListNet (Cao et al. 2007) devises a listwise loss to assess the dissimilarity between the predicted ranked list and the ground-truth permutation, given by the following expression:

$$\begin{aligned}&\mathcal{L}_{\text{ListNet}}\left(f;\boldsymbol{q},\mathcal{D},\mathcal{Y}\right) = \\ &\quad \text{KL}\left(P(\pi_f\mid\varphi(f(\boldsymbol{q},\mathcal{D})))\parallel P(\pi_{\mathcal{Y}})\right),\end{aligned} \quad (13)$$

where π_f is the permutation predicted by f, $\pi_{\mathcal{Y}}$ is the ground-truth permutation, and φ is a transformation function (an increasing and strictly positive function, e.g., linear, exponential or sigmoid). The probability of a permutation given the score list (Cao et al. 2007), is computed as follows,

$$P(\pi_f\mid\varphi(f(\boldsymbol{q},\mathcal{D}))) = \prod_{j=1}^{N_d}\frac{\varphi\left(f_{\pi(j)}(\boldsymbol{q},\mathcal{D})\right)}{\sum_{k=j}^{N_d}\varphi\left(f_{\pi(k)}(\boldsymbol{q},\mathcal{D})\right)}, \quad (14)$$

where $f_{\pi(i)}(\boldsymbol{q},\mathcal{D})$ denotes the similarity score predicted by f of the document, which is ranked at the i-th position with respect to the query $\boldsymbol{q}$. We define $\mathcal{L}_{adv}$ based on ListNet as,

$$\begin{aligned}&\mathcal{L}_{\text{adv}}^{\text{ListNet}} = \\ &\text{KL}\left(P(\pi_{f(\boldsymbol{q},\mathcal{D})}\mid f(\boldsymbol{q},\mathcal{D}_{adv}))\| P(\pi_{f(\boldsymbol{q},\mathcal{D})}\mid f(\boldsymbol{q},\mathcal{D}))\right),\end{aligned} \quad (15)$$

where $\pi_{f(\boldsymbol{q},\mathcal{D})}$ is the permutation computed by the ranking model f on data pair $(\boldsymbol{q},\mathcal{D})$.

(3) **Listwise function – ListMLE.** ListMLE (Xia et al. 2008) addresses the computational complexity of ListNet by optimizing the negative log-likelihood of the ground-truth permutation $\pi_{\mathcal{Y}}$, i.e.,

$$\mathcal{L}_{\mathrm{ListMLE}}\left(f; \boldsymbol{q}, \mathcal{D}, \mathcal{Y}\right) = -\log P(\pi_{\mathcal{Y}} \mid f(\boldsymbol{q}, \mathcal{D})). \quad (16)$$

Inspired by ListMLE, we design our adversarial loss $\mathcal{L}_{adv}$ to compute the negative log-likelihood of benign document permutation, i.e.,

$$\mathcal{L}_{\mathrm{adv}}^{\mathrm{ListMLE}} = -\log P(\pi_{f(\boldsymbol{q},\mathcal{D})} \mid f(\boldsymbol{q}, \mathcal{D}_{adv})), \quad (17)$$

where $\pi_{f(\boldsymbol{q},\mathcal{D})}$ represents the document permutation generated by the ranking model f for the list of documents that have not been attacked. This enables us to effectively align the ranked list after perturbations with the benign ranked list, thereby achieving adversarial robustness.

Experiments

We present our experimental setup and results in this section.

Experimental Setup

Dataset and target ranking models. We conduct experiments on the *MS MARCO Passage Ranking dataset*, which is a large-scale benchmark dataset for Web passage retrieval, with about 8.84 million passages (Nguyen et al. 2016). The relevant documents to user queries are obtained using Bing, thereby simulating real-world web search scenarios.

We choose several typical ranking models that achieve promising effectiveness, including traditional probabilistic models, e.g., *BM25* (Robertson and Walker 1994), interaction-focused NRMs, e.g., *ConvKNRM* (Dai et al. 2018b), and pre-trained models, e.g., *BERT* (Devlin et al. 2019) and *PROP* (Ma et al. 2021a), for adversarial attack.

Evaluation metrics.

(i) *CleanMRR@k* evaluates Mean Reciprocal Rank (MRR) performance on the clean dataset (Ma et al. 2021b; Yan et al. 2021). (ii) *RobustMRR@k* evaluates the MRR performance on the attacked dataset by WSRA. (iii) *Attack success rate (ASR)* (%) evaluates the percentage of the after-attack documents that are ranked higher than original documents (Wu et al. 2023). (iv) *Location square deviation (LSD)* (%) evaluates the consistency between the original and perturbed ranked list for a query, by calculating the average deviation between the document positions in the two lists (Sun, Li, and Zhao 2022).

The effectiveness of a ranking model is better with a higher CleanMRR. The robustness of a ranking model is better with a higher RobustMRR and a lower ASR and LSD.

Baselines. (i) *Standard training (ST)*: We directly optimize the ranking model via the natural ranking loss (Eq. 9) without defense mechanisms. (ii) *Data augmentation (DA)*: We augment each document in the collection with 2 new documents by uniformly replacing synonyms, and then use the normal hinge loss for training following (Wu et al. 2022). The number of replacement words equals the number of words perturbed by the WSRA attack. (iii) *Adversarial training (AT)*: We follow the vanilla AT method (Goodfellow, Shlens, and Szegedy 2015) to directly include the adversarial examples during training. (iv) *CertDR* is a certified defense method for NRMs (Wu et al. 2022), which achieves certified top-K robustness against WSRA attacks.

Implementation details. We implement target ranking models following previous work (Dai et al. 2018b; Devlin et al. 2019; Ma et al. 2021a; Liu et al. 2023b). First-stage retrieval is performed using the Anserini toolkit (Yang, Fang, and Lin 2018) with BM25, to obtain top 100 candidate passages. The ranked list is obtained by using the well-trained ranking model to re-rank the above initial candidate set.

We randomly sample 1000 Dev queries as target queries to attack their ranked lists for evaluation. For each sampled query, we randomly sample 1 document from 9 ranges in the ranked list following (Wu et al. 2023), i.e., $[11, 20], ..., [91, 100]$, respectively. We attack these 9 target documents to achieve their corresponding adversarial examples using WSRA. Finally, we evaluate the defense performance of ranking models using the attacked list with 9 adversarial examples and its query as an input. For BM25, we attack it using adversarial examples generated by the attack method in (Wu et al. 2023) designed for attacking BERT.

For adversarial training, considering the time overhead, we sample 0.1 million (1/10 of the total) training queries to generate adversarial examples. For each training query, we randomly sample 10 documents from its initial candidate set to construct adversarial examples using WSRA. Note the sampled documents are not ground-truth ones. We set the maximum number of word substitutions to 20, and other hyperparameters are consistent with Wu et al. (2023). The regularization hyperparameter λ is set to 0.5. We train the NRMs with a batch size of 100, maximum sequence length of 256, and learning rate of 1e-5.

By training the ranking model with different adversarial ranking losses, i.e., $\mathcal{L}_{\mathrm{adv}}^{\mathrm{KL}}$, $\mathcal{L}_{\mathrm{adv}}^{\mathrm{ListNet}}$, and $\mathcal{L}_{\mathrm{adv}}^{\mathrm{ListMLE}}$, we obtain three types of PIAT as *PIAT*$_{\mathrm{KL}}$, *PIAT*$_{\mathrm{ListNet}}$, and *PIAT*$_{\mathrm{ListMLE}}$, respectively.

Experimental Results

Defense comparison. Table 1 presents a comparison of the trade-off performance among four ranking models with different defenses. Observations on the defense baselines are: (i) Effectiveness and adversarial robustness of PROP is generally better than BERT, which in turn is stronger than ConvKNRM. This indicates that well-designed model architectures and pre-training objectives encourage a ranking model to achieve better trade-off performance. (ii) After being attacked, the ranking performance of the ST method without defense mechanisms, decreases significantly with a high ASR and LSD. Hence, it is imperative not only to focus on the effectiveness of existing NRMs when deploying them in real-world scenarios. (iii) CertDR ensures consistent ranking performance between clean and adversarial data. This could be attributed to CertDR's ability to guarantee the stability of the Top-K of the ranked list by certifying the Top-K robustness. (iv) DA and AT enhance the model's ranking performance on adversarial data, but this improvement comes at the cost of reduced performance on clean data. The finding is consistent with prior research in natural language processing and machine learning (Zhang et al. 2019; Rade and Moosavi-Dezfooli 2021; Bao, Wang, and Zhao 2021).

Model	Method	CleanMRR@10	CleanMRR@100	RobustMRR@10	RobustMRR@100	ASR↓	LSD↓
BM25	-	0.1874	0.1985	0.1624	0.1736	56.4	15.3
ConvKNRM	ST	0.2461	0.2592	0.1692	0.1741	95.1	33.2
	DA	0.2298	0.2378	0.1786	0.1829	71.2	23.6
	CertDR	0.1816	0.1935	0.1592	0.1632	65.3	19.3
	AT	0.2316	0.2410	0.1896	0.1953	61.3	17.6
	PIAT_{KL}	0.2498	0.2603	0.2008*	0.2073*	51.1*	12.6
	$\text{PIAT}_{\text{ListNet}}$	0.2513*	0.2621*	**0.2035***	**0.2009***	**48.3***	**11.5***
	$\text{PIAT}_{\text{ListMLE}}$	**0.2534***	**0.2645***	0.2018*	0.2091*	49.2*	11.9
BERT	ST	0.3831	0.3923	0.3225	0.3286	92.1	32.3
	DA	0.3705	0.3810	0.3315	0.3423	63.2	18.9
	CertDR	0.3202	0.3311	0.3026	0.3140	56.9	16.3
	AT	0.3743	0.3865	0.3451	0.3508	55.1	15.6
	PIAT_{KL}	0.3860	0.3948	0.3686*	0.3761*	41.2*	9.4
	$\text{PIAT}_{\text{ListNet}}$	0.3892	0.3981	**0.3728***	**0.3802***	**36.1***	**7.2***
	$\text{PIAT}_{\text{ListMLE}}$	**0.3910***	**0.4002***	0.3705*	0.3785*	38.3*	7.9
PROP	ST	0.3902	0.4061	0.3352	0.3478	90.3	30.9
	DA	0.3783	0.3930	0.3418	0.3538	60.4	16.8
	CertDR	0.3351	0.3489	0.3199	0.3220	52.8	13.4
	AT	0.3819	0.4002	0.3532	0.3611	51.2	12.8
	PIAT_{KL}	0.3943	0.4063	0.3749*	0.3853*	39.4*	8.2
	$\text{PIAT}_{\text{ListNet}}$	0.3971	0.4121	**0.3794***	**0.3890***	**35.0***	**6.2***
	$\text{PIAT}_{\text{ListMLE}}$	**0.3992***	**0.4148***	0.3767*	0.3864*	37.8*	7.8

Table 1: Trade-off performance of different ranking models under PIAT and defense baselines; For CertDR, the ASR is evaluated under conditional success rate (Wu et al. 2022); $*$ indicates significant improvements over the best baseline ($p \leq 0.05$).

When we look at PIAT, we find that: (i) In general, three types of PIAT exhibit superior effectiveness and adversarial robustness than baselines. This suggests that a combination of proposed supervised and semi-supervised training enables the effective utilization of information from extensive unlabeled documents to enhance trade-off performance. (ii) PIAT outperforms the baselines in terms of LSD, indicating increased resistance to perturbations across the entire ranked list. This highlights the efficacy of the perturbation-invariant ranking loss in facilitating NRMs to learn more robust ranking decision boundaries. (iii) PIAT_{KL} demonstrates comparatively lower effectiveness in comparison to the other two PIAT types, likely due to the fact that the KL divergence of the relevant scores serves as a soft constraint, rendering a relatively mild supervisory signal. (iv) $\text{PIAT}_{\text{ListMLE}}$ achieves a slightly inferior performance compared to $\text{PIAT}_{\text{ListNet}}$. The reason might be that ListMLE compromises precision by converting list-wise differences into an estimate of the probability distribution. Nevertheless, ListMLE exhibits higher training efficiency.

Effectiveness vs. robustness trade-off. λ is an important hyperparameter in our proposed method, since it plays a crucial role in determining the balance between effectiveness and robustness. Figure 1 shows a comparison of the effectiveness vs. robustness trade-off between PIAT and empirical defense baselines. CertDR is excluded from the comparison due to its inferior performance compared to empirical defenses in terms of both effectiveness and robustness, as indicated in Table 1. We conduct comparisons by examining CleanMRR@10 (for effectiveness) against

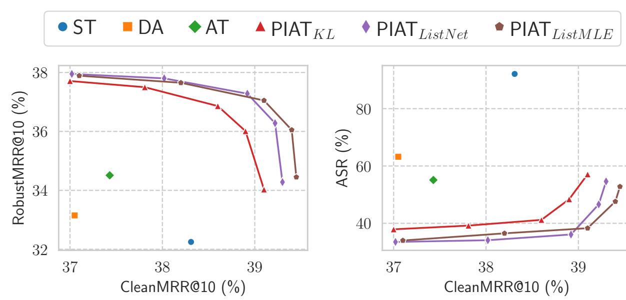

Figure 1: The sensitivity of trade-off parameter λ exhibited by different types of PIAT, compared with empirical defense methods. From left to right, we increase the trade-off parameter λ of PIAT from 0.2 to 0.8 with the step of 0.15.

RobustMRR@10 and ASR (for robustness) respectively, thereby visualizing the effectiveness-robustness trade-off.

We show the results of the BERT model; similar findings were obtained for other ranking models. Both DA and AT enhance robustness, but at the expense of effectiveness. This suggests they may not adequately consider the balanced relationship between effectiveness and robustness. When we look at the different types of PIAT we find that they achieve a heightened trade-off between effectiveness and robustness. This indicates that proper modeling and optimization of the boundary ranking error can guide NRMs to bolster robustness, while maintaining or even improving effectiveness. Furthermore, we note that with an excessively large λ, the model's robustness considerably decreases, while effective-

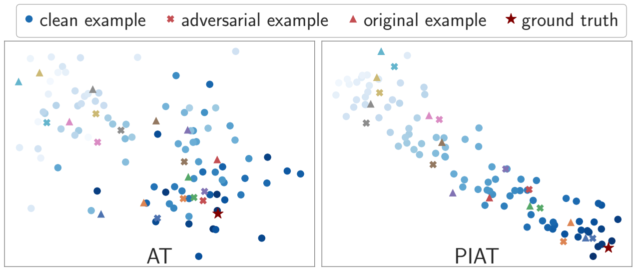

Figure 2: t-SNE plot of query-document representations for AT and PIAT. For clean examples, a darker blue circle represents a higher relevance score. Triangle and cross denote the original document and its corresponding adversarial example. Star denotes the ground-truth query-document pair.

ness exhibits marginal growth. Conversely, an excessively small λ shows a notable decline in effectiveness, while robustness experiences only a minor improvement. These findings emphasize the necessity of prioritizing the balance of effectiveness and robustness when training NRMs.

Visual analysis. We train the BERT model using AT and PIAT$_{ListMLE}$, respectively, on the MS MARCO passage ranking dataset, with inputs being query-document concatenations. The hidden states of [CLS] in BERT's final layer are utilized as query-document pair representations and visualized using t-SNE (Van der Maaten and Hinton 2008) to observe semantic space distributions. As a t-SNE example, we generate plots by sampling a query (QID=262232) and selecting its top 100 candidate documents, including 9 adversarial examples. Results in Figure 2 show that: (i) For AT, the distribution of adversarial examples in the latent space is relatively disordered. By examining data at the same relevance level (color depth), we find that the decision boundaries exhibit a certain degree of chaos. Some adversarial examples have managed to move away from their original positions and closer to the ground-truth. AT lacks clear distinctions in terms of modeling effectiveness and robustness, relying solely on pair-wise loss for simultaneous optimization. (ii) For PIAT, the ranking decision boundary not only distinguishes between data points of varying relevance levels, but also effectively constrains the adversarial examples to stay close to their original examples. This result emphasizes the fact that by using perturbation-invariant loss, tailored through analysis of boundary ranking errors, PIAT achieves remarkable adversarial robustness while maintaining effectiveness compared to the traditional AT. Similar observations were obtained with other ranking models.

Related Work

Neural ranking models. The emergence of deep learning has led to the popularity of NRMs (Onal et al. 2018; Guo et al. 2020), showcasing their superiority over traditional ranking models. There have been efforts to leverage pre-trained models for ranking tasks (Fan et al. 2022), further enhancing the effectiveness of NRMs. Additionally, studies have explored training NRMs using data augmentation techniques, such as hard negative mining (Xiong et al. 2021; Zhan et al. 2021), achieving new state-of-the-art performance. Despite these effectiveness improvements, these studies often overlook the adversarial robustness of NRMs.

Defense methods. Adversarial attacks aim to discover human-imperceptible perturbations that can deceive neural networks (Szegedy et al. 2014). In IR, there is growing interest in robustness (Liu et al. 2023a) and adversarial attacks. Wu et al. (2023) introduced the WSRA method of attacking black-box NRMs using word substitution. This study revealed the serious vulnerability of NRMs to synonym substitution perturbations. As a result, subsequent explorations of attack against NRMs have emerged (Liu et al. 2023c, 2022; Chen et al. 2023), inspired by this pioneering work.

In response to adversarial attacks, research has proposed various defense strategies to enhance adversarial robustness. These can be generally classified into *certified defenses* and *empirical defenses*. Certified defenses aim for theoretical robustness against specific adversarial perturbations (Raghunathan, Steinhardt, and Liang 2018). For instance, Wu et al. (2022) introduced a certified defense method that ensures the top-K robustness of NRMs via randomized smoothing. However, due to their theoretical nature, these methods often face limitations in practical applications and may not fully meet the desired performance requirements.

Empirical defenses aim to enhance the empirical robustness of models against known adversarial attacks, and this approach has been extensively explored in image classification (Madry et al. 2018; Wang et al. 2019) and text classification (Ye, Gong, and Liu 2020; Jia et al. 2019). Among these methods, adversarial training emerges as one of the most effective defenses. Adversarial training on adversarial examples remains empirically robust (Cui et al. 2021). However, the use of adversarial training as a defensive mechanism is often limited to simple classification scenarios, and its application in NRMs remains largely unexplored. Therefore, we propose an adversarial training method tailored for NRMs to improve the trade-off between effectiveness and robustness.

Conclusion

To the best of our knowledge, our study is the first study on the trade-off between effectiveness and adversarial robustness for neural retrieval models. Our theoretical analysis motivated the development of perturbation-invariant adversarial training, incorporating a new regularized surrogate loss. Experimental results have showcased the superior performance of our method in terms of effectiveness and robustness.

Broader impact and limitations. We aim for our initial exploration to serve as a benchmark for adversarial robustness and to inspire the IR community to further enhance the effectiveness-robustness trade-off. As to the limitations of our work, we currently only consider the popular attack of WSRA, and constructing adversarial training examples could be time-consuming. In future work, we will investigate the design of adversarial training methods to defend against other or unseen attacks, and create training examples with reduced time overhead. Besides, we will consider more benchmark datasets to simulate different retrieval scenarios.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 62006218 and 61902381, the Youth Innovation Promotion Association CAS under Grants No. 2021100, the project under Grants No. 2023YFA1011602, JCKY2022130C039 and 2021QY1701, the CAS Project for Young Scientists in Basic Research under Grant No. YSBR-034, the Innovation Project of ICT CAS under Grants No. E261090, and the Lenovo-CAS Joint Lab Youth Scientist Project. This work was also (partially) funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, and project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Bao, R.; Wang, J.; and Zhao, H. 2021. Defending Pre-trained Language Models from Adversarial Word Substitution Without Performance Sacrifice. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3248–3258.

Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; and Li, H. 2007. Learning to Rank: From Pairwise Approach to Listwise Approach. In *Proceedings of the 24th International Conference on Machine learning*, 129–136.

Chen, X.; He, B.; Ye, Z.; Sun, L.; and Sun, Y. 2023. Towards Imperceptible Document Manipulations against Neural Ranking Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6648–6664.

Cui, J.; Liu, S.; Wang, L.; and Jia, J. 2021. Learnable Boundary Guided Adversarial Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15721–15730.

Dai, Z.; and Callan, J. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of the 42nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, 985–988.

Dai, Z.; Xiong, C.; Callan, J.; and Liu, Z. 2018a. Convolutional Neural Networks for Soft-matching N-grams in Ad-hoc Search. In *Proceedings of the eleventh ACM international Conference on Web Search and Data Mining*, 126–134.

Dai, Z.; Xiong, C.; Callan, J.; and Liu, Z. 2018b. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*.

Fan, Y.; Xie, X.; Cai, Y.; Chen, J.; Ma, X.; Li, X.; Zhang, R.; and Guo, J. 2022. Pre-training Methods in Information Retrieval. *Foundations and Trends in Information Retrieval*, 16(3): 178–317.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Guo, J.; Fan, Y.; Pang, L.; Yang, L.; Ai, Q.; Zamani, H.; Wu, C.; Croft, W. B.; and Cheng, X. 2020. A Deep Look into Neural Ranking Models for Information Retrieval. *Information Processing & Management*, 57(6): 102067.

Gyöngyi, Z.; and Garcia-Molina, H. 2005. Web Spam Taxonomy. In *AIRWeb 2005: First International Workshop on Adversarial Information Retrieval on the Web*, 39–47.

Jia, R.; and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2021–2031.

Jia, R.; Raghunathan, A.; Göksel, K.; and Liang, P. 2019. Certified Robustness to Adversarial Word Substitutions. 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)*, 4129–4142.

Kurland, O.; and Tennenholtz, M. 2022. Competitive Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2838–2849.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep Learning. *Nature*, 521(7553): 436–444.

Lin, J.; Nogueira, R.; and Yates, A. 2022. *Pretrained Transformers for Text Ranking: Bert and Beyond*. Springer Nature.

Liu, J.; Kang, Y.; Tang, D.; Song, K.; Sun, C.; Wang, X.; Lu, W.; and Liu, X. 2022. Order-Disorder: Imitation Adversarial Attacks for Black-box Neural Ranking Models. In *2022 ACM SIGSAC Conference on Computer and Communications Security*, 2025–2039.

Liu, Y.-A.; Zhang, R.; Guo, J.; Chen, W.; and Cheng, X. 2023a. On the Robustness of Generative Retrieval Models: An Out-of-Distribution Perspective. In *Gen-IR@SIGIR*.

Liu, Y.-A.; Zhang, R.; Guo, J.; de Rijke, M.; Chen, W.; Fan, Y.; and Cheng, X. 2023b. Black-box Adversarial Attacks against Dense Retrieval Models: A Multi-view Contrastive Learning Method. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1647–1656.

Liu, Y.-A.; Zhang, R.; Guo, J.; de Rijke, M.; Chen, W.; Fan, Y.; and Cheng, X. 2023c. Topic-oriented Adversarial Attacks against Black-box Neural Ranking Models. In *46th international ACM SIGIR Conference on Research and Development in Information Retrieval*, 1700–1709.

Ma, X.; Guo, J.; Zhang, R.; Fan, Y.; Ji, X.; and Cheng, X. 2021a. Prop: Pre-training with Representative Words Prediction for Ad-hoc Retrieval. In *Proceedings of the 14th ACM international Conference on Web Search and Data Mining*, 283–291.

Ma, X.; Guo, J.; Zhang, R.; Fan, Y.; Li, Y.; and Cheng, X. 2021b. B-PROP: Bootstrapped Pre-training with Representative Words Prediction for Ad-hoc Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1513–1522.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. In *CoCo@ NIPS*.

Nogueira, R.; and Cho, K. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Onal, K. D.; Zhang, Y.; Altingovde, I. S.; Rahman, M. M.; Karagoz, P.; Braylan, A.; Dang, B.; Chang, H.-L.; Kim, H.; McNamara, Q.; Angert, A.; Banner, E.; Khetan, V.; McDonnell, T.; Nguyen, A. T.; Xu, D.; Wallace, B. C.; de Rijke, M.; and Lease, M. 2018. Neural Information Retrieval: At the End of the Early Years. *Information Retrieval*, 21(2–3): 111–182.

Rade, R.; and Moosavi-Dezfooli, S.-M. 2021. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*.

Raghunathan, A.; Steinhardt, J.; and Liang, P. 2018. Certified Defenses against Adversarial Examples. In *International Conference on Learning Representations*.

Robertson, S. E.; and Walker, S. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 232–241. Springer.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial Training for Free! *Advances in Neural Information Processing Systems*, 32.

Sun, K.; Li, Z.; and Zhao, H. 2022. Reorder and then Parse, Fast and Accurate Discontinuous Constituency Parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10575–10588.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations*, 2019.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, X.; Yang, Y.; Deng, Y.; and He, K. 2021. Adversarial Training with Fast Gradient Projection Method Against Synonym Substitution based Text Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13997–14005.

Wang, Y.; Ma, X.; Bailey, J.; Yi, J.; Zhou, B.; and Gu, Q. 2019. On the Convergence and Robustness of Adversarial Training. In *ICML*.

Wu, C.; Zhang, R.; Guo, J.; Chen, W.; Fan, Y.; de Rijke, M.; and Cheng, X. 2022. Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2128–2137.

Wu, C.; Zhang, R.; Guo, J.; de Rijke, M.; Fan, Y.; and Cheng, X. 2023. PRADA: Practical Black-box Adversarial Attacks Against Neural Ranking Models. *ACM Transactions on Information Systems*, 41(4): Article 89.

Xia, F.; Liu, T.-Y.; Wang, J.; Zhang, W.; and Li, H. 2008. Listwise Approach to Learning to Rank: Theory and Algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, 1192–1199.

Xiong, L.; Xiong, C.; Li, Y.; Tang, K.-F.; Liu, J.; Bennett, P.; Ahmed, J.; and Overwijk, A. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *arXiv preprint arXiv:2007.00808*.

Yan, M.; Li, C.; Bi, B.; Wang, W.; and Huang, S. 2021. A Unified Pretraining Framework for Passage Ranking and Expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4555–4563.

Yang, P.; Fang, H.; and Lin, J. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *Journal of Data and Information Quality (JDIQ)*, 10(4): 1–20.

Ye, M.; Gong, C.; and Liu, Q. 2020. SAFER: A Structure-free Approach for Certified Robustness to Adversarial Word Substitutions. *arXiv preprint arXiv:2005.14424*.

Zhan, J.; Mao, J.; Liu, Y.; Guo, J.; Zhang, M.; and Ma, S. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *International Conference on Machine Learning*, 7472–7482. PMLR.

Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; and Liu, J. 2019. FreeLB: Enhanced Adversarial Training for Natural Language Understanding. *arXiv preprint arXiv:1909.11764*.