UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models

Wu, C.; Zhang, R.; Guo, J.; Chen, W.; Fan, Y.; de Rijke, M.; Cheng, X.

DOI
10.1145/3511808.3557256

Publication date
2022

Document Version
Final published version

Published in
CIKM '22

License
Unspecified

Link to publication

Citation for published version (APA):
Wu, C., Zhang, R., Guo, J., Chen, W., Fan, Y., de Rijke, M., & Cheng, X. (2022). Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models. In *CIKM '22: proceedings of the 31st ACM International Conference on Information & Knowledge Management : October 17-21, 2022, Atlanta, GA, USA* (pp. 2128-2137). The Association for Computing Machinery. https://doi.org/10.1145/3511808.3557256

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:12 Oct 2026

Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models

Chen Wu
wuchen17z@ict.ac.cn
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China

Ruqing Zhang
Jiafeng Guo*
zhangruqing@ict.ac.cn
guojiafeng@ict.ac.cn
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China

Wei Chen
Yixing Fan
chenwei2022@ict.ac.cn
fanyixing@ict.ac.cn
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China

Maarten de Rijke
m.derijke@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Xueqi Cheng
cxq@ict.ac.cn
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China

ABSTRACT

Neural ranking models (NRMs) have achieved promising results in information retrieval. NRMs have also been shown to be vulnerable to adversarial examples. A typical Word Substitution Ranking Attack (WSRA) against NRMs was proposed recently, in which an attacker promotes a target document in rankings by adding human-imperceptible perturbations to its text. This raises concerns when deploying NRMs in real-world applications. Therefore, it is important to develop techniques that defend against such attacks for NRMs. In empirical defenses adversarial examples are found during training and used to augment the training set. However, such methods offer no theoretical guarantee on the models' robustness and may eventually be broken by other sophisticated WSRAs. To escape this arms race, rigorous and provable certified defense methods for NRMs are needed.

To this end, we first define the *Certified Top-K Robustness* for ranking models since users mainly care about the top ranked results in real-world scenarios. A ranking model is said to be Certified Top-K Robust on a ranked list when it is guaranteed to keep documents that are out of the top K away from the top K under any attack. Then, we introduce a Certified Defense method, named CertDR, to achieve certified top-K robustness against WSRA, based on the idea of randomized smoothing. Specifically, we first construct a smoothed ranker by applying random word substitutions on the documents, and then leverage the ranking property jointly with the statistical property of the ensemble to provably certify top-K robustness. Extensive experiments on two representative web search datasets demonstrate that CertDR can significantly outperform state-of-the-art empirical defense methods for ranking models.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; **Adversarial retrieval**.

KEYWORDS

Certified Top-K Robustness, Certified Defense, Word Substitution Ranking Attack, Ranking Models

ACM Reference Format:
Chen Wu, Ruqing Zhang, Jiafeng Guo, Wei Chen, Yixing Fan, Maarten de Rijke, and Xueqi Cheng. 2022. Certified Robustness to Word Substitution Ranking Attack for Neural Ranking Models. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557256

1 INTRODUCTION

Neural ranking models (NRMs) [6, 30, 36], especially pre-trained ranking models [22, 27, 34], have led to substantial performance improvements in a wide range of search tasks [14, 20, 27]. We have also seen NRMs being used in various practical usages in the enterprise [25]. Despite their success, recent observations [48, 49] show that NRMs are vulnerable to adversarial examples. A typical word substitution ranking attack (WSRA) [48] was proposed and proved successful for attacking NRMs. In this setting, an attacker promotes a target document in rankings by replacing important words in its text with their synonyms in a semantic-preserving way.

*Jiafeng Guo is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557256

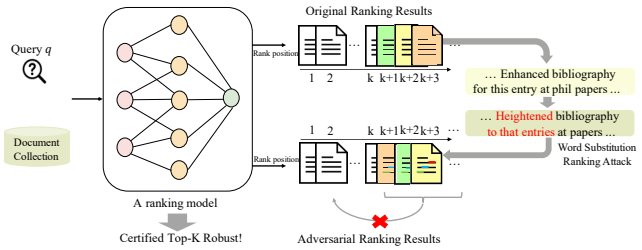

Figure 1: Illustration of the WSRA and *Certified Top-K Robustness*. Given a ranking model, it is said to be Certified Top-K Robust against WSRA on a ranked list if it is able to prevent documents outside top K from appearing in top K for all the possible WSRAs.

The adversarial documents generated by WSRAs are imperceptible to humans but can easily fool NRMs. This may bring great concerns when deploying these models to real-world applications. Thus, efficient methods for defending against such attacks are of critical importance for deploying modern NRMs to practical search engines.

Various approaches have been developed to defend against adversarial attacks. One representative type of defense is the empirical defense method [29, 47], where the set of perturbations is known at training time and adversarial examples are added to the training set [47]. However, it is insufficient when considering all possible WSRAs in which each word in a document can be replaced with any of its synonyms. Consequently, such defenses may eventually be broken by other sophisticated WSRAs [52]. To escape this arms race, procedures with rigorous and provable robustness guarantees are of special importance to the study of robustness of NRMs. In general, a model is said to be *certified robust* if such an attack is guaranteed to fail, no matter how the attacker manipulates the input. A line of work on certified defenses against any admissible adversarial attack has recently been introduced in image and text classification [5, 21, 44]. However, existing certified defense approaches are mainly for simple classification scenarios. They are quite distant from IR requirements, not just due to input differences (discrete text vs. continuous image [5]), but also due to the prediction type (ranked list vs. class label [52]). In this sense, the recent, promising advances in certified defenses have yet to bring similar robustness guarantees in how NRMs are approached.

Therefore, in this work, we make the first attempt to develop a certified defense method for NRMs so as to pursue adversarial immunization to WSRA to some extent. To this end, we need to answer two key questions: First, what is certified robustness in IR? Second, can we train NRMs that are robust in this sense?

To answer the first question, based on the previous definition and inspired by the IR properties, we propose to define the *Certified Top-K Robustness* of ranking models. In IR, it seems well-accepted that in many scenarios users mainly care about top ranked results [33, 50]. Some observations have shown that traffic (or click-through rate) falls off steeply as users work their way down the search results [19]. Many widely-used ranking metrics (e.g., MRR and nDCG) are concentrated on top ranked predictions. Therefore, as illustrated in Figure 1, a ranking model is said to be *Certified Top-K Robust against WSRA* on a ranked list if it is able to prevent documents outside the top K from appearing in top K for all possible WSRAs.

To answer the second question, we propose a novel **Cert**ified **D**efense method for **R**anking models, CertDR for short, to enhance a model's certified robustness against WSRA. To avoid exponential computational costs, our method is based on the idea of randomized smoothing [5, 52], which replaces the ranking model with a smoothed ranker for which it is easier to verify the certified robustness. Specifically, we first construct a smoothed ranker by averaging the output ranking scores of randomly perturbed documents. Then, we obtain a certification criterion to judge models' certified top-K robustness by leveraging the ranking property and statistical property of randomized ensembles. Finally, we design a practical certified defense algorithm, including a noise data augmentation strategy based on the perturbed documents for training and a rigorous statistical procedure to certify the top-K robustness.

We conduct experiments on two web search benchmark datasets, i.e., the MS MARCO document ranking dataset and the MS MARCO passage ranking dataset. We first compare the certified robustness among different ranking models (i.e., traditional probabilistic ranking models, and advanced neural ranking models) under CertDR. Based on the evaluation results, there clearly remains room for future certified robustness improvements. Besides, we compare with several state-of-the-art empirical defense methods and our experimental results show that CertDR can achieve the best defense performance against WSRA.

2 RELATED WORK

2.1 Text Ranking Models

Ranking models lie at the heart of IR. Many different ranking models have been proposed over the past decades, including probabilistic models [39, 43] (e.g., BM25 [43]) and learning-to-rank (LTR) models [26]. With the advance of deep learning, we have witnessed a substantial growth of interest in NRMs [6, 30, 36], achieving promising results in a variety of search tasks [14, 20, 27]. Recently, pre-trained ranking models such as BERT-based models [22, 34] have shown substantial performance improvements both in academic research and industry [25]. However, recent observations [48, 49] have shown that NRMs are vulnerable to adversarial examples. In this paper, we study how to defend against adversarial attacks for NRMs.

2.2 Adversarial Attacks

Adversarial attacks aim to generate human-imperceptible adversarial examples by perturbing inputs to maximally increase a model's risk of making wrong predictions. Adversarial examples were first discovered in the image domain [45], where early research has developed powerful white-box attack methods, e.g., Fast Gradient Sign Method [FGSM, 11] and Projected Gradient Descent [PGD, 28], for attacking continuous image data.

The existence and pervasiveness of adversarial examples have also been observed in the text domain [54]. Despite the fact that generating adversarial examples for texts has proven to be more challenging than for images due to their discrete nature, prior work has explored adversarial attacks for many language tasks, including text classification [10, 24], dialogue systems [4], and natural language inference [1]. Among these attacks, word substitution attacks [1, 41, 53], which replace words in a sentence with their

synonyms via a synonym table, have attracted considerable attention. The reason is they can preserve the syntactic and semantics of the original input to the most considerable extent and are very hard to discern, even from a human's perspective [8].

Document authors compete for more favorable positions in rankings [12]. The behavior is often referred to as search engine optimization (SEO) [15], which includes "illegitimate" black hat SEO [15] (e.g., web spam [3]) and "legitimate" white hat SEO [15]. Early work proposed methods to detect black-hat SEO behavior [35, 38]. Later, Goren et al. [12] proposed to attack LTR by replacing a passage in a document with other passages to promote its ranking. Recently, Wu et al. [48] proposed word substitution ranking attacks (WSRAs) to promote the ranking of a document, which easily escapes the detection of traditional anti-spamming methods. In this paper, we focus on defending against WSRAs.

2.3 Defense Methods

To defend against adversarial attacks, many defense methods have been proposed to make models more robust. These approaches can be classified into *empirical* defenses and *certified* defenses. Empirical defenses attempt to make models empirically robust to known adversarial attacks; this has been extensively explored in image [29, 47] and text classification [18, 52]. Data augmentation [17, 42] is a representative empirical defense by augmenting adversarial examples with the original training data. Since empirical defenses are only effective for certain attacks rather than all attacks, a competition emerges between adversarial attacks and defense methods.

To solve the attack-defense dilemma, researchers resort to certified defenses to make models provably robust to certain kinds of adversarial perturbations. Jia et al. [18] and Huang et al. [16] first proposed to certify the robustness to adversarial word substitutions by leveraging Interval Bound Propagation (IBP [9]) in NLP. These IBP-based methods are limited to continuous word embeddings and are not applicable to subword-level models like BERT. Ye et al. [52] recently adopted randomized smoothing to certify the robustness to word substitution attacks, by turning the original classifier into a smoothed classifier by adding noise to the input. The final class prediction of the smoothed classifier is decided by majority voting over the noised inputs. Randomized smoothing is the only certification method that scales up to large-scale neural networks like BERT [5] and provides tight bounds on large datasets. But existing certified defenses are limited to simple classification scenarios and NRMs are less well studied. Therefore, in this work, we develop a certified defense method for NRMs based on randomized smoothing.

3 CERTIFIED TOP-K ROBUSTNESS IN IR

We first introduce the WSRA attack we consider in this paper. Then, we introduce the definition of our proposed notion of Certified Top-K Robustness for ranking models to such attacks.

3.1 Word Substitution Ranking Attack

Attacks in Web Search. The web search eco-system is, perhaps, the largest-scale adversarial setting in which search methods operate [13]. For many queries in the web retrieval setting there exists an on-going ranking competition, i.e., many web document authors manipulate their documents deliberately to promote them in rankings [12]. This practice is often referred to as search engine optimization (SEO) [15]. The consequences of SEO are that the quality of search results may rapidly decrease since many irrelevant documents are ranked higher than they deserve.

Very recently, a typical black-box word substitution ranking attack (WSRA) [48] was proposed to simulate such real-world ranking competitions. Specifically, WSRA could successfully attack NRMs by generating human-imperceptible adversarial documents for rank promotion. The synonymous word substitution it employs could maximally maintain the naturalness and semantic similarity of the original document, making it easy for the generated adversarial documents to evade spam detection. Due to the popularity of NRMs and the challenges of defending against human-imperceptible perturbations, we focus on WSRA attacks and design a corresponding defense in this paper.

Notation. In ad-hoc retrieval, given a query q and a set of document candidates $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ selected from a collection C, a ranking model f aims to predict the relevance score $\{f(q, d_n) : n = 1, 2, \ldots, N\}$ between every pair of query q and candidate document for ranking the whole candidate set. For example, f outputs the ranked list $L = [d_N, d_{N-1}, \ldots, d_1]$ if it determines $f(q, d_N) > f(q, d_{N-1}) \cdots > f(q, d_1)$. In this paper, we assume the ranking score $f(q, d_n)$ is the probability of relevance from 0 to 1 [6], which can be easily achieved by adding a sigmoid operation on the output given by the ranking model.

In WSRA, an attacker replaces the important words in the document with their synonyms by maximizing the adversarial ranking loss to promote the target document in rankings. The number of important words is a hyper-parameter. Specifically, for any word w, we consider a pre-defined synonym set S_w containing the synonyms of w (including w itself). Following Ye et al. [52], we assume the synonymous relation is symmetric, that is, w is in the synonym set of all its synonyms. The synonym set S_w can be built based on GLOVE [37].

Definition 3.1. (δ-Word Substitution Ranking Attack). For an input document $d = \{w_1, w_2, \ldots, w_M\} \in \mathcal{D}$, a *$\delta$-word substitution ranking attack* constructs an adversarial document $d' = (w'_1, w'_2, \ldots, w'_M)$ by perturbing at most $\delta \cdot M$ ($\delta \leq 1$) words in d to any of their synonyms $w'_m \in S_{w_m}$. We denote the candidate set of adversarial documents as S_d, i.e.,

$$S_d := \{d' : \|d' - d\|_0 / \|d\| \leq \delta\},$$

where $\|d' - d\|_0 := \sum_{m=1}^{M} \mathbb{I}\{w'_m \neq w_m\}$ is the Hamming distance, with $\mathbb{I}\{\cdot\}$ the indicator function. $\|d\|$ denotes the number of words in the document d and $w'_m \in S_{w_m}$. Ideally, all $d' \in S_d$ have the same semantic meaning as d for human judges, but their ranks may be promoted by the ranking model f. The goal of the attacker is to find $d' \in S_d$ such that $f(q, d') > f(q, d)$. Note that we do not attack the documents ranked from 1 to K, since there is no need to attack user's top search results.

3.2 Definition of Certified Top-K Robustness

Certified Top-K Robustness. In general, a model is said to be *certified robust* if an attack is guaranteed to fail, no matter how the attacker manipulates the input [52]. In a real web search scenario, it is known that users usually care much more about the top ranking results than others [33]. For example, the traffic and click-through

rate (CTR) both fall off as users work their way down the search results in major search engines: while the first and second search results may achieve 36.4% and 12.5% CTR, the 10th search result may achieve a CTR of only 2.2%.[1] Moreover, many widely-used ranking metrics [2, 40] focus on the top-K ranking results, e.g., MRR@K and nDCG@K.

Therefore, protecting the results ranked at the top positions is of great importance [33, 50], not only for real-world applications, but also for the robustness guarantee of widely-used ranking metrics. Inspired by this, we define the *Certified Top-K Robustness* of ranking models in IR, where a ranking model f is said to be certified robust at the ranked list L if it is guaranteed that the documents ranked after top K will not be attacked to be ranked into top K in L. Since we focus on the WSRAs in this work, based on this basic definition, we further define *Certified Top-K Robustness to WSRA*.

Definition 3.2. (Certified Top-K Robustness to WSRA). Formally, a ranking model f is said to be *Certified Top-K Robust* against WSRAs on the ranked list L_q with respect to a query q if it can keep all the document $d \in L_q[K+1:]$ away from the top-K for all the possible δ-word substitution ranking attacks (as defined in Definition 3.1), i.e.,

$$Rank_{L_q}(f(q,d')) > K, \text{for all } d \in L_q[K+1:] \text{ and any } d' \in S_d, \quad (1)$$

where $Rank_{L_q}(f(q,d'))$ denotes the rank position of the adversarial document d' in L_q given by the ranking model f. It is highly challenging to judge if f is certified robust since all the candidate adversarial documents in S_d should be checked and the size of possible perturbations grows exponentially with δ. Following existing work [18, 52], we mainly consider the worst case when $\delta = 1$, which is the most challenging case.

4 OUR CERTIFIED DEFENSE METHOD

Based on the definition of certified top-K robustness to WSRA, we introduce a novel **Cert**ified **D**efense method for **R**anking models (CertDR) to enhance the certified robustness. We first introduce a randomized smoothing function for ranking and how to use it to certify the robustness theoretically. Then, we propose a practical certified defense algorithm for ranking models. Proofs are at the end of this section.

4.1 Randomized Smoothing Function for Ranking

To circumvent the computationally expensive combinatorial optimization (e.g., enumerating all the candidate adversarial documents in S_d), we borrow the idea from the randomized smoothing technique [5, 52], which could provably defend against the adversarial attacks by leveraging the voting of randomly perturbed samples derived from the original input. We target to replace the ranking model f with a smoothed ranking model $\bar{f}$ for which it is easier to verify the certified robustness.

Specifically, we construct the smoothed ranker $\bar{f}$ by averaging the output ranking scores of a set of randomly perturbed documents based on random perturbations, i.e.,

$$\bar{f}(q,d) = \mathbb{E}_{\mathbf{R}\sim\Pi_d} f(q,\mathbf{R}),$$

where R is a randomly perturbed document and Π_d is the corresponding probability distribution that prescribes a random perturbation around d. In our work, we define Π_d to be the uniform distribution on a set of random word substitutions following [52].

In previous classification tasks [5, 52], the output of the smoothed classifier is the class with the largest probability "voting" by all randomly perturbed inputs. Different from these works, we compute the output of the smoothed ranker by averaging the ranking scores of all randomly perturbed documents originated from the d, which is more suitable for the ranking problem. In this way, we could obtain the ranked list L_q^s based on the averaged scores produced by the smoothed ranker $\bar{f}$. Here, we leave the query q free from attack. In the future work, we would like to explore the defense against query attacks by focusing on q in this formulation.

To obtain random perturbations in defense methods effectively, we propose to build a perturbation set T_w for each word w. Specifically, we construct T_w from the synonym set S_w used in the attack method, i.e., WSRA in this work, by choosing the top J nearest neighbors via the cosine similarity of GLOVE vectors. Then, for a document $d = (w_1, w_2, \ldots, w_M)$, we define the perturbation distribution Π_d by perturbing each word w_i in d to a word in its perturbation set T_{w_i} randomly and independently, i.e.,

$$\Pi_d(\mathbf{R}) = \prod_{i=1}^{M} \frac{\mathbb{I}\{r_i \in T_{w_i}\}}{|T_{w_i}|},$$

where $\mathbf{R} = (r_1, \ldots, r_M)$ is the perturbed document and $|T_{w_i}|$ denotes the size of T_{w_i}.

4.2 Certifying Smoothed Ranking Models

Given the smoothed ranking model $\bar{f}$, in this section, we introduce how to certify its top-K robustness. For all the documents in $L_q^s[K+1:]$, if their adversarial documents could achieve lower scores than the document d_K ranked at the position K, we think these documents $L_q^s[K+1:]$ cannot be attacked into top K. Formally, the condition that $\bar{f}$ is certified top-K robust on L_q^s can be defined as,

$$\max_{d\in L_q^s[K+1:]} \max_{d'\in S_d} \bar{f}(q,d') < \bar{f}(q,d_K). \quad (2)$$

In general, there are two difficulties to complete the above certification case by case, i.e., the inner maximum and the outer maximum.

Inner Maximum. The first difficulty is to exam all candidate adversarial documents in S_d for the inner maximum, where the computation cost grows exponentially with the attacked word number δM. We address this problem in the following Theorem 4.1.

Theorem 4.1 (Certified Upper Bound). *Assume that the perturbation set T_w is constructed such that $|T_w| = |T_{w'}|$ for every word w and its synonym $w' \in S_w$. Define*

$$o_w = \min_{w'\in S_w} \frac{|T_w \cap T_{w'}|}{|T_w|},$$

where o_w indicates the overlap between the two different perturbation sets. For a given document $d = (w_1, \ldots, w_M)$, we sort the words according to o_w, such that $o_{w_{i_1}} \le o_{w_{i_2}} \le \cdots \le o_{w_{i_M}}$. Then

$$\max_{d'\in S_d} \bar{f}(q,d') \le \min(\bar{f}(q,d) + o_d, 1),$$

where $o_d := 1 - \prod_{j=1}^{E} o_{w_{i_j}}$.

[1] https://www.smartinsights.com/search-engine-optimisation-seo/seo-analytics/the-number-one-spot-how-to-use-the-new-search-curve-ctr-data/

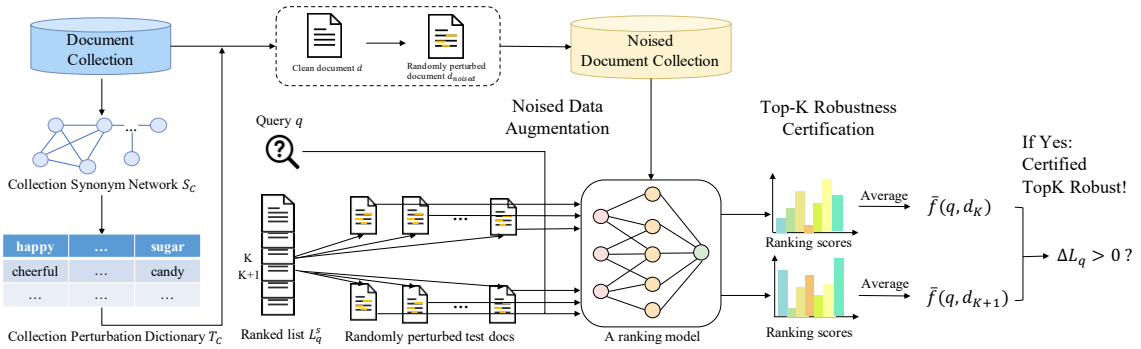

Figure 2: The overall architecture of the proposed practical certified defense method. We first generate the collection perturbation dictionary T_C over the whole document collection. Then, we train the original ranking model by a noised data augmentation strategy to increase the robustness of $\bar{f}$. Finally, we estimate the $\bar{f}(q, d_K)$ and $\bar{f}(q, d_{K+1})$ by Monte Carlo estimation. The criterion ΔL_q is computed, where if $\Delta L_q > 0$, we can certify that $\bar{f}$ is top-K robust at the ranked list L_q^s.

The idea is that for any adversarial document $d' \in S_d$, the upper bound of $\bar{f}(q, d')$ can be bounded by its original document ranking score $\bar{f}(q, d)$ with randomized smoothing. As a result, this theorem avoids the difficult adversarial optimization of $\bar{f}(q, d')$ on $d' \in S_d$, and only needs to evaluate $\bar{f}(q, d)$ at the original document d. Note that the difference between our Theorem 4.1 with Theorem 1 in [52] is that we extend the classification prediction in Theorem 1 to ranking scores between queries and documents, and prove the theorem in an upper bound situation, which has not been provided by Ye et al. [52]. The proof is provided in Section 4.4.1. Besides, the upper bound in Theorem 4.1 is sufficiently tight, which is shown in Section 4.4.2.

Outer Maximum. The second difficulty is to exam all documents in $L_q^s[K+1:]$ for the outer maximum, where the computational costs grow linearly with the list length N. We address the outer maximum in the following. To simplify our notation, we define $A_L = \bar{f}(q, d_K) - \max_{d \in L_q^s[K+1:]} \max_{d' \in S_d} \bar{f}(q, d')$, where Eq. (2) is equivalent to $A_L > 0$. By applying Theorem 4.1 to Eq. (2), we have

$$\begin{aligned} A_L &\geq \bar{f}(q, d_K) - \max_{d \in L_q^s[K+1:]} (\bar{f}(q, d) + o_d) \\ &\geq \bar{f}(q, d_K) - \bar{f}(q, d_{K+1}) - \max_{d \in L_q^s[K+1:]} o_d, \end{aligned}$$

where d_{K+1} denotes the document which is ranked at the position $K+1$. The proof is achieved by utilizing the ranking property that $\bar{f}(q, d_1) > \bar{f}(q, d_2) > \cdots$. The idea is that we can compute A_L (in other words, certifying $\bar{f}$) by comparing the ranking scores of documents ranked at K and $K+1$. Note that the computational cost of $\max_{d \in L_q^s[K+1:]} o_d$ is negligible.

Based on the above solutions of the inner and outer maximum in Eq. (2), it is possible to introduce a certification criterion for checking the certified top-K robustness for ranking models.

Proposition 4.1. *For a ranked list L_q^s with respect to a query q, under the condition of Theorem 4.1, we can certify that $Rank_{L_q^s}(\bar{f}(q, d')) > K$, for all $d \in L_q^s[K+1:]$ and any $d' \in S_d$ if*

$$\Delta L_q \stackrel{def}{=} \bar{f}(q, d_K) - \bar{f}(q, d_{K+1}) - \max_{d \in L_q^s[K+1:]} o_d > 0, \tag{3}$$

where ΔL_q can be estimated by Monte Carlo estimation as we show in the next section. We can certify whether the ranking model is top-K robust on the ranked list L_q^s by simply checking ΔL_q. If ΔL_q is positive, the model is certified top-K robust.

4.3 Practical Certified Defense Algorithm

Based on the above theoretical analysis, we now present a practical certified defense algorithm for ranking models. Formally, we write S_C for the synonym dictionary of the collection. Specifically, S_C contains the synonym set S_w for all words in the collection C and is often presented as a synonym network [52]. Similar to the process of obtaining T_w from S_w, we achieve the collection perturbation dictionary T_C by keeping the top J nearest neighbors in S_C for each word. The overall architecture is shown in Figure 2.

The certified defense algorithm contains two key steps, i.e., noised data augmentation and Top-K Robustness Certification. We describe the two steps in the following.

Noise Data Augmentation Strategy. The robustness certification holds regardless of how the original ranker f is trained. However, to rank the document d with respect to the q correctly and robustly by $\bar{f}$, it is expected that f properly ranks the perturbed document R (recall that $R \sim \Pi_d$) such that it is close to the rank position of the original document d. That is, the noise of R should have little effect on the ranking, making the ensembled ranking score $\bar{f}(q, d)$ close to the original ranking score $f(q, d)$. However, if f is trained via standard supervised learning without any noised documents, it will not necessarily learn how to rank R properly.

Inspired by previous works [23, 52, 55], we introduce a noise data augmentation strategy for ranking. Specifically, we first generate a perturbed document d_{noised} for each d in the collection C. The perturbation is achieved by randomly sampling every word from d using the perturbation distribution Π_d. Then, we train the original ranker f using the training triples equipped with the noised documents via the following objective:

$$L_{ndat} = \max(0, 1 - f(q, d^+_{noised}) + f(q, d^-_{noised})), \tag{4}$$

where $d^+_{noised}/d^-_{noised}$ is the perturbed document from d^+/d^-. And d^+/d^- denotes the positive/negative document in original training triples. Then, we can obtain a better smoothed ranker $\bar{f}$ by Monte Carlo estimation in the following.

Top-K Robustness Certification. In theory, since the perturbation space can be extremely large, it is impossible to exactly obtain

the prediction of $\bar{f}$ at each (q, d). Therefore, based on the ranker f obtained from the noised data augmentation strategy, we estimate $\bar{f}(q, d_K)$ and $\bar{f}(q, d_{K+1})$ by Monte Carlo estimation. Take $\bar{f}(q, d_K)$ as an example, we can estimate it like

$$\bar{f}(q, d_K) = \mathbb{E}_{\mathbf{R}_K \sim d_K} f(q, d_K) \approx \frac{1}{n} \sum_{i=1}^{n} f(q, \mathbf{R}_K^{(i)}),$$

where $\mathbf{R}_K^{(i)}$ are i.i.d. samples from Π_{d_K} and thus ΔL_q can be approximated accordingly. We can construct rigorous statistical procedures to reject the null hypothesis that $\bar{f}$ is not certified robust at L_q (e.g., $\Delta L_q < 0$) with a given significance level (e.g., 5%) following [52].

Finally, if $\Delta L_q > 0$, then $\bar{f}$ is certified top-K robust at the ranked list L_q^s. Otherwise, we will judge it is not certified top-K robust at the L_q^s. We can see that our practical certified defense algorithm could be achieved by assembling the ranking outputs and that it does not require any further information about the ranking models. Thus, it can be applied to any ranking model.

4.4 Proofs

Here we provide all the necessary proofs of Theorem 4.1 and the tightness of the bound in Theorem 4.1 with its proof. The upper bound of Theorem 4.1 is achieved by introducing an auxiliary function cluster based on the relevance between the query and document, and solving the constraint optimization problem by Lagrange and properties of randomly perturbed sets. Tightness is proved by constructing the randomized smoothing ranker that satisfies the desired property we want.

4.4.1 Proof of Theorem 4.1. Our goal is to calculate the upper bound $\max_{d' \in S_d} \bar{f}(q, d')$. The key idea is to frame the computation of the upper bound into a variational optimization.

Lemma 4.1. *Define $\mathcal{G}_{[0,1]}$ to be the set of all bounded functions mapping from $\mathcal{Q} \times \mathcal{D}$ to [0, 1]. For any $g \in \mathcal{G}_{[0,1]}$, define*

$$\Pi_d[g] = \mathbb{E}_{\mathbf{R} \sim \Pi_d}[g(q, \mathbf{R})].$$

Then we have for any $\mathbf{R}$,

$$\begin{aligned} \max_{d' \in S_d} \bar{f}(q, d') &\le \max_{d' \in S_d} \max_{g \in \mathcal{G}_{[0,1]}} \{\Pi_{d'}[g] \; s.t. \; \Pi_d[g] = \bar{f}(q, d)\} \\ &:= \bar{f}_{up}(q, d') \end{aligned}$$

Proof of Lemma 4.1. Define $g_0(q, d) = f(q, d)$. Then we have

$$\bar{f}(q, d) = \mathbb{E}_{\mathbf{R} \sim \Pi_d}[f(q, \mathbf{R})] = \Pi_d[g_0].$$

Therefore, g_0 satisfies the constraints in the optimization, which makes it obvious that

$$\bar{f}(q, d') = \Pi_{d'}[g_0] \le \max_{g \in \mathcal{G}_{[0,1]}} \{\Pi_{d'}[g] \; s.t. \; \Pi_d[g] = \bar{f}(q, d)\}.$$

Taking $\max_{d' \in S_d}$ on both sides yields the upper bound and thus the problem reduces to deriving bounds for the optimization problems.

Theorem 4.2. *Under the assumption of Theorem 4.1, for the optimization problem in Lemma 4.1, we have*

$$\bar{f}_{up}(q, d') \le \min(\bar{f}(q, d) + o_d, 1),$$

where o_d is the quantity defined in Theorem 4.1.

Proof of Theorem 4.2. For notation, we denote $p = \bar{f}(q, d)$. Applying the Lagrange multiplier to the constraint optimization problem and exchanging the min and max, we have

$$\begin{aligned} \bar{f}_{up}(q, d') &= \max_{d' \in S_d} \max_{g \in \mathcal{G}_{[0,1]}} \{\Pi_{d'}[g] \; s.t. \; \Pi_d[g] = \bar{f}(q, d)\} \\ &= - \min_{d' \in S_d} \min_{g \in \mathcal{G}_{[0,1]}} \{-\Pi_{d'}[g] \; s.t. \; \Pi_d[g] = \bar{f}(q, d)\} \\ &\le - \min_{d' \in S_d} \min_{g \in \mathcal{G}_{[0,1]}} \max_{\lambda \in \mathcal{R}} \lambda \Pi_d[g] - \Pi_{d'}[g] - \lambda p \\ &= \max_{d' \in S_d} \min_{\lambda \in \mathcal{R}} \max_{g \in \mathcal{G}_{[0,1]}} \Pi_{d'}[g] - \lambda \Pi_d[g] + \lambda p \\ &= \min_{\lambda \in \mathcal{R}} \lambda p + \max_{d' \in S_d} \max_{g \in \mathcal{G}_{[0,1]}} \int g(\mathbf{R})(d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R})). \end{aligned}$$

Note that

$$\max_{g \in \mathcal{G}_{[0,1]}} \int g(\mathbf{R})(d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R})) = \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+,$$

which is achieved by setting

$$g(\mathbf{R}) = \mathbb{I}\{d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}) \ge 0\},$$

where $(a)_+ = \max(a, 0)$ and $\mathbb{I}\{\cdot\}$ denotes the indicator function. Thus we have,

$$\bar{f}_{up}(q, d') = \min_{\lambda \in \mathcal{R}} \lambda p + \max_{d' \in S_d} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+.$$

For any $\lambda < 0$, we can show that

$$\begin{aligned} &\lambda p + \max_{d' \in S_d} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+ \\ &\qquad = \lambda p + \max_{d' \in S_d} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R})) \\ &\qquad = \lambda(p - 1) + 1 > 0(p - 1) + 1 = 1, \end{aligned}$$

which contradicts $\bar{f}(q, d') \le 1$. This implies that the minimum of $\lambda p + \max_{d' \in S_d} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+$ must be achieved at $\lambda \ge 0$. Thus we have

$$\bar{f}_{up}(q, d') = \min_{\lambda \ge 0} \lambda p + \max_{d' \in S_d} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+.$$

Now we calculate $\int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+$.

Lemma 4.2. *Given the words w, w', we write $n_w = |T_w|, n_{w'} = |T_w'|$, and $n_{w,w'} = |T_w \cap T_{w'}|$. We have the following identify*

$$\begin{aligned} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+ &= 1 - \prod_{j \in [M], w_j \ne w_j'} \frac{n_{w_j, w_j'}}{n_{w_j'}} \\ &+ \left(\prod_{j \in [M], w_j \ne w_j'} \frac{n_{w_j, w_j'}}{n_{w_j'}} \right) \left(1 - \lambda \prod_{j \in [M], w_j \ne w_j'} \frac{n_{w_j'}}{n_{w_j}} \right)_+ . \end{aligned}$$

As a result, under the assumption that $n_w = |T_w| = |T_{w'}| = n_{w'}$ for every word w and its synonym $w' \in S_w$, we have

$$\begin{aligned} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+ &= 1 - \prod_{j \in [M], w_j \ne w_j'} \frac{n_{w_j, w_j'}}{n_{w_j'}} \\ &+ \left(\prod_{j \in [M], w_j \ne w_j'} \frac{n_{w_j, w_j'}}{n_{w_j'}} \right) (1 - \lambda)_+ . \end{aligned}$$

Now we need to solve the optimization of $\max_{d' \in S_d} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+$.

Lemma 4.3. *For any word w, define $\tilde{w}^* = \arg\min_{w' \in S_d} n_{w,w'}/n_w$. For a given document $d = (w_1, \ldots, w_M)$, we define an ordering of the words $w_{p_1}, \ldots, w_{p_W}$ such that $n_{w_{p_i}, \tilde{w}^*_{p_i}}/n_{w_{p_i}} \leq n_{w_{p_j}, \tilde{w}^*_{p_j}}/n_{w_{p_j}}$ for any $i \leq j$. For a given d and $E = \delta M$, we define an adversarial perturbed document $d^* = (w^*_1, \ldots, w^*_M)$, where*

$$w^*_i = \begin{cases} \tilde{w}^*, & \text{if } i \in [w_1, \ldots, w_E] \\ w_i, & \text{if } i \notin [w_1, \ldots, w_E] \end{cases}$$

Then for any $\lambda > 0$, we have that d^* is the optimal solution of $\max_{d' \in S_d} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+$, that is,

$$\max_{d' \in S_d} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+ = \int (d\Pi_{d^*}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+.$$

Now by Lemma 4.3, the upper bound becomes

$$\begin{aligned} \bar{f}_{up}(q, d') &= \min_{\lambda \geq 0} \lambda p + \max_{d' \in S_d} \int (d\Pi_{d'}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+ \\ &= \min_{\lambda \geq 0} \lambda p + \int (d\Pi_{d^*}(\mathbf{R}) - \lambda d\Pi_d(\mathbf{R}))_+ \\ &= \min_{\lambda \geq 0} (\lambda p + o_d + (1 - o_d)(1 - \lambda)_+) \\ &= \min(p + o_d, 1), \end{aligned} \tag{5}$$

where o_d is consistent with the definition in Theorem 4.1:

$$o_d = 1 - \prod_{j \in [M], w_j \neq \tilde{w}^*_j} \frac{n_{w_j, \tilde{w}^*_j}}{n_{w_j}} = 1 - \prod_{j=1}^{E} o_{w_{i_j}}.$$

Here, Eq. (5) is calculated using the assumption of Theorem 4.1. The optimization of $\min_{\lambda \geq 0}$ in (5) is an elementary step: if $p + o_d > 1$, we have $\lambda^* = 0$ with solution 1; if $p + o_d \leq 1$, we have $\lambda^* = 1$ with solution $p + o_d$. For the proof of Lemma 4.2 and 4.3, we refer readers to [52, Lemma 2 and 3].

4.4.2 Tightness. Whether the bound in Theorem 4.1 is sufficiently tight is of great importance. In the following, we provide a theorem to state its tightness.

Theorem 4.3 (Tightness). *Assume the conditions of Theorem 4.1 hold. For a ranking model f that maps $Q \times D$ to [0, 1], there exists a model f_* such that its related smoothed ranker $\bar{f}_*$ satisfies*

$$\bar{f}_*(q, d) = \bar{f}(q, d),$$

and

$$\max_{d' \in S_d} \bar{f}_*(q, d') = \min(\bar{f}_*(q, d) + o_d, 1),$$

where o_d is defined in Theorem 4.1.

As shown in Theorem 4.3, the upper bound in Theorem 4.1 is tight and can not be further improved if we do not know any other structural information about f. In the following, we provide the proof of Theorem 4.3.

Proof of Tightness. We denote $\bar{f}(q, d) = p_r$ in this proof for simplicity. The d^* below is the optimal adversarial document defined in the proof of Lemma 4.3. Note that $o_d = |T_d - T_{d^*}|/|T_d|$ and $1 - o_d = |T_d \cap T_{d^*}|/|T_d|$ as defined in Theorem 4.1. Our proof is based on constructing a randomized smoothing ranker that satisfies the desired property we want to prove.

Case 1 $p_r \leq 1 - o_d$**.** Note that in this case $|T_d \cap T_{d^*}| = 1 - o_d \geq p_r$. Therefore, we can choose set U such that $U \subseteq T_d \cap T_{d^*}$ and $|U|/|T_d| = p_r$. We define the ranker:

$$f_*(q, \mathbf{R}) = \begin{cases} 1, & \text{if } \mathbf{R} \in U \cup T_{d^*} \\ 0, & \text{otherwise} \end{cases}$$

Case 2 $p_r > 1 - o_d$**.** We choose set U such that $U \subseteq T_d \cap T_{d^*}$ and $|U|/|T_d| = p_r$. We define the ranker

$$f_*(q, \mathbf{R}) = \begin{cases} 1, & \text{if } \mathbf{R} \in U \cup (T_{d^*} - T_d) \\ 0, & \text{otherwise} \end{cases}$$

It can easily be verified that for each case, the defined ranker satisfies all the conditions in Theorem 4.3. This indicates the bound can be achieved by learning a gold ranker that can judge some specific documents as relevant and others as irrelevant for the query.

5 EXPERIMENTAL SETUP

5.1 Datasets

To evaluate the effectiveness of our proposed methods, we conduct experiments on two representative web search benchmark datasets.

- **MS MARCO Document Ranking dataset** [32] (MS-MARCO-Doc) is a large-scale benchmark dataset for web document retrieval, with about 3.21M web documents and 0.37M training queries. The average length of the document is about 1129.
- **MS MARCO Passage Ranking dataset** [32] (MS-MARCO-Pas) is a large-scale benchmark dataset for passage retrieval, with about 8.84M passages from web pages and 0.5M training queries. The average length of the passage is about 58.

5.2 Baselines

Our CertDR can certify the top-K robustness of different ranking models. We first compare the certified top-K robustness among different ranking models (i.e, BM25 [43], Duet [30] and BERT [7]) under CertDR. Then, since the defense methods for NRMs have not been well explored yet, we adopt the representative empirical defense, i.e., **Data Augmentation (DA)**, in text classification task [17, 42], for NRMs as a baseline. Specifically, for each training document d, we augment the collection with 2 new documents $\tilde{d}$ by sampling $\tilde{d}$ uniformly from S_d, then train on the normal hinge loss following [18]. We do not use adversarial training [11] here because it would require running an adversarial search procedure at each training step, which would be prohibitively slow.

5.3 Evaluation Metrics

We evaluate the robustness to all WSRAs of models. We propose a metric to directly measure the **Certified Robust Query (CRQ)** percentage, the percentage of test queries for which the model is certified robust at the query q if all the documents out of top-K are not attacked into the top-K. Evaluating this exactly involves enumerating exponentially many perturbations, which is intractable (Section 4.2). Instead, we evaluate the *CRQ* under randomized smoothing, i.e.,

$$CRQ = \frac{\sum_{q \in Q} \mathbb{I}\{\Delta L_q > 0\}}{|Q|},$$

where ΔL_q is the criterion mentioned in Section 4.2. The ranking model is more certified robust with a higher CRQ (%).

To compare the defense ability of CertDR with empirical defense methods, we also leverage two metrics, i.e., success rate and conditional success rate. **Success Rate (SR) [48]** evaluates the percentage of the after-attack documents that are ranked higher than original documents. The robustness of a ranking model is better with a lower SR (%).

Inspired by CondAcc [46], which enables the comparison certified and empirical defense, we introduce **Conditional Success Rate (CondSR)**. CondSR evaluates whether the rankings of the adversarial documents in an attacked ranked list indeed cannot be improved when its counterpart clean ranked list is certified robust:

$$\text{CondSR} = \frac{\sum_{q\in Q}\mathbb{I}\{\Delta L_q > 0\}\frac{1}{N_q}\sum_{i=1}^{N_q}\mathbb{I}\{Rank_L(d_i+p) < Rank_L(d_i)\}}{\sum_{q\in Q}\mathbb{I}\{\Delta L_q > 0\}}.$$

While CRQ is evaluated on clean ranked lists to show certified robustness, CondSR is tested on attacked ranked lists to show the empirical robustness of models on these certified ranked lists. The robustness of a ranking model is better with a lower CondSR (%).

5.4 Implementation Details

We implement ranking models following previous work [6, 49]. For the MS-MARCO-Doc collection, we use the official top 100 (i.e., $N = 100$) ranked documents retrieved by the QL model. For the MS-MARCO-Pas, initial retrieval is performed using the Anserini toolkit [51] with the BM25 model to obtain the top 100 ranked passages. We evaluate all ranking models on 200 queries (i.e., $|Q| = 200$) randomly sampled from the dev set of two datasets following [48].

For the Monte Carlo estimation of ΔL_q, we use 1,000 random perturbed documents to accept $\Delta L_q > 0$ with probability of at least 0.95. The corresponding estimation error is 0.086 and is considered during the estimation following Ye et al. [52]. Further implementation details and the code can be found online.[2]

6 EXPERIMENTAL RESULTS

We evaluate our defense method to address the following research questions: (**RQ1**) What is the certified robustness among different ranking models via CertDR? (**RQ2**) How does the randomized smoothed ranker perform compared with the original ranker? (**RQ3**) How does K affect certified top-K robustness? (**RQ4**) How does CertDR perform compared with empirical defense baselines?

6.1 Certified Top-K Robustness of Different Ranking Models

To answer **RQ1**, we analyze certified top-K robustness of different ranking models using CertDR on MS-MARCO-Doc and MS-MARCO-Pas. See Figure 3. We find that (i) Overall, the certified robustness of the ranking model is lower than that of the text classification models [18, 52], indicating that ranking models are vulnerable to adversarial attacks. There are two potential reasons: First, the text ranking task itself needs to model cross-document interactions to capture query-document relevance, which is more complex than classifying a single sentence independently as in text classification. Second, certified top-K robustness imposes requirements on the ranked list, which is demanding than the point-wise classification scenario. (ii) Pre-trained model BERT generally outperforms other models, indicating that BERT is more certified robust than other ranking models. The reason might be that pre-training on a large text corpus can improve the out-of-distribution generalizability to adversarial examples attacked by synonyms substitution. Therefore, it is worthwhile to leverage pre-training techniques to enhance the robustness of NRMs in the future. (iii) BM25 is less certified robust than Duet and BERT on MS-MARCO-Doc when K is small, while it is more certified robust than Duet on MS-MARCO-Pas for all Ks. BM25 depends on exact matching signals between the query and document; therefore, a possible explanation is that there are fewer options of attacked words in the short passage than the long document, contributing to the robustness on short texts. Besides, we leave the analysis of different Ks to Section 6.3.

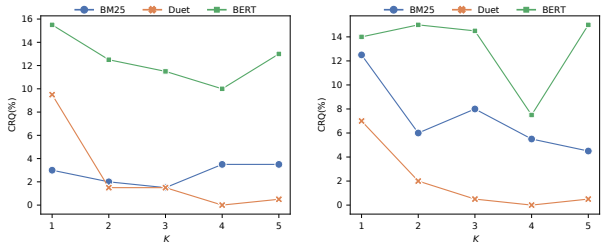

Figure 3: Certified top-K robustness of different ranking models in terms of CRQ (%) on MS-MARCO-Doc (a) and MS-MARCO-Pas (b).

Table 1: Comparing the ranking performance between the original and randomized smoothed ranker in terms of the MRR@10 and MRR@100 on MS-MARCO-Doc. * denotes significant degradation w.r.t. the randomized smoothed ranker w/o noise data augmentation (p-value<0.05).

Method	MRR@10	MRR@100
Original f	0.4428*	0.4470*
Smoothed $\bar{f}$ w/o noise data aug	0.2259	0.2416
Smoothed $\bar{f}$	0.3635*	0.3722*

6.2 Smoothed Ranker vs. Original Ranker

To answer **RQ2**, we compare the ranking performance of the randomized smoothed ranker with the original ranker. We select BERT as the original ranker and conduct experiments on MS-MARCO-Doc. We also show the ranking performance of the randomized smoothed ranker without the noised data augmentation.

From Table 1, we observe that: (i) The ranking performance of smoothed ranker without noised data augmentation drops dramatically (e.g., 0.2259 vs. 0.4428 in terms of MRR@100). The reason might be that the smoothed ranker ranks documents based on the ensemble ranking scores of perturbed documents, which are far away from the original documents. (ii) By applying a noised data augmentation strategy, the ranking performance of the smoothed ranker improves significantly and becomes closer to the original ranker. The reason might be that the augmented training documents are generated from the same perturbation distribution with

[2] https://github.com/ict-bigdatalab/CertDR/

the perturbed documents, which helps the smoothed ranker learn to rank the perturbed documents properly. (iii) The smoothed ranker has a moderate drop in terms of MRR@100 compared with the original ranker with normal training (0.3722 vs. 0.4470). Similar drops on clean acc (accuracy on clean examples) are also seen for robust models in previous work [18, 31]. Future work could explore how to achieve the trade-off between clean and robust performance.

Table 2: The CRQ (%) of different ranking models with different K on MS-MARCO-Doc and MS-MARCO-Pas.

K	MS-MARCO-Doc			MS-MARCO-Pas		
	BM25	Duet	BERT	BM25	Duet	BERT
1	3.0	9.5	15.5	12.5	7.0	14.0
2	2.0	1.5	12.5	6.0	2.0	15.0
3	1.5	1.5	11.5	8.0	0.5	14.5
4	3.5	0	10.0	5.5	0	7.5
5	3.5	0.5	13.0	4.5	0.5	15.0
10	1.5	0	3.0	2.5	0	9.5
20	1.5	0	0	0.5	0	3.0
30	0.5	0	0	0	0	1.5
40	0.5	0	0	1.5	0	0
50	0.5	0	0	0.5	0	0
60	0.5	0	0	0	0	0
70	0.5	0	0	1.0	0	0
80	0.5	0	0	0	0	0
90	0.5	0	0	0.5	0	0

6.3 Analysis of the Effect of K

To answer **RQ3**, we analyze the effect of K for CertDR when we certify the top-K robustness. Specifically, we analyze the ranking performance of different ranking models in terms of CRQ, and set K to 14 different values. As shown in Table 2, we can find that: (i) Overall, the model becomes less certified top-K robust with the increase of K on both datasets. Intuitively, it is more difficult to attack a document to a higher rank position than a lower rank position. (ii) However, an interesting finding is that the certified top-K robustness with a larger K is even greater than a smaller K in a certain range. For example, the CRQ of BERT is 15.0 with $K = 5$ while 7.5 with $K = 4$ on the MS-MARCO-Pas dataset. By conducting further analysis, we find that although documents ranked out of top 5 are not easily attacked into the top 5, the 5-th document could be attacked into the top 4 easily. (iii) While the CRQ of NRMs reduces to 0 after a point (e.g, the CRQ of Duet reduces to 0 after K=10), it is interesting to find that the CRQ of BM25 remains at a low positive value when K is very large (e.g., CRQ of BM25 remains at 0.5 when K=30 to 90 on MS-MARCO-Doc). The reason may be that BM25 relies on the statistical features, which is more robust than word embeddings of NRMs against adversarial attacks. This is consistent with the findings in [49].

6.4 Comparison with Empirical Defenses

Based on the above analysis of certified robustness of different models, we further compare CertDR with baseline empirical defense methods (i.e., DA) following [46]. The WSRA is conducted by PRADA [48], and we set $K = 10$ and 5 for CertDR on MS-MARCO-Doc and MS-MARCO-Pas, respectively.

Table 3: Comparisons between our proposed CertDR and the baseline on the BERT. Adversarial attacks are conducted by PRADA [48]. ADV corresponds to no defense. ADV and DA are evaluated under SR (%) and CertDR is evaluated under CondSR (%).

Dataset	ADV	DA	CertDR
MS-MARCO-Doc	96.7	57.0	40.0
MS-MARCO-Pas	91.4	64.6	57.4

To answer **RQ4**, as shown in Table 3, we observe that: (i) The SR is very high (i.e., up to 96.7% on MS-MARCO-Doc) if we do not take any defense, indicating that it is important to develop defense methods for NRMs to fight against adversarial attacks. (ii) Empirical defense method DA reduces the SR to some extent. However, it performs worse than CertDR. Hence, simply augmenting the training documents (as in NLP) is not a robust defense against attacks in IR. Future work should explore more adequate empirical defense methods in IR. Importantly, empirical defense methods do not provide rigorous certified robustness guarantees and the performance may significantly depend on the datasets and specific attacks. (iii) CertDR achieves the lowest CondSR on both datasets, indicating that our CertDR could certify the robustness theoretically while enhancing the robustness empirically for ranking models.

7 CONCLUSION

In this paper, we defined the notion of Certified Top-K Robustness for ranking models focusing on the characteristics of IR. We proposed a certifiably robust defense method called CertDR, based on randomized smoothing. The key idea is to smooth the ranking model with random word substitutions, and construct provable certification bounds based on the ranking property. Extensive experiments validate that CertDR outperforms existing defense methods and improves the certifiable robustness of ranking models.

In future work, it is worth to strengthen the notion of Certified Top-K Robustness to guarantee that the order of top-K ranking results remains unchanged. We hope that our study helps to put concerns about the robustness of NRMs on the research agenda and to motivate new defense ideas, including empirical and certified defenses of ranking models.

ACKNOWLEDGMENTS

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. 20144310 and 2021100, the Innovation Project of ICT, CAS under Grants No. E261090, the Young Elite Scientist Sponsorship Program by CAST under Grants No. YESS20200121, 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 Dutch Research Council, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998* (2018).

[2] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*. 89–96.

[3] Carlos Castillo and Brian D Davison. 2011. *Adversarial web search*. Vol. 4.

[4] Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In *NAACL*.

[5] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *ICML*.

[6] Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *SIGIR*. 985–988.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

[8] Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. 2020. Towards Robustness Against Natural Language Word Substitutions. In *International Conference on Learning Representations*.

[9] Krishnamurthy Dvijotham, Sven Gowal, Robert Stanforth, Relja Arandjelovic, Brendan O'Donoghue, Jonathan Uesato, and Pushmeet Kohli. 2018. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265* (2018).

[10] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *SPW*. IEEE, 50–56.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[12] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2018. Ranking robustness under adversarial document manipulations. In *SIGIR*.

[13] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2020. Ranking-Incentivized Quality Preserving Content Modification. In *SIGIR*.

[14] Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *CIKM*. 2041–2044.

[15] Zoltan Gyongyi and Hector Garcia-Molina. 2005. Web spam taxonomy. In *AIRWeb*.

[16] Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. 2019. Achieving Verified Robustness to Symbol Substitutions via Interval Bound Propagation. In *EMNLP/IJCNLP (1)*.

[17] Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *EMNLP*.

[18] Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified Robustness to Adversarial Word Substitutions.. In *EMNLP/IJCNLP (1)*.

[19] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[20] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *TACL* 8 (2020), 64–77.

[21] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. 2017. Reluplex: An efficient SMT solver for verifying deep neural networks. In *International conference on computer aided verification*. Springer, 97–117.

[22] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *SIGIR*.

[23] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE.

[24] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2017. Deep text classification can be fooled. *arXiv preprint arXiv:1704.08006* (2017).

[25] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies* 14, 4 (2021), 1–325.

[26] Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval*. Springer Science & Business Media.

[27] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Yingyan Li, and Xueqi Cheng. 2021. B-PROP: Bootstrapped pre-training with representative words prediction for ad-hoc retrieval. *arXiv preprint arXiv:2104.09791* (2021).

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.

[30] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *WWW*.

[31] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725* (2016).

[32] 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 *CoCo@ NIPS*.

[33] Shuzi Niu, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2012. Top-k learning to rank: labeling, ranking and evaluation. In *SIGIR*. 751–760.

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

[35] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*. 83–92.

[36] Kezban Dilek Onal, Ye Zhang, Ismail Sengor Altingovde, Md Mustafizur Rahman, Pinar Karagoz, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Thanh Nguyen, Dan Xu, Byron C. Wallace, Maarten de Rijke, and Matthew Lease. 2018. Neural information retrieval: At the end of the early years. *Information Retrieval Journal* 21, 2–3 (June 2018), 111–182.

[37] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[38] Jakub Piskorski, Marcin Sydow, and Dawid Weiss. 2008. Exploring linguistic features for web spam detection: a preliminary study. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*. 25–28.

[39] Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *SIGIR*. 275–281.

[40] Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating Web-based Question Answering Systems.. In *LREC*. Citeseer.

[41] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *ACL*. 1085–1097.

[42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *ACL*.

[43] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*. Springer, 232–241.

[44] Gagandeep Singh, Rupanshu Ganvir, Markus Püschel, and Martin Vechev. 2019. Beyond the single neuron convex barrier for neural network certification. *NIPS* (2019).

[45] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[46] Wenjie Wang, Pengfei Tang, Jian Lou, and Li Xiong. 2021. Certified robustness to word substitution attack with differential privacy. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1102–1112.

[47] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. 2019. On the Convergence and Robustness of Adversarial Training. In *ICML*.

[48] 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. *arXiv preprint arXiv:2204.01321* (2022).

[49] Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2021. Are Neural Ranking Models Robust? *arXiv preprint arXiv:2108.05018* (2021).

[50] Fen Xia, Tie-Yan Liu, and Hang Li. 2009. Statistical consistency of top-k ranking. *NIPS* 22 (2009).

[51] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality (JDIQ)* 10, 4 (2018), 1–20.

[52] Mao Ye, Chengyue Gong, and Qiang Liu. 2020. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. *arXiv preprint arXiv:2005.14424* (2020).

[53] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2019. Word-level textual adversarial attacking as combinatorial optimization. *arXiv preprint arXiv:1910.12196* (2019).

[54] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *TIST* 11, 3 (2020), 1–41.

[55] Mo Zhou, Zhenxing Niu, Le Wang, Qilin Zhang, and Gang Hua. 2020. Adversarial ranking attack and defense. In *ECCV*.