Stop Hardening Everything: A Training-Free Neuron-Level Defense for Neural Ranking Models

Yu-An Liu[1,2,3], Ruqing Zhang[1,2,3]*, Hongru Song[1,2,3], Jiafeng Guo[1,2,3], Yixing Fan[1,2,3], Xueqi Cheng[1,2,3]

[1]State Key Laboratory of AI Safety

[2]Institute of Computing Technology, Chinese Academy of Sciences

[3]University of Chinese Academy of Sciences

{liuyuan21b, zhangruqing, songhongru24s, guojiafeng, fanyixing, cxq}@ict.ac.cn

Abstract

While neural ranking models (NRMs) have achieved state-of-the-art performance in information retrieval, they remain highly vulnerable to imperceptible adversarial perturbations. Existing defenses are predominantly *data-centric*, exemplified by adversarial training, which requires constructing large collections of adversarial examples. By treating NRMs as black boxes and indiscriminately optimizing all model parameters, these methods incur substantial computational cost and often degrade performance on clean data due to overfitting. In this paper, we advocate that adversarial vulnerability is not uniformly distributed across model parameters, but instead originates from specific internal units. We propose a paradigm shift toward a *model-centric* defense that addresses vulnerability at its architectural source, without requiring costly retraining or adversarial data generation. Specifically, we introduce *Search in the Model*, a novel training-free framework that performs fine-grained identification and rectification of vulnerable neurons directly within the model. By formulating neuron identification as a ranking problem, we develop a maximum marginal vulnerability criterion to precisely locate the top-K neurons most responsible for model vulnerability, and apply targeted neuronal inverse perturbation to correct them. Extensive experiments on MS MARCO and TREC 19 show that our approach outperforms state-of-the-art baselines in both defense efficiency and robustness to seen and unseen attacks, while preserving strong performance on clean data.

1 Introduction

With the rise of deep learning (LeCun et al., 2015), the effectiveness of information retrieval (IR) technology has been significantly enhanced (Guo et al., 2020; Fan et al., 2022). Neural Ranking Models (NRMs), particularly those based on pre-trained language models, have achieved state-of-the-art performance by capturing complex semantic relationships between queries and documents (Lin et al., 2022; Guo et al., 2020; Xiong et al., 2021).

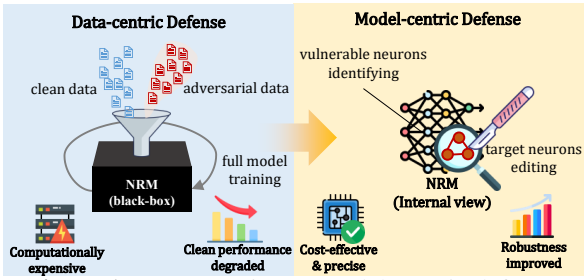

Figure 1: Data-centric v.s. model-centric defense.

Vulnerability of NRMs. Despite their strong relevance modeling capability, NRMs inherit the robustness limitations of neural networks and are highly susceptible to adversarial attacks (Wu et al., 2023; Liu et al., 2022; Chen et al., 2023; Bigdeli et al., 2025). Recent studies show that even imperceptible perturbations, such as inserting synonymous words into documents, can significantly manipulate ranking results and promote malicious content (Liu et al., 2025b,a; Bigdeli et al., 2025; Wu et al., 2023). Such vulnerabilities pose serious threats to the reliability and security of search engines, potentially enabling spam, misinformation propagation, and ultimately eroding user trust (Liu et al., 2025a; Gyöngyi and Garcia-Molina, 2005).

Limitations of existing data-centric defenses. To mitigate these vulnerabilities, prior work has explored adversarial defenses for NRMs (Liu et al., 2024b; Wu et al., 2022). Most existing approaches follow a data-centric paradigm, as shown in Figure 1, treating the NRM largely as a black box and improving robustness by synthesizing large amounts of data to retrain the entire model. A representative example is adversarial training based on external data augmentation (Liu et al., 2024b), which mixes clean and adversarial examples during training to harden all model parameters. However, such data-centric defenses suffer from two

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23474–23484

fundamental limitations. (i) Firstly, generating sufficient adversarial samples and jointly optimizing all model parameters is computationally expensive, especially as NRMs continue to scale in size. Moreover, prior studies suggest that not all neurons contribute equally to a model's decision for a given input, indicating substantial redundancy in full-parameter optimization (Ghorbani and Zou, 2020). (ii) Secondly, adversarial training often overfits to specific attack patterns, leading to degraded ranking performance on clean data and an unfavorable robustness–effectiveness trade-off.

From data-centric to model-centric defense. As NRMs grow increasingly large, relying on data-centric defenses becomes progressively inefficient, much like attempting to maneuver a massive ship through constant external force rather than correcting its internal steering mechanism. This observation motivates a shift in perspective: *instead of continuously generating more adversarial data and retraining entire models, can we directly identify and fix the internal sources of vulnerability in NRMs?*

Therefore, in this paper, we advocate a model-centric defense paradigm. We advance that, similar to previous findings that not all neurons contribute equally to effectiveness (Ghorbani and Zou, 2020), adversarial vulnerability is also not uniformly distributed across all neurons. Rather, a small subset of vulnerable neurons is disproportionately responsible for model fragility under adversarial perturbations. Consequently, blindly hardening the entire network is inefficient; selectively correcting neuron-level weaknesses without additional training offers a more cost-effective solution.

Search in the model. To operationalize the model-centric paradigm, we propose Search in the Model, a novel training-free defense framework for NRMs. The core idea is to cast vulnerable neuron identification as a ranking problem and directly edit the most vulnerability-inducing neurons without retraining the model. Inspired by maximal marginal relevance (MMR) from fairness in learning to rank (Xia et al., 2015), we introduce maximum marginal vulnerability, which identifies a set of neurons whose combined influence maximizes adversarial impact. "Search in the Model" includes three phases: (i) *Adversarial scanning*, which constructs a small number of adversarial examples to probe neuron-level vulnerability; (ii) *Combinatorial vulnerability identification*, which ranks neurons using maximum marginal vulnerability to select the most critical subset; and (iii) *Neuronal inverse perturbation*, which corrects vulnerable neurons by applying targeted parameter updates based on deviations between clean and adversarial responses. Together, this framework provides an efficient alternative to data-centric defenses by addressing adversarial vulnerability through direct neuron-level intervention rather than large-scale retraining.

Experimental results. We conduct extensive experiments on the MS MARCO benchmark (Nguyen et al., 2016), targeting two representative categories of adversarial attacks (Wu et al., 2023; Bigdeli et al., 2025). The results show that our method significantly improves robustness against both seen and unseen attacks over 13%, while maintaining, and in some cases improving, ranking performance on clean data compared to state-of-the-art baselines without additional training.

2 Background

Adversarial attacks against NRMs. The web is a canonical example of a competitive search environment (Kurland and Tennenholtz, 2022). Originating from search engine optimization (Gyöngyi and Garcia-Molina, 2005), adversarial attacks aim to increase the exposure of low-quality documents by boosting their rankings with imperceptible perturbations (Wu et al., 2023; Liu et al., 2025a). As NRMs directly determine the final ranked list, they are often at the front line of such threats.

Given a query q and a target document d, the objective of an adversarial attack against an NRM f_θ is to generate an imperceptible perturbation p:

$$\min_{p} \; f_\theta\left(q, d \oplus p\right) \quad \text{s.t. } p \in \mathcal{V}, \tag{1}$$

where $f_\theta\left(q, d \oplus p\right)$ denotes the ranking position of the perturbed document $d \oplus p$ in the ranked list generated by f_θ with respect to query q. $\mathcal{V}$ defines the allowable perturbation space (e.g., synonym replacement constraints).

Adversarial attacks on NRMs are commonly categorized into two types: (i) *Gradient-based attacks* (Wu et al., 2023; Liu et al., 2022, 2023), which typically rely on surrogate models to identify token-level perturbations that strongly influence the target model's gradients; and (ii) *In-context-learning-based attacks* (Liu et al., 2025b; Bigdeli et al., 2025), which leverage large language models (LLMs) to uncover ranking-sensitive linguistic patterns and generate sentence-level perturbations. In our experiments, we select two repre-

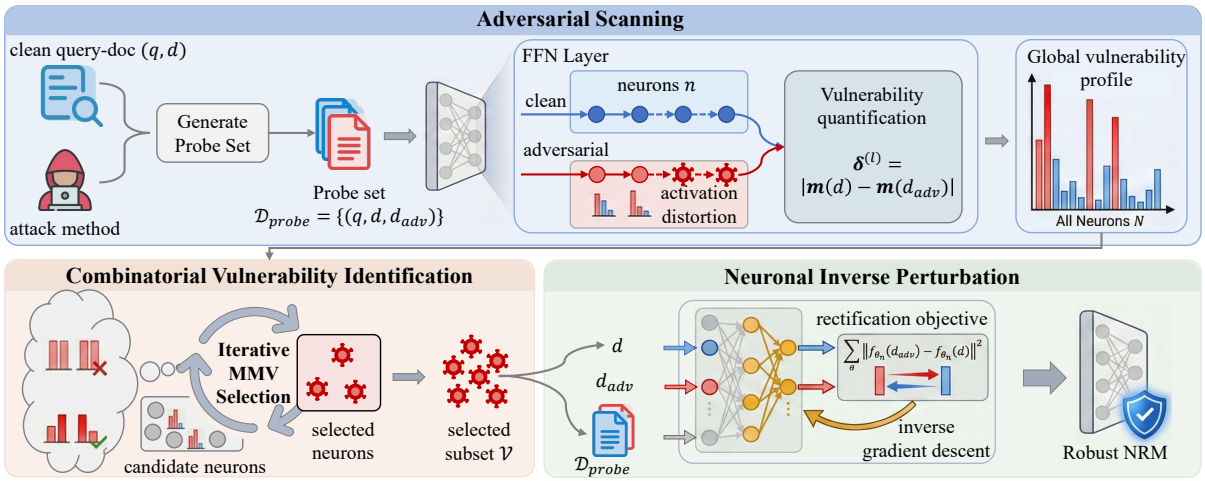

Figure 2: Overview of the proposed "Search in the model" framework.

sentative attacks from these categories to evaluate defense effectiveness: MARA (Liu et al., 2024a) as a gradient-based ranking attack, and FSAP (Bigdeli et al., 2025) as an in-context-learning-based attack.

Adversarial defense of NRMs. Corresponding to adversarial attacks, the objective of adversarial defense for NRMs is to restore the correct ranking positions of adversarially perturbed documents while preserving ranking effectiveness on clean data (Wu et al., 2022; Liu et al., 2024b).

Most existing defenses adopt a *data-centric* paradigm, exemplified by adversarial training, which treats the model as a black box and optimizes the entire parameter θ with large-scale augmented data (Wu et al., 2022; Liu et al., 2024b). Formally, data-centric adversarial defense is commonly formulated as a min–max optimization problem:

$$\min_{\theta} \mathbb{E}_{(q,d)\sim\mathcal{D}_{\text{train}}} \left[\max_{p\in\mathcal{V}} \mathcal{L}\left(f_\theta(q, d \oplus p; \theta), y\right) \right], \tag{2}$$

where $\mathcal{L}$ is ranking loss and y is ground-truth.

In contrast, our approach follows a model-centric perspective. Rather than retraining all parameters, we aim to identify a small vulnerable neuron set $\mathcal{V} \subset \theta$ (where $|\mathcal{V}| \ll |\theta|$) and apply a targeted inverse perturbation τ to these neurons only. The goal is to align the model's behavior on adversarial inputs with its behavior on clean inputs, while keeping most parameters fixed. It can be formulated as:

$$\min_{\mathcal{V},\tau} \sum_{(q,d)} \left\| f_\theta(q, d \oplus p; \theta \odot \mathcal{V}^\tau) - f_\theta(q, d; \theta) \right\|^2, \tag{3}$$

where $\theta \odot \mathcal{V}^\tau$ denotes the operation of applying the correction τ only to the neurons in indices $\mathcal{V}$ while keeping other parameters fixed. A robust defense should not only withstand known (seen) attack strategies but also generalize to unseen attacks. Accordingly, we defend against one representative attack during method design and evaluate robustness against a different, previously unseen attack.

3 Our method: Search in the model

We propose "Search in the Model", a training-free adversarial defense framework that protects NRMs by directly operating in the model parameter space. Our method explicitly identifies and selectively strengthens a small set of vulnerable neurons that are most responsible for adversarial fragility. Compared to conventional data-centric defenses, our method offers an efficient and robust alternative: it requires no retraining, avoids indiscriminate parameter updates, and reduces the risk of overfitting to specific attack patterns. As shown in Figure 2, our approach operates in three distinct phases, which are detailed below.

3.1 Adversarial scanning

To enable model-centric hardening at the neuronal level, we first identify a candidate set of neurons whose behavior is sensitive to adversarial perturbations. Rather than treating all parameters uniformly, our goal is to quantify how individual neurons respond to adversarial noise and to isolate those that exhibit disproportionate instability. To this end, we leverage existing attack methods to construct a small probe set and use it to measure neuron-level vulnerability in a training-free manner.

Neuron definition in NRMs. We consider a Transformer-based architecture as the backbone of the NRM. Following prior work (Geva et al., 2021), the feed-forward networks (FFNs) can be viewed as key–value memories that store semantic

and task-specific patterns. Accordingly, we focus our analysis on FFN neurons, which provide a natural and interpretable unit for model introspection.

For the l-th layer, given the hidden state $\mathbf{h}$, the intermediate activation vector $\mathbf{m}^{(l)} \in \mathbb{R}^{d_{ff}}$ is:

$$\mathbf{m}^{(l)} = \sigma\left(\mathbf{h}\mathbf{W}_1^{(l)} + \mathbf{b}_1^{(l)}\right), \tag{4}$$

where $\mathbf{W}1^{(l)}$ and $\mathbf{b}1^{(l)}$ are the parameters of the first linear transformation, and d_{ff} is the intermediate dimension. We define the i-th element of $\mathbf{m}^{(l)}$ as a single neuron, denoted by $n_{l,i}$. The full set of neurons in the model is therefore defined as:

$$\mathcal{N} = \{n_{l,i} \mid 1 \le l \le L, 1 \le i \le d_{ff}\}. \tag{5}$$

Vulnerability quantification. We next quantify neuron-level vulnerability in NRMs by measuring how internal activations change under adversarial perturbations. Let f_θ denote the NRM with parameters θ. For a given query q and document d, let $\mathbf{m}^{(l)} \in \mathbb{R}^{D_l}$ represent the activation vector of layer l. (i) We first construct a probe set $\mathcal{D}_{probe} = \{(q, d, d_{adv})\}$ using a known attack method, where d_{adv} is the adversarial example of d; (ii) Then, we define the activation distortion vector, $\boldsymbol{\delta}^{(l)}$, for layer l as the element-wise difference between activations on clean and adversarial inputs:

$$\boldsymbol{\delta}^{(l)}(q, d) = \left|\mathbf{m}^{(l)}(q, d) - \mathbf{m}^{(l)}(q, d_{adv})\right|; \tag{6}$$

and (iii) Finally, by averaging this distortion over the probe set, we obtain the global vulnerability profile, $\mathbf{V} \in \mathbb{R}^N$, where N is the total number of neurons in the target layers:

$$\mathbf{V}_i = \mathbb{E}_{(q,d)\sim\mathcal{D}_{probe}}\left[\boldsymbol{\delta}_i(q, d) \cdot \omega_i\right], \tag{7}$$

where $\omega_i = \left|\frac{\partial \mathcal{L}}{\partial h_i}\right|$ is a gradient-based importance weight, which emphasizes neurons that are not only sensitive to adversarial perturbations but also exert a strong influence on the final ranking loss.

3.2 Combinatorial vulnerability identification

A naive strategy is to select the top-K neurons with the highest individual vulnerability scores $\mathbf{V}_i$. However, this approach is suboptimal in practice, as vulnerable neurons often exhibit highly correlated failure patterns. When multiple neurons respond similarly to adversarial perturbations, correcting all of them leads to diminishing returns and inefficient use of the intervention budget. To identify an effective combination of vulnerable neurons, we draw inspiration from maximal marginal relevance (MMR), a classic principle from fairness-aware learning to rank (Xia et al., 2015).

Maximal marginal relevance in IR. MMR is a widely adopted criterion in IR for balancing relevance with diversity (Xia et al., 2015). Rather than independently selecting top-ranked items, MMR employs a greedy iterative strategy to penalize redundancy: it selects candidates that maximize relevance to the query while minimizing similarity to the already selected set. This mechanism provides a theoretical basis for selecting candidates that are individually important yet collectively diverse, motivating our neuron selection strategy.

Neuron combination identification with maximal marginal vulnerability. Building on MMR, we introduce maximum marginal vulnerability (MMV) to guide neuron selection. Here, individual neuron vulnerability serves as the utility signal, while similarity between neuron failure patterns represents redundancy. Our goal is to select a subset of neurons $\mathcal{V} \subset \theta, |\mathcal{V}| = K$, that maximizes vulnerability coverage. We adopt a greedy selection strategy, iteratively adding the neuron that maximizes marginal vulnerability:

$$\hat{n} = \arg\max_{n_i \in \theta \setminus \mathcal{V}} \left[\lambda \cdot \mathbf{V}_{n_i} - (1 - \lambda) \cdot \max_{n_j \in \mathcal{V}} \mathrm{Sim}(\mathbf{u}_{n_i}, \mathbf{u}_{n_j})\right], \tag{8}$$

where $\mathcal{V}$ is the set of selected neurons, $\mathbf{V}_{n_i}$ is the individual vulnerability score of candidate neuron n_i, $\mathbf{u}_n \in \mathbb{R}^{|\mathcal{D}_{probe}|}$ is the behavior vector of n, representing its distortion across all samples in probe set, $\mathrm{Sim}(\cdot, \cdot)$ is the cosine similarity measuring how correlated the failure patterns of two neurons are, and λ is a hyperparameter balancing pure vulnerability against diversity. By iteratively solving this equation until $|\mathcal{V}| = K$, we obtain a compact neuron set that captures the broadest spectrum of adversarial failure modes, rather than repeatedly selecting neurons that break in the same way.

3.3 Neuronal inverse perturbation

After identifying the vulnerable neuron set $\mathcal{V}$, the next step is to correct their behavior. A straightforward solution is to suppress these neurons using generic regularization techniques such as weight decay (Loshchilov and Hutter, 2017). However, indiscriminately weakening neurons risks degrading ranking performance, as the same neurons may also contribute to clean relevance signals.

Instead, we adopt an inverse perturbation strategy that directly counteracts adversarial effects. Rather than diminishing neuron influence, our goal is to realign their responses so that adversarial inputs elicit internal activations consistent with clean inputs. This preserves useful ranking signals while

neutralizing adversarial noise.

Neuron-level behavioral alignment. For each neuron $n \in \mathcal{V}$, we seek a targeted parameter update that minimizes the discrepancy between its responses to clean and adversarial documents. (i) First, we define the rectification objective $\mathcal{L}_{rect}$ as

$$\mathcal{L}_{rect} = \sum_{n \in \mathcal{V}} \| f_{\theta_n}(d_{adv}) - f_{\theta_n}(d) \|_2^2 ; \quad (9)$$

(ii) We then compute the inverse gradient solely for the parameters associated with the subset $\mathcal{V}$. The model parameters are updated using a dedicated gradient descent step,

$$\theta_n \leftarrow \theta_n - \eta \cdot \text{sign}\left(\nabla_{\theta_n} \mathcal{L}_{rect}\right). \quad (10)$$

By aligning internal activations rather than re-optimizing rankings, the method stabilizes vulnerable neurons while preserving the behavior of robust components. As a result, the model becomes resistant to adversarial perturbations δ without sacrificing effectiveness on clean data.

4 Experimental settings

Dataset and target ranking models. We conduct experiments on *MS MARCO Passage Ranking* (MS MARCO) (Nguyen et al., 2016) and TREC DL 2019 (TREC19) (Craswell et al., 2020). They are based on a large-scale corpus for Web retrieval, with about 8.84 million passages. Following Bigdeli et al. (2025), we chose two typical ranking models that achieve promising effectiveness, i.e., monoBERT (Nogueira et al., 2020) with an encoder-only architecture and monoT5 (Nogueira et al., 2020) with an encoder-decoder architecture.

Attack methods. To evaluate the proposed defense against seen and unseen attacks, we selected two representative attack methods of different types: (i) For gradient-based attacks, we adopt *MARA* (Liu et al., 2024a), which identifies token-level vulnerability from the model's gradients and generates multi-granular perturbations; and (ii) For in-context-learning-based attacks, we adopt *FSAP* (Bigdeli et al., 2025), which leverages LLMs to uncover ranking-sensitive linguistic patterns and generates sentence-level perturbations. We sequentially developed defense strategies tailored to each attack method and evaluated the defense performance under both of the two attack scenarios.

Evaluation metrics. Following (Liu et al., 2024b), we adopt four metrics: (i) *CMRR* evaluates mean reciprocal rank (MRR) (Ma et al., 2021; Yan et al., 2021) on the clean dataset. (ii) *RMRR* evaluates the MRR performance on the after-attack dataset. (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 et al., 2022). For CMMR and RMMR, we focus on the top 10 results. A higher CMRR indicates better effectiveness of the ranking model. Better robustness is reflected by a higher RMRR alongside lower ASR and LSD.

Baselines. (i) *Standard training (ST)*: We directly train NRM without defense mechanisms. (ii) *Adversarial training (AT)*: We follow the vanilla AT method (Goodfellow et al., 2015) to directly include the adversarial examples during training. (iii) *CertDR* is a certified defense method for NRMs (Wu et al., 2022), which achieves certified top-K robustness against word subsititution attacks. (iv) *PIAT* is a theory-driven adversarial training method that optimizes the trade-off between ranking performance and defense capabilities (Liu et al., 2024b). (v) *SNS* is an interpretable method that identifies sensitive neurons and regularizes them to enhance model interpretability.

Implementation details. To identify vulnerable neurons, we generate the probe set $\mathcal{D}_{probe}$ containing 500 query-document pairs. For combinatorial vulnerability identification, we set the diversity hyperparameter $\lambda = 0.7$, which provides a balanced trade-off between the magnitude of individual neuron sensitivity and the coverage of diverse failure modes. We experiment with $K \in \{32, 64, 128, 256\}$, representing the number of "targeted neurons" to be edited. In our final reported results, $K = 128$ is used as the default, which accounts for less than 0.01% of the total model parameters. For neuronal inverse perturbation, we use a very small learning rate of 5×10^{-6}.

We implement target ranking models as in previous work (Liu et al., 2024b; Nogueira et al., 2020; Liu et al., 2023). First-stage retrieval uses Anserini (Yang et al., 2018) (BM25) to retrieve the top 100 candidates, and the final ranked list is generated by re-ranking this initial set with the trained model.

For MS MARCO, we randomly sample 1000 Dev queries as targets, attacking their ranked lists for evaluation. For TREC 19, we adopt all queries in the evaluation. For each query, we

Dataset & NRM	Method	Seen Attack Scenario				Unseen Attack Scenario			
MS MARCO		CMRR	RMRR	ASR↓	LSD↓	CMRR	RMRR	ASR↓	LSD↓
Seen Attack: Gradient-based Attack									
monoBERT	ST	38.3	30.2	95.3	36.5	38.3	28.5	98.5	46.2
	CertDR	32.1	28.4	59.2	18.2	32.1	27.1	68.8	29.5
	AT	37.5	32.3	58.2	17.3	37.5	31.0	67.5	27.8
	PIAT	38.6	35.4	48.3	13.5	38.6	33.8	59.1	24.2
	SNS	37.2	35.9	43.2	10.2	37.2	34.2	54.7	19.6
	Ours	**38.8**	**36.5**	**39.2**	**8.3**	**38.8**	**35.1**	**47.6**	**17.4**
monoT5	ST	39.4	31.5	94.4	35.8	39.4	29.8	97.6	45.4
	CertDR	33.2	29.7	58.5	17.5	33.2	28.2	68.1	28.7
	AT	38.6	33.4	57.3	16.6	38.6	32.1	66.8	26.9
	PIAT	39.5	36.6	47.6	12.8	39.5	34.9	58.3	23.4
	SNS	38.3	37.1	42.4	9.4	38.3	35.4	54.0	18.8
	Ours	39.5	**37.7**	**38.4**	**7.5**	39.5	**36.3**	**46.8**	**16.7**
In-context-learning-based Attack									
monoBERT	ST	38.3	29.8	97.4	38.2	38.3	28.1	96.4	44.1
	CertDR	31.0	27.9	61.5	20.4	31.0	26.8	67.2	27.8
	AT	36.1	31.7	60.1	19.5	36.1	30.7	65.4	26.2
	PIAT	38.0	34.8	50.4	15.6	38.0	33.4	57.8	22.5
	SNS	36.9	35.3	45.7	12.3	36.9	33.9	52.1	17.4
	Ours	**38.2**	**36.1**	**41.8**	**10.5**	**38.2**	**34.7**	**45.5**	**15.2**
monoT5	ST	39.4	30.9	96.8	37.5	39.4	29.4	95.8	43.7
	CertDR	32.2	29.1	60.2	19.8	32.2	27.9	66.5	27.1
	AT	35.3	32.8	59.4	18.9	35.3	31.8	65.1	25.4
	PIAT	38.6	35.9	49.8	14.2	38.6	34.6	56.9	21.8
	SNS	36.9	36.4	44.1	11.2	36.9	35.1	51.4	16.9
	Ours	38.4	**37.2**	**40.7**	**9.4**	38.4	**36.1**	**44.3**	**14.8**
TREC 19		**Seen Attack Scenario**				**Unseen Attack Scenario**			
		CMRR	RMRR	ASR↓	LSD↓	CMRR	RMRR	ASR↓	LSD↓
Seen Attack: Gradient-based Attack									
monoBERT	ST	82.5	63.8	90.2	34.6	82.5	59.2	92.1	43.1
	CertDR	68.4	59.5	56.1	17.3	68.4	56.4	64.3	27.5
	AT	80.7	68.2	54.8	16.5	80.7	64.3	62.9	25.9
	PIAT	83.2	75.1	45.4	12.8	83.2	70.8	55.2	22.6
	SNS	79.8	76.3	40.8	9.7	79.8	71.5	51.3	18.3
	Ours	**83.9**	**77.8**	**37.1**	**7.9**	**83.9**	**73.6**	**44.5**	**16.2**
monoT5	ST	85.1	66.5	88.5	33.9	85.1	61.9	91.4	42.4
	CertDR	70.9	62.1	55.4	16.6	70.9	58.6	63.8	26.8
	AT	83.2	70.5	54.1	15.7	83.2	66.8	62.3	25.1
	PIAT	85.3	77.8	45.1	12.1	85.3	72.9	54.6	21.9
	SNS	82.6	78.9	40.1	8.9	82.6	74.2	50.5	17.6
	Ours	85.0	**80.4**	**36.3**	**7.1**	85.0	**75.9**	**43.7**	**15.6**
Seen Attack: In-context-learning-based Attack									
monoBERT	ST	82.5	62.9	92.2	36.1	82.5	58.3	90.4	41.2
	CertDR	65.8	58.7	58.3	19.4	65.8	55.6	62.9	25.9
	AT	77.4	66.8	56.9	18.5	77.4	63.8	61.2	24.5
	PIAT	81.8	73.6	47.7	14.8	81.8	69.8	54.1	21.1
	SNS	79.3	74.8	43.3	11.6	79.3	70.8	48.8	16.3
	Ours	82.3	**76.7**	**39.6**	**9.9**	82.3	**72.6**	**42.6**	**14.1**
monoT5	ST	85.1	65.3	91.4	35.4	85.1	60.9	89.6	40.8
	CertDR	68.6	61.2	57.1	18.8	68.6	57.9	62.1	25.4
	AT	75.6	69.2	56.3	17.9	75.6	66.1	60.8	23.7
	PIAT	83.2	76.1	47.2	13.5	83.2	72.3	53.2	20.4
	SNS	79.3	77.3	41.8	10.6	79.3	73.4	48.1	15.8
	Ours	82.7	**79.1**	**38.6**	**8.9**	82.7	**75.4**	**41.5**	**13.9**

Table 1: Ranking and defense performance of our method and the baselines across the seen and unseen attack scenarios on two benchmark datasets

sample 1 document from 9 ranked ranges (i.e., [11,20],...,[91,100]) as in (Wu et al., 2023; Liu et al., 2024b). These 9 target documents are attacked to generate adversarial examples, and defense performance is evaluated on the attacked lists paired with their queries. For baselines, we sample 0.1 million training queries (to reduce time overhead) as in (Wu et al., 2022; Liu et al., 2024b); for each, 10 documents are randomly sampled from its initial candidate set to form training examples.

5 Experimental results

5.1 Main results

Table 1 compares ranking effectiveness and defense robustness across different methods. Observations on the data-centric defense baselines are: (i) monoT5 slightly outperforms monoBERT in ranking effectiveness, likely due to the stronger modeling capacity of T5. (ii) In the absence of defensive mechanisms, ST suffers a substantial degradation in ranking performance under high ASR and LSD, indicating that NRM is highly vulnerable to adversarial interference in realistic settings. (iii) CertDR and AT markedly improve robustness but at the cost of degraded performance on clean data, whereas PIAT preserves original performance but incurs significantly higher training overhead. This suggests that data-centric defenses are prone to overfitting and limited defense efficiency. (iv) All methods exhibit weaker robustness against in-context learning–based attacks than gradient-based attacks, along with poorer generalization. This implies that LLMs can more effectively exploit NRM preferences via large-scale training data and advanced reasoning, posing greater challenges to robust ranking.

Focusing on model-centric defenses, we observe that: (i) Overall, model-centric methods (SNS and ours) consistently outperform data-centric defenses in both ranking effectiveness and adversarial robustness, indicating that repairing vulnerabilities at the neural level can improve robustness without sacrificing clean performance. (ii) Our method exhibits substantially lower performance degradation under unseen attacks than existing approaches, suggesting that precise vulnerability identification and optimization at the neural network level provides more fundamental protection. (iii) Our approach consistently surpasses SNS in both original ranking quality and defense robustness, demonstrating that locating maximum vulnerability combinations via MMV and repairing neural behaviors constitutes a more effective repair paradigm for NRM.

5.2 Ablation study

To further validate our approach, we introduced two variants: (i) Ours$_{TopK}$ replaces the MMV-based fragile neuron selection with greedy selection; (ii) Ours$_{Mask}$ discards identified vulnerable neurons instead of modifying them. We take the monoBERT under in-context-learning-based attack under MS MARCO dataset as an example.

Method	CMRR	RMRR	ASR↓	LSD↓
Ours	38.2	36.1	41.8	10.5
Ours$_{TopK}$	38.0	35.5	43.6	11.1
Ours$_{Mask}$	37.2	34.6	46.9	13.4

Table 2: Comparison of our method with its variants.

As shown in Table 2, we find that, (i) Greedy neuron selection based solely on vulnerability scores degrades defensive performance and harms the model's discriminative ability, whereas considering inter-neuronal relationships enables more effective repair while reducing unintended side effects; (ii) Directly discarding vulnerable neurons severely degrades performance, as these neurons still encode useful functions. Neuron behavior alignment preserves their original capabilities while enabling effective repair.

5.3 Impact of the number of edited neurons K

K is a key hyperparameter, as it determines the number of vulnerable neurons selected. We take the monoBERT under in-context-learning-based attack under MS MARCO as an example. Figure 3 shows a comparison of defense performance between our method and SNS. The results reveal a limitation of SNS's greedy strategy: individually vulnerable neurons often exhibit redundant failure patterns, leading to diminishing returns when repaired. In contrast, by maximizing coverage of diverse vulnerability types and minimizing redundancy, our MMR-based strategy achieves more comprehensive robustness.

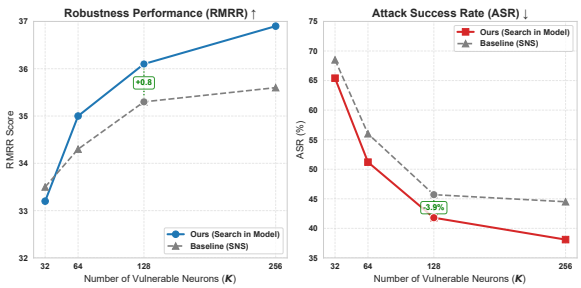

Figure 3: Impact of the number of edited neurons K exhibited by our method, compared with SNS.

5.4 Cost comparison

Table 3 shows a comparison between our method and selected baselines in terms of time overhead, the scale of constructed data (Data size), and the size of perturbed neurons (Optim. size). We take the monoBERT under in-context-learning-based attack under MS MARCO as an example. From the result, we can intuitively find that, compared with traditional data-centric methods, our approach

Method	Time overhead	Data size	Optim. size
CertDR	2h	20K	100%
PIAT	15h	1M	100%
Ours	**0.4h**	**500**	$<$ **0.01%**

Table 3: Time and computational overhead comparison between our method and selected baselines.

significantly reduces computational costs and accelerates model hardening by precisely adjusting individual neurons. In practice, adversarial attack-defense is a time-sensitive cat-and-mouse game, and our method enables more agile deployment, thereby gaining an edge for service providers.

6 Related Work

Neural ranking models. The rise of deep learning has accelerated the adoption of NRMs (Onal et al., 2018; Guo et al., 2020), which have consistently demonstrated clear advantages over traditional learning-to-rank approaches. Recent work has incorporated pre-trained language models into ranking (Fan et al., 2022), further strengthening the effectiveness of NRMs. In parallel, researchers have improved NRM training via data augmentation strategies, such as hard negative mining (Xiong et al., 2021; Zhan et al., 2021), pushing performance to new state-of-the-art levels. However, despite these gains in effectiveness, existing studies largely neglect the adversarial robustness of NRMs.

Competitive web search. The web is a canonical competitive search setting, featuring authors optimizing content for better rankings (Kurland and Tennenholtz, 2022)—a practice called search engine optimization (SEO) that boosts webpage visibility for specific queries (Gyöngyi and Garcia-Molina, 2005). Accordingly, extensive research now focuses on adversarial attacks against NRMs to simulate real-world SEO. They aim to inject human-imperceptible perturbations into documents that can deceive neural networks (Szegedy et al., 2014). The most representative among them include gradient-based attacks (Wu et al., 2023; Liu et al., 2022, 2024a) and in-context-learning-based attacks (Liu et al., 2025b; Bigdeli et al., 2025). Through different approaches, these studies reveal that current NRMs face severe vulnerability.

Adversarial defense for NRMs. Existing defenses against adversarial attacks on NRMs are largely data-centric, relying on large-scale data synthesis for retraining (Wu et al., 2022; Liu et al., 2024b). For instance, adversarial training enhances a model's robustness against adversarial attacks by incorporating adversarial examples into the training process (Liu et al., 2024b); while certified defense improves NRMs' resistance within specific ranges by adding a certain proportion of random smoothing noise to training data (Wu et al., 2022). However, these defense mechanisms often come with high computational costs and optimize all model parameters indiscriminately. This leads to inefficient defense and frequently compromises the model's inherent ranking capabilities. Therefore, we propose model-centric defense, i.e., by identifying vulnerable neurons in the model that contribute to misbehavior, we can perform surgical-level precision repairs for NRMs.

Model editing. Model editing aims to update pre-trained models with targeted knowledge without full retraining (Meng et al., 2022a; Mitchell et al.). Early methods such as ROME (Meng et al., 2022a) perform rank-one updates on selected FFN layers to localize and modify factual knowledge, while MEMIT (Meng et al., 2022b) extends this paradigm to efficient batch editing via gradient-based updates. Subsequent work focuses on mitigating error accumulation in sequential edits to preserve previously edited knowledge (Wang et al., 2023; Li et al., 2023). To date, model editing has been primarily applied to knowledge updating (Meng et al., 2022a,b) and bias mitigation (Yu and Ananiadou, 2025; Nadeem et al., 2025). In contrast, we argue that model vulnerabilities are also concentrated in a small subset of neurons. Based on this insight, we identify neuron combinations that contribute most to ranking vulnerability and repair them from a ranking-oriented perspective.

7 Conclusion

This paper presents "Search in the Model," a novel framework that shifts the defense paradigm of NRMs from data-centric augmentation to model-centric introspection. By identifying "vulnerable neurons" as the root cause of adversarial fragility, we move beyond black-box defenses toward surgical parameter rectification. Specifically, we leverage MMV-based combinatorial selection to pinpoint diverse coalitions of sensitive neurons and employ an inverse perturbation strategy to stabilize their activations. Our results confirm that securing NRMs does not require exhaustive retraining, but rather a precise "healing" of the model's internal weak links.

Limitations

Our work has several limitations to address in future research. (i) The identification of vulnerable neurons depends on the initial probe set. While our fairness-based selection improves diversity, the identified neurons may still be biased toward the specific attack types used during the scanning phase. (ii) While effective for Transformer-based NRMs, the distribution and behavior of vulnerable neurons may vary across different architectures (e.g., GNNs or lightweight ranking models), requiring further cross-model validation; and (iii) Our current method treats vulnerability as static. In a "cat-and-mouse" game with adaptive attackers, the model might develop new weak links after the primary vulnerable neurons are rectified, necessitating a more dynamic or iterative monitoring approach.

Ethical considerations

We approach ethics with great care. In this paper, all the models we use are open-source. For datasets, we construct benchmarks based on the open-source dataset. We ensured that all data we use was desensitized. Additionally, the methods we propose aim to enhance the robustness of NRMs and do not encourage or induce the model to produce any harmful information or leakage of user data.

Acknowledgements

This work was funded by the National Key Research and Development Program of China under Grants No. 2023YFA1011602, the National Natural Science Foundation of China under Grants No. 62472408, U25B2076, 62372431 and 62441229, and the Strategic Priority Research Program of the CAS under Grant No. XDB0680102. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Amin Bigdeli, Negar Arabzadeh, Ebrahim Bagheri, and Charles LA Clarke. 2025. Adversarial attacks against neural ranking models via in-context learning. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 211–220.

Xuanang Chen, Ben He, Zheng Ye, Le Sun, and Yingfei Sun. 2023. Towards imperceptible document manipulations against neural ranking models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6648–6664.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. corr abs/2003.07820 (2020). *arXiv preprint arXiv:2003.07820*.

Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang, and Jiafeng Guo. 2022. Pre-training methods in information retrieval. *Foundations and Trends in Information Retrieval*, 16(3):178–317.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Amirata Ghorbani and James Y Zou. 2020. Neuron shapley: Discovering the responsible neurons. *Advances in neural information processing systems*, 33:5922–5932.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067.

Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Web spam taxonomy. In *AIRWeb 2005: First International Workshop on Adversarial Information Retrieval on the Web*, pages 39–47.

Oren Kurland and Moshe Tennenholtz. 2022. Competitive search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2838–2849.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Yifan Li, Yuxi Zhang, Bingqian Li, Jinsong Su, Zhiyuan Wang, Yutao Hou, and Lichao Wang. 2023. Alphaedit: Null-space constrained model editing for language models. *arXiv preprint arXiv:2310.05737*.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2022. *Pretrained Transformers for Text Ranking: Bert and Beyond*. Springer Nature.

Jiawei Liu, Yangyang Kang, Di Tang, Kaisong Song, Changlong Sun, Xiaofeng Wang, Wei Lu, and Xiaozhong Liu. 2022. Order-disorder: Imitation adversarial attacks for black-box neural ranking models. In *2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2025–2039.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. 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*, pages 1647–1656.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024a. Multi-granular adversarial attacks against black-box neural ranking models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1391–1400.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2025a. *ACM Transactions on Information Systems*, 44(1):1–48.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2025b. Attack-in-the-chain: bootstrapping large language models for attacks against black-box neural ranking models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12229–12237.

Yu-An Liu, Ruqing Zhang, Mingkun Zhang, Wei Chen, Maarten de Rijke, Jiafeng Guo, and Xueqi Cheng. 2024b. Perturbation-invariant adversarial training for neural ranking models: improving the effectiveness-robustness trade-off. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8832–8840.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1513–1522.

Kevin Meng, David Bau, Alex Andonian, Yossi Belinkov, Yossi Belinkov, Yifeng Zhou, Mikel Artetxe, Mike Lewis, Luke Zettlemoyer, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Rank-one model editing for large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 28097–28110.

Kevin Meng, David Bau, Alex Andonian, Yossi Belinkov, Yifeng Zhou, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Mass-editing memory in gpt. In *NeurIPS 2022 Workshop on Machine Learning Safety*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Learning Representations*.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Context-aware fairness evaluation and mitigation in llms. *arXiv preprint arXiv:2510.18914*.

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*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 708–718.

Kezban Dilek Onal, Ye Zhang, Ismail Sengor Altingovde, Md. Mustafizur Rahman, Pinar Karagoz, Alexander 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*, 21(2–3):111–182.

Kailai Sun, Zuchao Li, and Hai Zhao. 2022. Reorder and then parse, fast and accurate discontinuous constituency parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10575–10588.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Yuxin Wang, Zhengyang Liu, Jinsong Su, Zhiyuan Wang, Xu Han, Lijie Wu, Yutao Hou, and Lichao Wang. 2023. O-edit: Orthogonal subspace editing for language model sequential editing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13144–13156.

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 & Knowledge Management*, pages 2128–2137.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. Prada: Practical black-box adversarial attacks against neural ranking models. *ACM Transactions on Information Systems*, 41(4):Article 89.

Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2015. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 113–122.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold

Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Ming Yan, Chenliang Li, Bin Bi, Wei Wang, and Songfang Huang. 2021. A unified pretraining framework for passage ranking and expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4555–4563.

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

Zeping Yu and Sophia Ananiadou. 2025. Understanding and mitigating gender bias in llms via interpretable neuron editing. *arXiv preprint arXiv:2501.14457*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.