Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2025.050109
ISSN 2790-0983 Vol. 5, Issue 1: 87-98

Optimization of Generative AI Intelligent Interaction System Based on Adversarial Attack Defense and Content Controllable Generation

Yuxin Wu

College of Engineering, Carnegie Mellon University, Moffett Field 94035, CA, United States

Email:yuxinwu202507@163.com

Keywords: Information retrieval; Neural Text Sorting Model; Robustness assessment; Adversarial information manipulation; Random mask smoothing

Abstract: This article focuses on the robustness research of information retrieval models, constructing a multi scenario (grey box/black box) and multi strategy (character/word/phrase level perturbation) adversarial evaluation system, proposing a knowledge guided grey box attack method (KnowAttack) and a black box evaluation framework based on transfer substitution models. At the defense level, the system optimizes model reliability by integrating empirical strategies (feature defense, adversarial training) and provable strategies (RobustMask random mask smoothing). Theoretical contributions include: 1) expanding the framework for evaluating machine learning robustness; 2) Compare the robustness differences between traditional bag of words models and neural ranking models; 3) Building a bridge between information manipulation theory and retrieval technology. At the practical level, we will improve the full scenario evaluation system and propose a system defense strategy of "knowing oneself and knowing the enemy" (data augmentation, adversarial training, etc.) to effectively reduce the risk of malicious information manipulation. Research promotes the transformation of information retrieval models from "usable" to "reliable and trustworthy", providing empirical evidence for AI governance. In the future, provable defense methods such as multimodal models, user cognitive impact analysis, and causal reasoning will be expanded.

1. Introduction

As a core technology of artificial intelligence, information retrieval has evolved from Boolean retrieval to neural text sorting models. Although pre trained models have improved performance, the problem of susceptibility to adversarial perturbations is becoming increasingly prominent. There are shortcomings in the existing research on the correlation between robustness evaluation frameworks, defense strategies, and information manipulation theories. This article proposes the construction of a multi scenario and multi strategy robustness evaluation system, and proposes optimization strategies to promote the transformation of information retrieval models towards "reliable, trustworthy, and responsible". Its contribution expands the machine learning robustness evaluation system and provides reference for model improvement. In the field of e-commerce, Anyi Chen pointed out that under the trend of diversified and personalized consumer demands, traditional demand analysis methods (such as RFM model) lack a systematic classification framework. To this

end, Anyi Chen innovatively combines text mining with the IPA-KANO model to construct a hierarchy of requirements analysis system. Through data-driven mining of explicit and potential user needs, they are classified into basic needs, necessary needs, and attractive needs, and their importance and satisfaction impact are evaluated. This method breaks through traditional limitations and achieves deep integration of text mining and IPA-KANO model for the first time, providing methodological references for industries such as fresh agricultural products and intelligent hardware. Interdisciplinary technological innovation is deeply exploring model robustness, decision scientificity, and system reliability: Saillenfest et al. proposed a density matching method for nonlinear concept erasure; Wang improves the accuracy of fuzzy AHP decision-making by using trapezoidal fuzzy pairwise comparison matrix; The hybrid supervised fine-tuning technology of long and short thinking chains developed by Yu et al. stimulates the reasoning ability of large-scale language models; Zillmann et al.'s complex valued asymmetric weighted overlapping addition filtering scheme suppresses channel interference; Madhavi et al. improved the QoS of wireless sensor networks by integrating ELECTRE with bipolar fuzzy PROMOTHEE. These studies collectively reflect the emphasis on interdisciplinary technology integration in the fields of artificial intelligence and communication, providing theoretical support and practical references for complex system optimization and application innovation.

2. Correlation theory

2.1. Research on the Robustness of Information Retrieval Models from the Perspective of Adversarial Information Manipulation

This study focuses on the robustness of information retrieval models in the context of artificial intelligence governance, and systematically explores their evaluation and improvement methods from the perspective of adversarial information manipulation. The core research objectives include two aspects: firstly, to construct a multi strategy (character/word/phrase level operation) robustness evaluation system covering white box, gray box, and black box scenarios, and comprehensively detect model vulnerabilities; Secondly, based on the evaluation results, propose optimization strategies at both the empirical and theoretical levels to reduce the risk of information manipulation and promote the transformation of information retrieval technology towards a "reliable, trustworthy, and responsible" direction.In terms of methodology, the study clarifies the definitions and strategies of three threat models: white box scenarios assume that attackers have complete control over the model architecture, parameters, and training data, and reveal vulnerabilities in the worst-case scenario through carefully designed perturbations; Combining knowledge distillation and Learning to Rank ideas in grey box scenarios, a transfer replacement model evaluation method and a Pairwise attack generation strategy for anchored documents are proposed to balance prior knowledge and actual attack capabilities; The black box scenario simulates general information manipulation through semantically coherent attack text generation (such as KnowAttack method) to evaluate the robustness of the model in the absence of internal information. To improve robustness, systematic strategies are proposed from two perspectives: empirical (feature defense, learnable detection, adversarial training) and theoretically provable (RobustMask random mask method), revealing the applicability and limitations of different methods.The results and advantages are reflected in: expanding the robustness evaluation of machine learning to the field of information retrieval at the theoretical level, comparing the robustness differences between traditional bag of words models and neural text sorting models, and building a bridge between information manipulation theory and information retrieval models; At the practical level, the comprehensive evaluation system for all scenarios is improved, and optimization strategies such as data augmentation, adversarial training, and random mask smoothing are proposed to effectively reduce the risk of malicious information

manipulation, providing empirical evidence for model improvement; Using case studies such as WebGPT and WebGLM to illustrate the downstream task risks that may arise from insufficient robustness of information retrieval models (such as malicious enhancement of candidate options), and to strengthen the practical significance of the research. Limitations include: the grey box evaluation data scenarios (disease themes, illegal pharmacy promotion) are relatively limited and need to be extended to areas such as illegal gambling, pornography, politics, as well as platforms such as Quora and Reddit; Manual evaluation did not fully incorporate factors such as text consistency, logical correctness, and user cognitive behavior; The defense methods that do not integrate detection and correction prediction, and the comprehensive defense framework need to be explored. Future research will expand to dense retrieval, full-text retrieval, and multimodal scenarios, deepen user cognitive impact analysis, and improve theoretically provable defense methods (such as causal inference) to enhance attack type coverage and defense efficiency.

2.2. The Supporting Role of Machine Learning and Information Retrieval Theory in Model Robustness

Machine learning theory provides core methodological support for the robustness research of information retrieval models: supervised learning establishes input-output mappings through labeled data, guiding model training, attack detector design, and high-quality adversarial sample generation; Unsupervised learning utilizes clustering, dimensionality reduction, and other techniques to provide a feature extraction and threshold partitioning framework for attack detection in unlabeled scenarios; Comparative learning utilizes the contrastive loss mechanism of "bringing similar samples closer and pushing dissimilar samples farther" to generate pseudo labeled data for optimizing ranking model training (such as the Pairwise method), while also improving the search generation efficiency of adversarial text in black box scenarios; Combining knowledge distillation and transfer learning with the Learning to Rank approach, weakly supervised student model training is achieved through relative order relationship transformation, balancing performance and efficiency. The theory of information retrieval has solidified the foundation of the model: Learning to Rank transforms sorting problems into machine learning tasks, and its Pointwise (regression/classification), Pairwise(relative order optimization), and Listwise (direct optimization evaluation index) methods provide objective functions and loss design references for robustness evaluation; Deep learning based models (such as DSSM, DRMM) and pre trained models (BERT, RoBERTa) improve sorting performance through query document representation learning and interaction mechanisms. The MonoBERT (query document concatenation input) and DuoBERT (dual document relative comparison) architectures provide specific carriers for analyzing the impact of adversarial perturbations. The integration of Information Manipulation Theory (IMT) and Social Engineering Theory further deepens the research dimension: IMT guides the design of micro manipulation strategies (such as character/word/phrase level operations and false association construction) by violating the four communication principles of quality, quantity, relevance, and conciseness (such as false information, topic shift, and vague expression); Social engineering utilizes human cognitive limitations and psychological weaknesses (such as persuasion principles such as reciprocity, authority, and social identity) to optimize the directionality and semantic coherence of attack texts through attack chain design such as information collection, relationship building, and trust utilization (such as generating targeted attack prompts by combining internal knowledge of large models). The collaboration between the two not only deepens the understanding of actual vulnerabilities in information retrieval systems, but also provides theoretical basis for the design of evaluation systems (white box to black box scenarios) and defense methods (such as detection templates and random masks), promoting robustness research from the technical level to

the social behavior intervention level, and jointly constructing a full process research framework from attack generation, vulnerability detection to defense optimization.

3. Research method

3.1. Research on Robust Evaluation Method of Grey Box Information Retrieval Model Based on Knowledge Guidance

This study proposes a novel evaluation method based on knowledge guidance to address the key issue of gray box scenarios in the robustness assessment of information retrieval models. In the gray box scenario, although attackers cannot obtain all the information of the model (such as architecture and parameters), they can use some of the transmitted information (such as input-output details and correlation scores) to manipulate it. Existing research mostly focuses on image retrieval or white box/black box scenarios, with insufficient exploration of gray box robustness in text retrieval models. Traditional attack methods often rely on gradient or scoring guidance, ignoring contextual knowledge between the model and corpus, resulting in low fluency and easy detection of attack texts. With the development of Large Language Models (LLMs), their knowledge reserves and task understanding capabilities provide new ideas for attack design. This study, for the first time, combines large models such as ChatGPT to infer key nodes related to the target query through the Chain of Thought method, and then uses the relevance ranking ability of the grey box target model to screen high relevance nodes. Finally, the target query is integrated with document paragraphs to generate smooth and natural attack perturbation text. The text needs to meet three objectives simultaneously: improving the ranking of the target document, possessing camouflage to evade detection, and content that naturally does not arouse user suspicion. The experimental results show that this method can effectively improve the sorting of target documents and carry manipulation information, while avoiding detection mechanisms and successfully detecting model adversarial security vulnerabilities. This study fills the gap in the robustness evaluation of text retrieval models in gray box scenarios. By combining the knowledge guidance ability of large models with the characteristics of gray box scenarios, it provides a more practical and efficient attack generation framework for robustness evaluation, promoting the full process research of information retrieval models from theoretical vulnerability detection to practical defense optimization.

3.2. Robustness Assessment and Ethical Framework for Grey Box Scenarios

This chapter focuses on the gray box attack scenario and systematically explores for the first time the robustness evaluation method of information retrieval models in this scenario. In gray box scenarios, although attackers cannot obtain model architecture, parameters, or training data, they can obtain ranking results and candidate document relevance scores by calling the target model infinitely, and use this information to construct attacks. The attack target is defined as reversing the original sorting result (such as changing from $s(q, pi) > s(q, pj)$ to $s(q, pi) < s(q, pj)$) by inserting a trigger containing key target information (to maintain semantic coherence to avoid detection), while ensuring the concealment of the trigger. The robustness evaluation is based on the criterion of "adversarial failure": if the model can still maintain the original correct ranking after the attack ($s(q, pi) > s(q, pj)$), it is considered robust. The experiment used the MiniLM-L-12 model fine tuned with MS MARCO and TREC DL datasets as the evaluation object. This model was selected as a representative paragraph sorting model due to its excellent performance in public rankings. On an ethical level, the research follows the ACM Code of Ethics to avoid attacks on real systems (such as Wikipedia) and only conducts controllable testing on local models. The aim is to promote the development of defense algorithms through public vulnerabilities, rather than causing

actual harm. This process is analogous to the vulnerability disclosure mechanism of white hat hackers.

3.3. Experimental design and result analysis of robustness evaluation of grey box scene information retrieval model

This study focuses on the robustness evaluation of information retrieval models in gray box scenarios, and constructs a full process experimental framework including datasets, benchmark models, experimental evaluation indicators, and prompt engineering. The effectiveness of the proposed method is verified through experiments. In terms of the dataset, with the promotion of illegal online pharmacies as the target, 125 pharmacies in the Concocted Pharma dataset were selected as attack targets, and 445 query terms (average length 1.75) were generated based on Wikipedia disease-related terms. The search results were reordered based on the target ranking model and divided into three subsets: Top (top 10 items, 1960 items), All (randomly sampled 10 items, 1960 items), and Tail (last 5 items, 890 items). The benchmark model adopts two comparative methods: heuristic attack method Query+(query by keyword stacking and concatenation, target pharmacy and drug information) and white box gradient guided method Collision (generate semantically unrelated but model judged relevant conflicting content). The experimental evaluation indicators include: ranking improvement success rate (ASR, defined as the proportion of adversarial modifications that successfully improve ranking, with a focus on the top 10 items),% top-10/% top-20 (the proportion of target paragraphs entering the top 10/20), detection rate (% Detected, the proportion of anomaly detector recognition based on fine-tuning the RoBERTa base model, with a testing accuracy of 87.6%), topic relevance rate (% Topic Rel, the proportion of adversarial modifications with a similarity of more than 0.33 to the original paragraph topic), and average perplexity (avg. PPL, a content fluency indicator calculated by the Wikicorpus fine-tuning language model).In terms of prompt engineering, a three-stage Prompt instruction is designed: disease-related query screening (extracting disease-related queries from 3000 candidate terms), related node queries (associating drugs through thought chain reasoning to screen top-2 candidates), and target information insertion fusion (naturally integrating pharmacy names and drug information into the original paragraph to maintain content coherence). The experiment generates adversarial text through small-scale batch processing and ChatGPT API calls to verify its ranking improvement, concealment, and semantic consistency effects.

Table 1 Performance Comparison of Grey Box Attack Methods in Information Retrieval Models

Search Result Set	MethodASR↑ (%)	%top-10↑ (%)	%top-20↑ (%)
TopQuery+	31.6	64.3	91.3
Collision	12.7	44.5	82.1
KnowAttack	21.1	39.6	66.9
TailQuery+	99.9	24.8	51.3
Collision	58.3	18.3	35.4
KnowAttack	96.3	25.6	49.5
AllQuery+	68.2	37.3	64.6
Collision	22.1	13.9	37.2
KnowAttack	59.0	28.9	55.6

Query+achieved the highest ASR (31.6%) and% top-10 (64.3%) on the Top dataset, but had poor

concealment (% Detected reaching 86.9%); Collision has the worst performance across corpora/models due to semantic conflicts with the target model (ASR only 12.7%); KnowAttack has an ASR (96.3%) close to Query+on the Tail dataset, and its overall stealthiness is superior to other methods (% Detected lowest at 81.7%,% Topic Rel highest at 89.3%). The perplexity distribution is close to the original text, and there are fewer grammar errors. Overall, KnowAttack achieves a better balance between attack effectiveness and stealthiness through the guidance of large model knowledge and thought chain reasoning, effectively detecting adversarial vulnerabilities in information retrieval models in gray box scenarios.

4. Results and discussion

4.1. Robust evaluation method for black box information retrieval model based on knowledge transfer

In response to the robustness evaluation requirements of information retrieval models in black box scenarios, this study proposes a knowledge transfer based attack framework that generates transferable adversarial text by constructing an imitation substitute model to detect adversarial vulnerabilities in the target model. In black box scenarios, attackers can only obtain the ranking result list of the target model through input queries, and cannot access its internal structure, parameters, or training data, making traditional white box methods ineffective. Existing research mostly focuses on image retrieval or classification models, while the black box robustness evaluation in the field of text retrieval still needs to be explored - the data types (continuous vs discrete) and encoding mechanisms of images and texts are significantly different, making it difficult to directly transfer attack methods. Inspired by knowledge distillation and the transferability of adversarial samples, this study combines the Learning to Rank approach to construct an imitation surrogate model to approximate the behavior of the target model, thereby generating attack text and verifying its transfer effect. Specifically, the attack process is divided into two stages: Stage 1 uses a black box model to simulate and construct an alternative model, and Stage 2 generates attack text based on Pairwise perturbations of anchored documents. In threat modeling, given a query q and a candidate document set P, the sorting model outputs a sorted list$p_1 \succ p_2 \succ \cdots \succ p_k$the attack target is to insert trigger t into the documentp_j, causing it to rank higher than p_i (i.e $s(q, p_j') > s(q, p_i)$) .Simultaneously maintaining semantic consistency and concealment of triggers, attackers can only obtain the ranking results of the target model through queries and cannot access internal information. Sampling positive/negative triplets from the top-N ranking results of the target model during the black box model imitation phase（$[q, p_i, p_j]$ marked as 1

$[q, p_j, p_i]$mark as 0) build a dataset and train PairLM (a BERT architecture based Pairwise encoder) to mimic the ranking preferences of the target model. PairLM concatenates queries with candidate options and outputs correlation scores through the linear layer (Formula 1 $s_i = W \cdot [CLS]q, p_i, p_j[SEP]$)

,the loss function is cross entropy(Formula 2

$$\mathcal{L} = -\sum y_{m,i,j} \log(s_i/(s_i + s_j))$$

The trained PairLM can approximate the sorting behavior of the target model. In the Pairwise perturbation generation stage based on anchored documents, given the query q, target document, and anchored document, triggers are generated through gradient optimization and bundle search. The optimization objective is combined with sorting relevance score, language model fluency, and semantic consistency (Formula 3:

$$\text{argmax}_x[\lambda_1 s_{pos} + \lambda_2 f_{nsp}]$$

The trigger is generated through softmax and bundle search to ensure naturalness and attack effectiveness. The experiment selected BERT base and MiniLM-L-12 as target models to verify the effectiveness and attack transfer effect of the imitation substitution model. This method provides a new framework for evaluating the robustness of text retrieval models through knowledge transfer in black box scenarios, balancing attack effectiveness and concealment requirements.

4.2. Model experiment

This study proposes a robust evaluation framework for black box information retrieval models based on knowledge transfer, and verifies its effectiveness through multi dataset experiments. The experiment used four datasets: MS MARCO DEV (a sparse label dataset based on Bing queries), TREC DL 2019 (43 queries with graded correlation labels), NQ (an external dataset based on Wikipedia), and TREC MB 2014 (a social media short text dataset). The benchmark models included BM25, TK, BERT Base/Large, DistilBERTCAT, MiniLM, and our study's PairLM (a BERT based truncated Pairwise encoder). The experiment revolves around three core issues: sampling strategy (such as the comparison between "top-25+others" and "top-15+others"), the impact of pre training data (domain specific MS MARCO and domain specific NQ), and the influence of model architecture differences (such as BERT base and MiniLM) on ranking similarity. Key indicators include MRR@10 and, &, as well as, plus NDCG@10 and, &, as well as, plus Inter@10 (Top 10 overlap rates) and RBO@1K (Weighted ranking overlap).

Table 2 Performance comparison of MiniLM imitate model under 1000 triplet sampling

Model Type	Sampling Strategy	MS MARCO DEV (MRR@10/NDCG@10)	TREC DL 2019 (MRR@10/NDCG@10)	Inter@10	RBO@1K
MiniLM (V2) Target Model	-	39.7/45.6	90.1/74.3	-	-
Zero Imitate V2	Top15+others	35.7/41.9	83.6/68.4	74.1	57.4
Zero Imitate V2	Top20+others	36.1/42.3	87.4/69.0	74.4	57.0
Zero Imitate V2	Top25+others	36.9/43.2	86.2/70.8	75.2	62.7
Zero Imitate V2	Top29+others	37.1/43.5	86.8/70.4	75.8	63.2
MS (ID) ↓ Imitate V2	Top15+others	37.3/43.5	85.3/70.1	74.8	58.1
MS (ID) ↓ Imitate V2	Top20+others	36.5/42.7	87.6/70.6	76.5	63.2
MS (ID) ↓ Imitate V2	Top25+others	37.3/43.6	86.8/71.6	77.2	63.4
MS (ID) ↓ Imitate V2	Top29+others	37.2/43.5	89.5/71.4	75.8	62.3
NQ (OOD) ↓ Imitate V2	Top15+others	36.0/42.1	88.3/70.5	74.4	57.5
NQ (OOD) ↓ Imitate V2	Top20+others	37.3/43.6	88.1/71.0	76.7	61.5
NQ (OOD) ↓ Imitate V2	Top25+others	36.7/42.9	88.5/69.7	74.9	60.4
NQ (OOD) ↓ Imitate V2	Top29+others	37.0/43.3	88.8/71.7	76.5	60.7

Table 2 shows the performance of the MiniLM imitate model under 400 triplet samples. The results showed that the "top-25+others" strategy performed the best: for example, the model pre trained based on NQ (OOD) achieved the best performance on TREC DL 2019 MRR@10 =89.9 Inter@10 =77.0 RBO@1K =66.1, Approaching the Target Model (MiniLM V2) MRR@10 =90.1.

Table 3 Performance Comparison of Multiple Models in MS MARCO DEV and TREC MB 2014

model name	MS MARCO DEV (MRR@10/NDCG@10)	TREC MB 2014 (AP/P@30)
BM25	18.7/23.4	41.4/62.6
TK	33.1/38.4	-/-
BERT-Base	35.2/41.5	45.4/68.0
BERT-Large (V1)	37.1/43.3	44.9/67.4
DistilBERT	38.2/44.2	44.6/66.5
MiniLM (V2)	39.7/45.6	47.5/70.9
Pairwise BERT (ID)	34.4/40.5	41.9/65.2
Imitate ID's Triples	32.6/39.0	41.4/65.2
Zero imitate V1	35.7/41.7	39.5/63.1
ID→Imitate V1	36.2/42.4	45.0/67.5
OOD→Imitate V1	36.2/42.5	42.6/65.4
Zero imitate V2	37.0/43.1	41.3/65.1
ID→Imitate V2	38.4/44.5	45.9/68.4
OOD→Imitate V2	*38.3/44.5*	*45.2/68.2*

Table 3 compares the results of 1000 triplet samples, although the performance is slightly lower than that of 400 triplet samples (such as the OOD pre trained model on TREC DL 2019) MRR@10 =88.5 Inter@10 =74.9), but still shows strong imitation ability.

4.3. Effect analysis

This study focuses on improving the robustness of information retrieval models. In response to the significant achievements and adversarial vulnerabilities of neural network ranking models (NRMs) in information retrieval tasks, an empirical based method framework is proposed to systematically explore robustness enhancement strategies. Although neural network sorting models can capture complex patterns to achieve high-quality sorting, they are vulnerable to information manipulation attacks (such as adding query content, semantically conflicting text, etc.), which can lead to performance degradation and threaten information security. Current research mainly focuses on attack detection (white box, gray box, black box), with few defense methods and a lack of systematic comparison. Therefore, this study starts from an empirical perspective and constructs a framework that includes feature-based defense, learnable adaptive defense, and adversarial training based defense, taking into account attack detection and prediction correction. The framework for improving the robustness of information retrieval models based on empiricism demonstrates the research approach: feature-based methods use perplexity TF-IDF、 Unsupervised methods such as linguistic features are used to detect attack samples; Learnable adaptive methods for training detection models (such as fine-tuning RoBERTa) to recognize normal and attack texts; Adversarial training enhances the model's resistance to attacks by introducing discrete (such as Collision, PAT) or continuous (PGD) adversarial samples. The experimental design revolves around three research questions: (1) Can adversarial training improve model robustness? (2) What is the performance of feature-based detection? (3) What is the detection capability of the learnable detector? Using datasets such as MSMARCO PASSAGE DEV, TREC DL 2019, and COLA, with BERT base as the target model, attack samples such as Collision (Aggressive/Natural variant) and PAT were

generated to test the effectiveness of different defense methods.

Table 4 Comparison of Adversarial Training and Attack Robustness in TREC DL 2019

Method	TREC DL 2019 (NDCG@10/MRR@10)	Attack (Aggressive) - ASR/%r≤100/avg.B	Attack (Natural) - ASR/%r≤100/avg.B
BERT	70.9/87.1	100/12.2/671.4	100/13.2/673.6
adv+Caggr	68.5/81.8	50.2/0.0/11.4	91.2/23.4/451.5
adv+Cnat	66.2/76.4	90.2/0.5/75.7	79.0/3.4/179.4
adv+PAT	62.5/73.1	84.4/0.0/75.9	88.3/15.1/354.3
adv+PGD	66.9/82.7	96.6/1.9/192.1	85.8/6.8/262.8

Table 4 compares the robustness improvement effects of various adversarial training methods. Adversarial training can significantly reduce the success rate of attacks (ASR), but has strong specificity. For example, adv+Caggr has the best defense effect against Caggr attacks (ASR=50.2%), but its effectiveness against Cnat attacks decreases (ASR=91.2%); Adv+PGD (continuous perturbation) has a generalized defense ability against both Caggr and Cnat attacks (ASR is 96.6% and 85.8%, respectively), but the overall effect is weaker than targeted discrete adversarial training. In addition, adversarial training can lead to a decrease in the ranking performance of the model on normal data (such as adv+Cnat) NDCG@10 The decrease from 70.9 to 66.2 is due to the interference of adversarial samples in semantic correlation judgment.

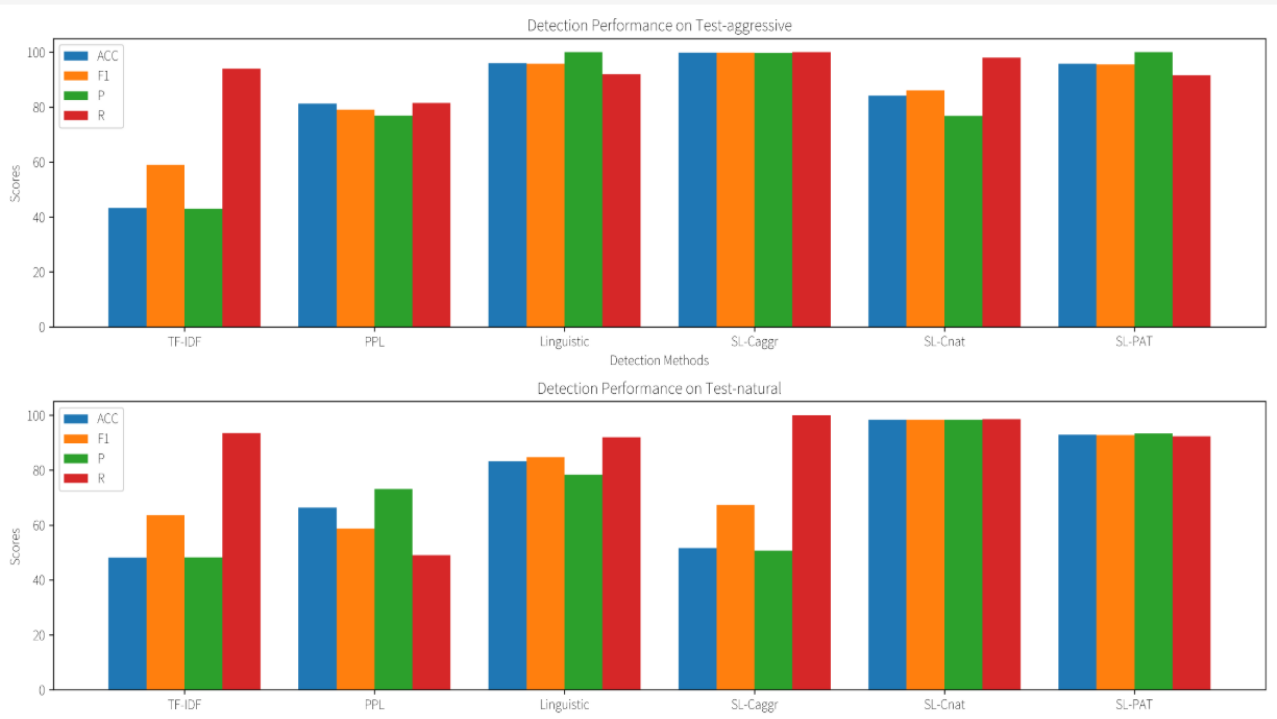

Figure 1 Performance comparison of different detection methods on attack/natural test sets

The comparison of detection performance between feature-based detector and supervised fine-tuning detector in Figure 1 shows that different detection methods have significant specificity: TF-IDF has a good detection effect on keyword stuffing (Query+) (ACC=50.7%), but a poor effect on Caggr (ACC=43.3%); Confusion detection is effective for Caggr (ACC=81.3%), but its effectiveness decreases for Cnat (ACC=66.3%) and PAT (ACC=58.5%); The Linguistic feature detector is optimal for Caggr detection (ACC=96.0%), but its effectiveness is weakened for Cnat

(ACC=83.3%) and PAT (ACC=78.7%). In the learnable detector (SL), SL-PAT has strong generalization detection ability for attribute based attacks (such as Caggr, Cnat, PAT) due to the inclusion of contextual information (ACC is 99.8%, 98.4%, 93.2%, respectively), while SL Caggr and SL Cnat are only effective for attacks of the same type, and the detection quality of other attacks decreases. The experimental conclusion shows that although empirical methods can improve robustness, they have limitations: feature-based and learnable detectors are difficult to handle all types of attacks, adversarial training requires targeted samples and may lead to a decrease in ranking performance. In the future, it is necessary to explore provable robustness methods to ensure the reliability of model predictions within specific disturbance ranges through theoretical modeling and statistical verification.

5. Conclusion

This study is based on the perspective of adversarial information manipulation, and systematically conducts research on the robustness of information retrieval models, focusing on the two core issues of "how to evaluate" and "how to improve". In terms of evaluation, model vulnerabilities are detected through information manipulation attack methods in gray box and black box scenarios: the KnowAttack method guided by knowledge is proposed in gray box scenarios, which utilizes large models to generate semantically coherent attack texts; Combining knowledge distillation and Learning to Rank ideas in black box scenarios, an evaluation method based on transfer substitution model and a Pairwise attack generation method anchored to documents are proposed, effectively improving the attack amplitude. In terms of enhancing robustness, systematic strategies are proposed from two perspectives: empirical (feature defense, learnable detection, adversarial training) and theoretically provable (RobustMask random mask method), revealing the effectiveness and limitations of different methods. In terms of theoretical contribution, we have constructed a robustness research framework from the perspective of adversarial information manipulation, extended machine learning robustness evaluation to the field of information retrieval, systematically compared the robustness differences between traditional bag of words models and neural text sorting models, and built a bridge between information manipulation theory and information retrieval models, promoting new breakthroughs in the robustness research of neural text sorting models; In terms of practical contributions, we will improve the robustness evaluation system for all scenarios (white box, gray box, black box), provide empirical evidence for model improvement, propose optimization strategies based on experience and theory, reduce the risk of malicious information manipulation, promote the supervision and governance of artificial intelligence services based on information retrieval, and promote the transformation of technology towards reliability, trustworthiness, and responsibility. There are still limitations in the research: the data scenarios in grey box evaluation (disease themes, illegal pharmacy promotion) are relatively limited and need to be expanded to areas such as illegal gambling, pornography, politics, as well as platforms such as Quora and Reddit; The granularity of manual evaluation is relatively coarse, and insufficient consideration has been given to text consistency, logical correctness, and user cognitive behavior factors; The defense methods that do not integrate detection and correction prediction, and the comprehensive defense framework need to be explored. Future work will expand research objects to models such as dense retrieval and full-text retrieval, and validate and optimize evaluation and improvement methods; Research on the robustness of multimodal (audio, video) information retrieval models; Deeply explore the impact of robustness on user cognition (emotional tendencies, decision-making behavior) and text credibility; Improve the robustness enhancement methods that can be proven by theory (such as causal inference), expand the coverage of attack types, and enhance defense efficiency and completeness.

References

[1] *Saillenfest A, Lemberger P. Nonlinear Concept Erasure: a Density Matching Approach[J]. 2025.*

[2] *Tang X, Wu X, Bao W. Intelligent Prediction-Inventory-Scheduling Closed-Loop Nearshore Supply Chain Decision System[J]. Advances in Management and Intelligent Technologies, 2025, 1(4).*

[3] *Madhavi S, Praveen R, Jagatheswari S, et al. Hybrid ELECTRE and bipolar fuzzy PROMOTHEE-based packet dropping malicious node mitigation technique for improving QoS in WSNs[J]. International Journal of Communication Systems, 2025, 38(2). DOI:10. 1002/dac. 5974.*

[4] *Wu X, Bao W. Research on the Design of a Blockchain Logistics Information Platform Based on Reputation Proof Consensus Algorithm[J]. Procedia Computer Science, 2025, 262: 973-981.*

[5] *Z Zhong. AI-Assisted Workflow Optimization and Automation in the Compliance Technology Field [J]. International Journal of Advanced Computer Science and Applications (IJACSA), 2025, 16(10): 1-5.*

[6] *Liu X. Emotional Analysis and Strategy Optimization of Live Streaming E-Commerce Users Under the Framework of Causal Inference[J]. Economics and Management Innovation, 2025, 2(6): 1-8.*

[7] *Lai L. Risk Control and Financial Analysis in Energy Industry Project Investment[J]. International Journal of Engineering Advances, 2025, 2(3): 21-28.*

[8] *Chen X. Research on Architecture Optimization of Intelligent Cloud Platform and Performance Enhancement of MicroServices[J]. Economics and Management Innovation, 2025, 2(5): 103-111.*

[9] *Yuan S. Application of Network Security Vulnerability Detection and Repair Process Optimization in Software Development[J]. European Journal of AI, Computing & Informatics, 2025, 1(3): 93-101.*

[10] *Sun Q. Research on Accuracy Improvement of Text Generation Algorithms in Intelligent Transcription Systems[J]. Advances in Computer and Communication, 2025, 6(4).*

[11] *Su H, Luo W, Mehdad Y, et al. Llm-friendly knowledge representation for customer support[C]//Proceedings of the 31st International Conference on Computational Linguistics: Industry Track. 2025: 496-504.*

[12] *Liu Y. Blockchain Future in Cloud Computing: The Challenges to Implement Blockchain Technology in Cloud Computing[J]. Journal of Computer, Signal, and System Research, 2025, 2(5): 15-23.*

[13] *Zhang K. Research on the Application of Homomorphic Encryption-Based Machine Learning Privacy Protection Technology in Precision Marketing[C]//2025 3rd International Conference on Data Science and Network Security (ICDSNS). IEEE, 2025: 1-6.*

[14] *Li W. Building a Credit Risk Data Management and Analysis System for Financial Markets Based on Blockchain Data Storage and Encryption Technology[C]//2025 3rd International Conference on Data Science and Network Security (ICDSNS). IEEE, 2025: 1-7.*

[15] *Li, W. (2025). Discussion on Using Blockchain Technology to Improve Audit Efficiency and Financial Transparency. Economics and Management Innovation, 2(4), 72-79.*

[16] *Lai L. Data-Driven Credit Risk Assessment and Optimization Strategy Exploration[J]. European Journal of Business, Economics & Management, 2025, 1(3): 24-30.*

[17] *Yan J. Research on Application of Big Data Mining and Analysis in Image Processing[J]. Pinnacle Academic Press Proceedings Series, 2025, 2: 130-136.*

[18] Xiu L. Research on the Design of Modern Distance Education System Based on Agent Technology[J]. Pinnacle Academic Press Proceedings Series, 2025, 2: 160-169.

[19] Lu, C. (2025). The Application of Point Cloud Data Registration Algorithm Optimization in Smart City Infrastructure. European Journal of Engineering and Technologies, 1(1), 39-45.

[20] Zhu, Z. (2025). Cutting-Edge Challenges and Solutions for the Integration of Vector Database and AI Technology. European Journal of AI, Computing & Informatics, 1(2), 51-57.