Towards Robustness Analysis of E-Commerce Ranking System

Ningfei Wang
University of California, Irvine
Irvine, USA
ningfei.wang@uci.edu

Yupin Huang
Amazon
Palo Alto, USA
huayupin@amazon.com

Han Cheng
Amazon
Palo Alto, USA
chenghan@amazon.com

Jiri Gesi
Amazon
Palo Alto, USA
jirigesi@amazon.com

Xiaojie Wang
Amazon
Palo Alto, USA
xiojie@amazon.com

Vivek Mittal
Amazon
Palo Alto, USA
vivekmit@amazon.com

ABSTRACT

Information retrieval (IR) is a pivotal component in various applications. Recent advances in machine learning (ML) have enabled the integration of ML algorithms into IR, particularly in ranking systems. While there is a plethora of research on the robustness of ML-based ranking systems, these studies largely neglect commercial e-commerce systems and fail to establish a connection between real-world and manipulated query relevance. In this paper, we present the first systematic measurement study on the robustness of e-commerce ranking systems. We define robustness as the consistency of ranking outcomes for semantically identical queries. To quantitatively analyze robustness, we propose a novel metric that considers both ranking position and item-specific information that are absent in existing metrics. Our large-scale measurement study with real-world data from e-commerce retailers reveals an open opportunity to measure and improve robustness since semantically identical queries often yield inconsistent ranking results. Based on our observations, we propose several solution directions to enhance robustness, such as the use of Large Language Models. Note that the issue of robustness discussed herein does not constitute an error or oversight. Rather, in scenarios where there exists a vast array of choices, it is feasible to present a multitude of products in various permutations, all of which could be equally appealing. However, this extensive selection may lead to customer confusion. As e-commerce retailers use various techniques to improve the quality of search results, we hope that this research offers valuable guidance for measuring the robustness of the ranking systems.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Robustness; Ranking system; Measurement study; Metric

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore.

ACM ISBN 979-8-4007-0172-6/24/05…$15.00
https://doi.org/10.1145/3589335.3648335

ACM Reference Format:
Ningfei Wang, Yupin Huang, Han Cheng, Jiri Gesi, Xiaojie Wang, and Vivek Mittal. 2024. Towards Robustness Analysis of E-Commerce Ranking System. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589335.3648335

1 INTRODUCTION

Information retrieval (IR) [13] is a crucial task in various applications such as Web search [24], etc. Unlike other domains, IR is unique in leveraging ranking algorithms to prioritize the relevance of retrieved resources. Therefore, the development of effective ranking systems has consistently occupied a central position in IR. Machine learning (ML) technologies have brought about significant breakthroughs in various longstanding tasks, such as natural language processing (NLP) [50]). Therefore, the integration of ML into ranking systems has emerged as a compelling and imperative trend.

However, ML generally lacks robustness [26], with inherent vulnerabilities to adversarial manipulations [5, 33, 74]. This problem has ignited considerable concerns regarding the robustness and reliability of ranking systems [47]. Thus, recent research has focused on ranking robustness by generating adversarial examples to evaluate the model's response disparities [15, 28, 30, 31]. However, these works have limitations: ignore evaluation in commercialized ranking systems, which is generally more robust than ML models [12, 44, 53], and fail to substantiate the relevance between real-world and manipulated queries. To overcome the first limitation above, we perform the first systematic investigation into the robustness of a leading commercialized e-commerce ranking system. To address the second limitation mentioned above, we use millions of historical search and ranking data sourced from anonymous users in real-world scenarios to assess the system's robustness.

To initiate the measurement study mentioned above, establishing a formulation for robustness is imperative. Existing formulations highly rely on ground-truth data [7, 28], which is labeled in the existing dataset by the human judge or approximation from user data. However, in our case, where we leverage real-world data, obtaining reliable and precise ground truth can be challenging or even undesirable, given the vast and diverse range of selections available and nuanced preferences of customers. Therefore, we define ground-truth orthogonal robustness in commercial e-commerce ranking systems as *the consistency of ranking outcomes for semantically identical queries*. This notion gains significance for the following

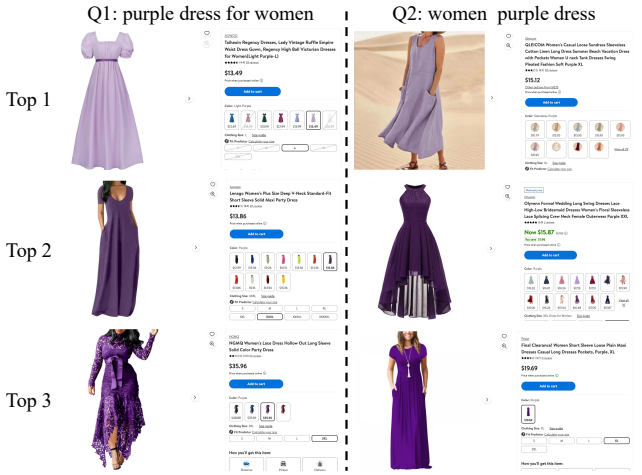

Figure 1: Motivating examples: the two queries are semantically same but their top 3 ranking results are totally different.

reason: a lack of robustness can severely compromise model accuracy, which is defined as the distance between model prediction and real ground truth [41]. For instance, low robustness manifests as inconsistent ranking results for semantically identical queries. This inconsistency indicates that the model prediction of one of the queries is likely to exhibit a significant departure from the ground truth. Such non-robust behavior has the potential to undermine user experience and thus reduce the customer's confidence in trusting the ranking results from the e-commerce systems.

To quantitatively analyze the robustness defined above, an appropriate metric is necessary. With a comprehensive review of the prior metrics, we have identified noteworthy shortcomings in all existing metrics. They tend to equally penalize errors occurring at any position within the ranking list and struggle to handle cases where items appear in one list but not the other. With these, we propose a novel metric to assess the disparity between two ranking lists, considering both the position and item-specific information.

We conduct the first large-scale measurement study on the ranking system in one of the leading e-commerce retailers, with our newly proposed metric and historical real-world data. We study more than several million queries and group them together as query pairs by their semantics with rule-based normalization techniques [3] and predictive modeling method [17]. An example is shown in Fig. 1, where we analyze a pair of semantically identical queries: 'purple dress for women' and 'women purple dress'. Notably, the top 3 ranking results were entirely different. From a model training perspective, this finding implies that at least one query is likely to yield rankings divergent from the ground truth due to semantic consistency but ranking inconsistency. Our large-scale study reveals new and open research opportunities to improve robustness of ranking system, as it frequently generates divergent ranking outcomes for semantically identical queries. The issue of robustness discussed herein does not constitute an error or oversight. Instead, in situations characterized by a wide range of options, it is possible to offer multiple products in various configurations, each appealing to different customers. Nevertheless, this abundance of choices may cause consumer confusion.

To ensure the semantic identity of our evaluated query pairs, we conducted a user study. The results of this evaluation indicate that a significant majority of the query pairs analyzed in our study are indeed perceived as semantically identical by human participants. In order to systematically explore the robustness of the ranking system, we also present a comprehensive taxonomy delineating various patterns observed in non-robust query pairs, the temporal evolution of robustness, and the integration of robustness into commercialized models. More details are demonstrated in §4.

Building upon the insights garnered from our measurement study, we propose several solutions and outline future research directions to improve ranking system robustness. Specifically, we investigate the potential of Large Language Models (LLMs) and model ensembles as avenues for enhancements in the robustness of e-commerce ranking systems. We hope that the insights will serve as valuable guidance for the design of e-commerce ranking systems, facilitating enhancements in both their robustness and accuracy.

To sum up, this paper makes the following contributions:

- We conduct the first large-scale measurement study on the robustness analysis of a leading commercialized e-commerce ranking system, utilizing millions of real-world data from users.
- We propose a novel metric designed to quantitatively assess the disparity between two ranking lists for the ranking systems.
- We systematically measure the robustness of the e-commerce ranking system by addressing five research questions. Additionally, we present ten measurement observations that could guide future robustness and accuracy improvement of ranking systems.

2 RELATED WORK

Text-based model robustness. Recent works find that ML models are generally not robust in different NLP tasks [34, 36, 46, 63, 68, 70, 72]. In response, a plethora of studies have endeavored to enhance their robustness, such as adversarial training [18, 29, 52, 75] and certified robustness [16, 20, 23, 45, 54, 64, 67]. Given that ranking systems share the ML nature, they inherit similar robustness challenges, which can impact user experiences and harm a company's reputation. However, currently, there is a limited number of research studies dedicated to investigating the robustness issues of ranking systems [7, 28], especially on the robustness improvement. Thus, we conduct the first large-scale measurement study to assess the robustness of commercialized e-commerce ranking systems.

Ranking metrics. To evaluate the performance of ranking models, a wide range of metrics, e.g., normalized discounted cumulative gain (NDCG) [19, 61] and expected reciprocal rank (ERR) [6], have been introduced [55, 57]. However, these conventional metrics rely on ground-truth data, making them unsuitable for the problem settings in this paper. Furthermore, metrics such as Kendall's τ [21], τ_{AP} [69], τ_a [49], τ_b [49], and Spearman's rank correlation coefficient [43] have been developed to quantify the disparity between two ranking lists [56]. However, these existing metrics exhibit limitations, including the equal penalization of errors across all positions in the list and inadequate handling of items that appear in one list but are absent in the other. We provide detailed discussion in §3. Based on their drawbacks, we introduce a novel metric specifically tailored to evaluate the robustness of ranking systems in e-commerce settings, which will be detailed in §3.

3 METHODOLOGY

3.1 Ranking Model

Ranking models are essential for ordering candidate items by relevance to a query. Traditional methods such as BM25 [42] struggle to model human language effectively. Recent advancements in ML [50] have led to ML-based models emerging as state-of-the-art solutions for various ranking tasks[8, 10, 71]. Nonetheless, these ML-based ranking models exhibit robustness issues [28, 30, 31, 60, 62]. Notably, the literature has not thoroughly evaluated the robustness of commercial e-commerce ranking systems or delineated the distinction between real-world and manipulated queries. Commercial ranking systems, such as Google or Amazon, often exhibit greater robustness than standalone ML models [12, 44, 53]. For example, these systems usually include query rewrite, auto-completion, and search facet to create a smooth search journey, which can potentially avoid the robustness problem from tweaking characters [27] in the input. Thus, it is unclear how robust the commercial e-commerce ranking models are with real-world queries. Thereby, we perform the first measurement study on the commercial e-commerce ranking system robustness with real-world queries to fill in these research gaps.

We provide the formulation of the ranking system. Given a query q, the ranking systems generate a list of ranked items, denoted as $<i_1, i_2, \cdots, i_n>$, along with their respective positions in the list, denoted as $<p_1, p_2, \cdots, p_n>$. Typically, these positions are determined based on the relevance scores predicted by the ranking models [28]. These results are subsequently presented to users through systems.

3.2 Robustness Scope and Motivating Example

Robustness scope. We define robustness in e-commerce ranking systems as the *consistency of ranking results between two semantically identical queries*. Under an ideal scenario with a ranking system that is both robust and accurate, two distinct but semantically equivalent queries would yield identical ranking outcomes. Conversely, a system that produces different ranking results for such queries—despite all items being relevant—is considered non-robust. Such a problem is critical considering the following scenarios. If ground truth for the ranking model is available, the lack of robustness can compromise the model's accuracy when compared to the ground truth, owing to the inconsistency in ranking outcomes; inevitably, one of the queries is likely to deviate significantly from the ground truth. Moreover, considering that e-commerce models are generally trained on user behavior data, non-robust outcomes generated by the model could induce users to engage in suboptimal behaviors. This, in turn, could result in biased data, thereby affecting the subsequent training and performance of the model.

Motivating example. We consider the two semantically identical queries: Q_1 as 'purple dress for women' and Q_2 as 'women purple dress' in Fig. 1. When searching on one of the leading commercialized e-commerce retailers, the ranking lists for these queries were completely dissimilar. Fig. 1 shows a comparison of the top 3 ranking results for both queries. This stark contrast in ranking outcomes for semantically equivalent queries suggests that there is at least one query (Q_1 or Q_2) for which the model predictions differ from the ground truth, despite the queries being semantically equivalent. This observation underscores the fundamental issue studied in this paper, which we refer to as model robustness.

3.3 Robustness Study Method

To address the robustness problem in §3.2, the initial and indispensable step is to identify queries that contain different forms or alternative expressions of the same concept. We explore two types of data in this paper: 1) Rule-based normalization query pair, and 2) Predictive modeling query pair.

Rule-based Normalization Data. We leverage query normalization techniques [35] that include tokenization, filtering, and stemming to consolidate the queries. Further, we observe that shopping intent queries are typically concise. Reordering the tokens within these queries usually does not alter the underlying concept. Thus, in Text Processing and Sorting (TPS) data, queries sharing the same normalized tokens are considered as semantically identical query pairs, such as "battery AA" and "AA battery".

Predictive Modeling Data. To improve the behavior feature with query cluster and feature coverage by mapping queries to the closest query, Q2Q (in-house trained query to query similarity prediction) models are proposed with collected similar queries and created labels [17]. Q2Q can provide similarity scores between the two query pairs, and thus, can be used in our measurement study.

For 1), our approach begins by collecting vast data of query pairs. With that, we select a subset in which the queries exhibit the same TPS. The selection process is crucial, as it ensures that the retained query pairs share a meaningful level of semantic equivalence. In 2), we adopt the following procedure: one query designated as query 1 is held constant, and we then identify the top k query 2 candidates with the highest similarity scores to create query pairs. Then, we employ these query pairs to retrieve historical data from e-commerce systems. This process allows us to identify the ranked items with their unique identifiers and average ranking positions.

3.4 Metric Design

As introduced in §2, most representative metric designs such as Kendall's τ and τ_{AP} suffer from several drawbacks within the context of our study: 1) equally penalizing errors that occur at any position in the list, which fails to fully consider the order information; 2) difficulty in handling cases where items appear in one list but not in the other. Kendall's τ is particularly deficient in both aspects, while τ_{AP} is mainly affected by the latter issue.

For instance, regarding the first drawback 1), consider three ranking lists: R1: <1, 2, 3, 4>, R2: <1, 2, 4, 3>, and R3: <2, 1, 3, 4>. Given the significance of item position in ranking lists, R1 and R2 should intuitively be more similar than R1 and R3. However, when calculating Kendall's τ, we obtain the same result of 0.67 for both comparisons, highlighting the limitation in addressing this first drawback. In the context of the second drawback 2), we examine another three ranking lists: R1: <1, 2, 3, 4>, R2: <1, 2, 3, 4>, and R3: <1, 2, 5, 6>. R1 and R2 should be more similar than R1 and R3 since R1 and R2 are exactly the same. However, both Kendall's τ and τ_{AP} yield values of 1.00 for these comparisons. This experimental analysis underscores the limitations of the representative existing metric in handling this second drawback. These limitations are not unique to these two metrics but also affect other state-of-the-art ranking metrics, such as Spearman's rank correlation coefficient [43]. These limitations hinder the applicability of these metrics in our context.

Addressing the second limitation requires a method to account for missing items in ranked lists. One straightforward approach is

to append these missing elements to the end of the list. Integrating this method with specific metrics, such as τ_{AP}, may offer a solution to both identified defects. However, this direct appending method introduces unique corner cases, wherein the appended items create discrepancies in position that can result in counterintuitive errors. Consider the evaluation of three distinct ranking lists: R1: <1, 2, 3, 4>, R2: <1, 2, 4, 3>, and R3: <1, 2, 5, 6>. The τ_{AP} value stands at 0.33 for R1 and R2. When we apply the appending strategy to R1 and R3, the process entails calculating the τ_{AP} between the lists <1, 2, 3, 4, 5, 6> and <1, 2, 5, 6, 3, 4>. Notably, this also yields a τ_{AP} of 0.33, identical to the value derived from comparing R1 and R2. Intuitively, the similarity between R1 and R2 should be larger than that between R1 and R3. Thus, this equivalence in τ_{AP} values underscores an inherent flaw in the appending approach. Thereby, the weakness of this straightforward strategy renders it unsuitable for addressing the nuances of our problem context.

With limitations observed in existing metrics, we propose a novel metric named RDS (Ranking Distance Score) defined in Eq (1) that addresses these issues. When comparing two ranking lists, R_1 and R_2, RDS incorporates a position-based decay using a logarithmic function and also checks for the presence of items in both lists.

$$\sum_{r \in R_1 \cup R_2} RDS(p(r, R_1), p(r, R_2)) \tag{1}$$

where the function $p(*)$ is to retrieve the position of an item in the ranking list and returns -1 if the item does not exist in the ranking list. The function $RDS(*)$ is to calculate the score defined in Eq (2).

$$RDS(p_1, p_2) = \begin{cases} |\frac{1}{\log_2(p_1+1)} - \frac{1}{\log_2(p_2+1)}|, & p_1 \neq -1 \ \wedge \ p_2 \neq -1 \\ |\frac{1}{\log_2(2)} - \frac{1}{\log_2(p_{1_{max}}+1)}| + \frac{1}{\log_2(p_1+1)}, & p_1 \neq -1 \\ |\frac{1}{\log_2(2)} - \frac{1}{\log_2(p_{2_{max}}+1)}| + \frac{1}{\log_2(p_2+1)}, & p_2 \neq -1 \end{cases} \tag{2}$$

where the $p_{*_{max}}$ represents the maximum position in the ranking list $*$ (either 1 or 2). The term $|\frac{1}{\log_2(2)} - \frac{1}{\log_2(p_{*_{max}}+1)}|$ is to address the second drawback, which is as a penalty term to overcome the limitation of handling items that appear in one list but not the other.

RDS can effectively address the limitations mentioned above For instance, consider three ranking lists: R1: <1, 2, 3, 4>, R2: <1, 2, 3, 4>, and R3: <1, 2, 5, 6>. After normalizing the scores to a range of 0 to 1 and calculating the similarity score (i.e., 1 minus the RDS), we find that the score between R1 and R2 is 1.00, while the score between R1 and R3 is 0.38. For the case: R1: <1, 2, 3, 4>, R2: <1, 2, 4, 3>, and R3: <1, 2, 5, 6>, the similarity score between R1 and R2 is 0.95, while the similarity score between R1 and R3 is 0.38. These simulated results clearly demonstrate that our metric outperforms existing ones even the existing metric with a straightforward design in effectively capturing the differences in ranking list similarity.

4 MEASUREMENT STUDY

4.1 Methodology and Setup

Data preparation. This study utilizes two distinct datasets: 1) a collection of real-world historical search and ranking data retrieved weekly from the ranking system, and 2) a dataset derived from the in-house Q2Q model. The first dataset comprises several billion data points that facilitate the generation of TPS data pairs, as introduced in §3.3, and aid in constructing evaluation sets for ranking lists. These lists comprise items and their corresponding average positions given a specific query. Within this context, we have generated over 26 million query pairs, their relevance determined based on whether each pair shares identical TPS. The second dataset, encompassing roughly several billion query pairs, allows for the generation of more than 4 million query pairs. This process is achieved by holding one query constant and pairing it with the top k queries based on the highest similarity scores within the Q2Q model. We assume that both TPS and Q2Q model similarity scores can offer a semantically identical guarantee to a certain degree, which will be further explored in §4.4. All data used in this study is sourced from real-world users, lending a high degree of real-world applicability to our findings. Note that all the data is anonymous, and no other potentially private data is included.

Data filter. To enhance the quality of the data, we apply a series of filters to eliminate noise. First, we restrict our data originating from the United States and written in English to control regional and linguistic biases. Second, we eliminate the bottom 20% of query-item pairs based on their searched frequency from users within the week to remove low-frequency errors that could affect ranking. Third, we consider only those cases where the ranking list length exceeds 20, specifically the top 20 ranking results. Last, for query lists sharing the same TPS, we include only the top three queries based on weekly search count to control the bias of query frequency.

Large-scale measurement study. We perform our large-scale evaluation using PySpark [11] on the Amazon EMR cluster. The study focuses on the e-commerce system, generating ranked lists from historical real-world anonymous user query data.

4.2 Research Question

We first pose several research questions (RQs) to assess the robustness and semantic consistency of e-commerce ranking systems:

- RQ1: How robust is the e-commerce ranking system?
- RQ2: To what extent can semantic consistency be ensured?
- RQ3: How has the robustness of the e-commerce ranking system evolved over time?
- RQ4: How is robustness incorporated into the in-house Q2Q model used for the e-commerce ranking system?
- RQ5: What types of query pairs challenge the robustness of the e-commerce ranking system?

4.3 RQ1: Robustness of E-Commerce System

We employ RDS metrics to evaluate the robustness of the e-commerce ranking system, utilizing both TPS and Q2Q data types described in §4.1. Our findings, illustrated in Fig. 2, reveal substantial discrepancies in the ranking outcomes provided by the e-commerce ranking system for most query pairs in both data types. Notably, the Q2Q data exhibits a frequency rate for RDS between 0.9 and 1.0 that is more than twice as high as that in the TPS data. This suggests that the Q2Q data imposes fewer constraints on the semantics of the query pairs, making it less robust compared to TPS. Thus, in the following analysis, we mostly use TPS data since it can help to understand how robust the current e-commerce system is with tight constraints on the semantics. With that, we can conclude:

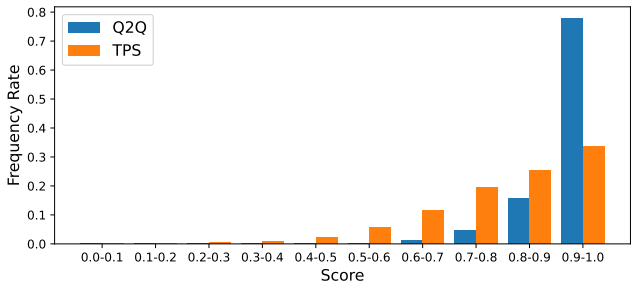

Figure 2: Histogram of TPS and Q2Q data evaluated with RDS metric on millions of query pairs.

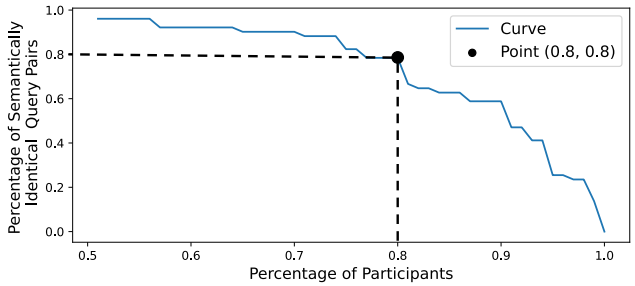

Figure 3: Illustration of participants and semantically identical query pairs for user study. Specifically, 80% of participants regarded 80% of the query pairs as semantically identical.

Observation 1

The current commercialized e-commerce ranking system displays a lack of robustness as it often produces divergent ranking outcomes for semantically identical queries.

4.4 RQ2: Semantic Consistency Guarantee

To systematically understand whether the two queries are semantically identical, we perform a user study.

Methodology and setup. We recruited 50 human subjects through Prolific [40], a platform specializing in research-related crowd-sourcing. All participants, ranging in age from 19 to 72, were verified to have experience shopping on one of the leading e-commerce retailers and proficiency in English. Each subject was presented with 50 sets of query pairs, randomly sampled from our full TPS data. We describe the shopping scenarios and ask the survey question: "in a hypothetical shopping scenario online, one might input two semantically identical queries into the search bar with the expectation of receiving identical search results. We hope that you can help to judge whether you are expecting to get the same results from e-commerce retailers giving two queries." Then, we provide different groups of query pairs and ask the users to provide binary answers, i.e., "Yes" or "No".

Results. As depicted in Fig. 3, our survey indicates that 80% of the participants perceive 80% of the query pairs to be semantically identical and expected e-commerce retailers to return identical results for these query pairs. This finding corroborates our initial hypothesis regarding semantic similarity among the query pairs. Based on the results, we can formulate the following observation:

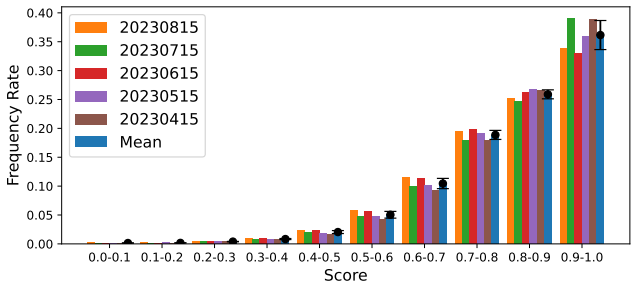

Figure 4: Histogram for RDS from TPS data over a five-month period with more than 20 million query pairs for each time.

Observation 2

The majority of query pairs are perceived as semantically identical by human subjects, strengthening the observation of a lack of robustness in the e-commerce ranking system.

4.5 RQ3: Evolution of Robustness

To address RQ3, we assess the evolution of e-commerce ranking system robustness over a period of four months, spanning from April 15, 2023, to August 15, 2023. During this period, we conduct evaluations on a weekly basis for one selected week each month, and the results are presented in Fig. 4. We use TPS data for analysis.

In Fig. 4, we also report key statistical measures, including the average score and the standard deviation (STD). The figure includes a detailed histogram illustrating these statistics. The frequency rate difference across bins is observed to be minimal. Thus, the STD values are notably low. Specifically, the maximum STD observed is 0.02, which is associated with an average normalized frequency rate of 0.35. Given the small magnitude of the STD, we conclude that the variance is negligible in comparison to the original values.

Observation 3

Our analysis indicates that the e-commerce system's robustness has remained stable over the observed period, suggesting that the robustness has not been considered to be improved.

4.6 RQ4: Robustness Integration on Q2Q

To investigate RQ4, we examine the relationship between the similarity scores utilized in in-house Q2Q models [17] and our RDS metric. We use the *Pearson correlation coefficient* metrics, as implemented in the Scipy Python library [51].

Our results indicate that the absolute value of the *Pearson correlation coefficient* is 0.345, thereby indicating a low correlation between the two metrics [38]. Additionally, a *p-value* of 0.0 confirms that the observed correlation is statistically significant [9]. Given that in-house Q2Q models are trained and evaluated based on similarity scores [17], it is evident that robustness considerations are largely absent from both the training and evaluation processes.

Observation 4

Our findings suggest that robustness factors are generally not incorporated into the training or evaluation of in-house Q2Q models in e-commerce ranking systems.

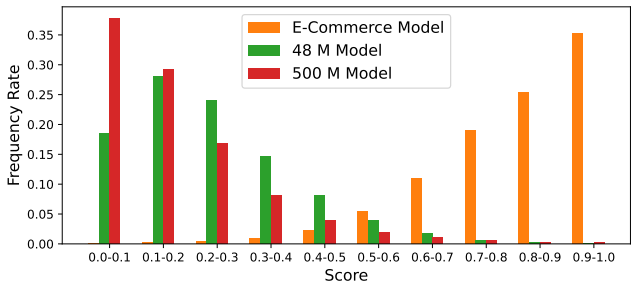

Figure 5: Comparison of histograms between the original e-commerce ranking model and two larger models on millions of TPS query pairs. Lower scores indicate greater robustness.

4.7 RQ5: Taxonomy of Non-Robust Cases

Addressing RQ5 is pivotal for systematically improving the robustness of current e-commerce ranking systems. To this end, we conduct a novel classification study using historical data from real users, focusing on the worst cases of which RDS is 1.0.

Taxonomy methodology. Due to large amounts of TPS query pairs, i.e., more than 20 million, it is impractical to manually look up the entire data and perform the taxonomy. Thus, we begin by randomly sampling 500 query pairs from the entire data and conducting manual classification. To validate our taxonomy, we randomly sample another 1,000 query pairs from the entire data as validation. We manually check whether the 1,000 query pairs can fit into the generated taxonomy. If not, we will include the new class in the existing ones and repeat the validation process. This process iterated until no new classes emerged, validating our taxonomy.

Taxonomy result. We taxonomize them into 8 categories (C1 to C8) summarized in Table 1. For each category, we provide a case study including the query pairs, where we search Query 1 and Query 2 on one of leading e-commerce retailers and compare the top 3 items. The analysis is as follows.

- C1: Rewriting phrases using prepositions with semantic integrity confuses the system such as 'T-shirt for men' vs. 'men T-shirt'.
- C2: The use of abbreviations such as 'volt' to 'v' or 'inch' to 'in' confuses the ranking algorithm.
- C3: Switching between singular noun and plural noun, e.g., 'T-shirt for man' vs. 'T-shirts for men', leads to divergent results.
- C4: Queries where changing word order doesn't alter the semantics but affects rankings, such as 'battery AA' vs. 'AA battery'.
- C5: Including articles such as 'the' in 'the heels' produces entirely different ranking lists, especially related to movies and books.
- C6: Similar to articles, punctuations mainly affect book and movie titles and thus the ranking results.
- C7: Different treatment of spaces, e.g., '1 mm ring' vs. '1mm ring', leads to ranking differences.
- C8: User or bot-generated queries might use different characters such as '+' or 'x' to link words, affecting the ranking results.

Observation 5

Our taxonomy categorizes semantically identical query pairs challenging the e-commerce ranking system's robustness into eight classes, highlighting areas needing improvement.

5 SOLUTIONS AND FUTURE DIRECTIONS

5.1 S1: Large Language Models

As shown in Table 1, our analysis reveals that the prevalent issues in current e-commerce ranking models largely arise from their limited semantic understanding of user queries. This limitation often leads these models to incorrectly differentiate between semantically identical query pairs. This, thus, results in significant ranking discrepancies as noted in our Observations 1 and 3. Thus, a promising improvement method for addressing these shortcomings involves the integration of Large Language Models (LLMs) [25]. These models excel in capturing both the semantic and syntactic nuances of queries, compared to the existing e-commerce ranking.

To substantiate this claim, we conduct preliminary tests using GPT-3.5 and GPT-4.0 [25]. We randomly sampled 50 semantically identical query pairs from each taxonomy in § 4.7, each of which received disparate rankings from the existing e-commerce system. The LLMs are then tasked with determining whether the pairs are indeed semantically identical. The outcomes are summarized in Table 2. Notably, GPT-4.0 achieved a 99.5% average accuracy in correctly identifying these pairs, while GPT-3.5 can provide 77.0% accuracy on average. Thereby, LLMs potentially enhance the query understanding component of current e-commerce ranking systems.

Preliminary tests with GPT models indicate the potential benefits of using LLMs for improving ranking robustness. However, the impact of these improvements in an end-to-end setting remains unclear. Thus, we conduct additional experiments on two in-house ranking models, comprising billions of data points. We then compare their performance with that of a current e-commerce ranking model (§4), which is less complex than the two in-house models.

Methodology and setup. We select two representative in-house language ranking models with sizes of 48 million and 500 million parameters, respectively. For evaluation, we select at least 2 million query pairs, including all types in Table 1. Our measurement methodology and experimental setup align with those in §4.1. Note that we do not differentiate among various categories during the evaluation due to the high labor costs. Instead, we conduct large-scale testing involving millions of queries, rendering our results comparable to those of e-commerce ranking systems in §4.

Result. The comparison results on these three models are shown in Fig. 5. The histogram clearly indicates that both large-parameter ranking models demonstrate enhanced robustness in comparison to the current e-commerce model. Notably, the model comprising 500M parameters shows approximately twice as many queries falling within the 0.0 - 0.1 RDS range (0.0: a perfect match) as compared to its 48M counterpart. In contrast, the e-commerce ranking model scarcely exhibits cases within this optimal range. Additionally, the e-commerce model manifests a high frequency of queries with RDS exceeding 0.9 (1.0: no match between two ranked lists), which is scarcely observed in the two large-parameter models. Thus, large models offer significant potential for enhancing robustness.

While the current LLM results show effectiveness in addressing robustness issues, it is crucial to consider the latency requirements for services, such as requiring millisecond-level responsiveness [1]. LLMs lead to substantial computational costs, necessitating a plethora of machines and GPUs [32]. Such overhead may be prohibitive for small companies due to budget constraints.

Table 1: Taxonomy and case study of robustness issues in e-commerce ranking system: The table classifies various types of query pairs and includes a case study with query pairs from one of leading e-commerce retailers.

ID	Taxonomy	Case study	
		Query 1	Query 2
C1	Preposition	purple dress for women	women purple dress
C2	Abbreviation	30" marble top	30 inch marble top
C3	Singular and plural	electric thing for kids	electric things for kids
C4	Word order	red watch	watch red
C5	Article	heels	the heels
C6	Punctuation	funding	funding.
C7	Space	24 x 20 outdoor cushion	24x20 outdoor cushion
C8	Words connection	black swing coat	black+swing+coat

Table 2: Evaluation of semantically identical query pairs using LLMs on various robustness issues in Table 1. Each class has 50 query pairs, and the LLMs answer 'Yes' or 'No'.

LLM	C1	C2	C3	C4	C5	C6	C7	C8	Average
GPT-3.5	76%	88%	36%	70%	68%	92%	98%	88%	77.0%
GPT-4.0	98%	100%	100%	100%	100%	100%	100%	98%	99.5%

Observation 6

Large Language Models offer the potential to improve the robustness of e-commerce ranking systems but their high computational and financial costs constrain universal applicability.

5.2 S2: Model Ensemble

We integrate model ensemble [66] during inference by leveraging a few ranking models. The objective of this approach is to refine prediction quality by using aggregation techniques such as majority voting [39]. In the context of e-commerce ranking systems, the final output can be determined by averaging the positions across various ranking lists. To validate this hypothesis, we conduct an evaluation.

We assume that the e-commerce ranking system undergoes monthly updates. To evaluate the system's performance over time, we use 1,000 semantically identical query pairs that persist across a five-month period, specifically from April to August 2023. Subsequently, we aggregate the results over this time to calculate the smoothed RDS and compare it with the RDS for a single month. This assesses the efficacy of the model ensemble during inference.

The detailed histogram of RDS is presented in Table 3. Specifically, in the most favorable case (i.e., a RDS range of 0.0–0.2), the rate of query pairs with model ensemble is 4.4 times higher than the rate observed without model ensemble. Conversely, in the least favorable case, the rate of query pairs with model ensemble is 1.2 times lower than that without model ensemble. These findings suggest that model ensemble enhances the robustness of the e-commerce ranking system during the inference stage, constituting a promising avenue for future research.

Observation 7

Model ensemble serves as a cost-effective method to improve the robustness of e-commerce ranking systems without requiring model training, offering a promising future direction.

Table 3: RDS histogram with and without ensemble.

	0.0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-0.1.0
No ensemble	0.5%	2.5%	10.0%	30.3%	57.7%
Ensemble	2.2%	4.7%	13.9%	32.1%	47.1%

5.3 S3: User Behavior Features Improvement

Based on the information in Table 1, different semantically identical queries can significantly impact the ranking outcomes in e-commerce systems with ML algorithms. Essentially, these variations induce disparities in the feature sets used for ranking, thus, altering the model's output [22]. State-of-the-art e-commerce ranking systems are heavily dependent on features derived from user behavior. However, these features are prone to various forms of bias. For example, users are constrained to select items ranked by the e-commerce system, which generally relies on historical user behavior data for training. Thus, this can inadvertently introduce position bias into the new model. A ranking system that lacks initial robustness risks entering a self-perpetuating cycle of non-robustness, where flawed user behavior data and ranking outcomes mutually reinforce each other's weaknesses. Therefore, introducing robustness during the training phase by refining the feature set can break the ill cycle and thus lead to significant improvements in the model's performance. While the details of our e-commerce ranking features cannot be disclosed due to legal constraints, we can discuss potential future directions.

Leveraging our framework and metrics, we can systematically and automatically identify query pairs that exhibit non-robust behavior. Armed with this, we can conduct controlled experiments by inputting these pairs into the ranking system and analyzing the feature differences. This is crucial for worst-case scenarios (RDS is 1.0) that may necessitate immediate remedial action in productions. Such experiments can elucidate which features are susceptible to robustness issues. The findings can guide the model training to improve robustness and accuracy. Additionally, our RDS could be integrated into training as a penalty, which is future research.

Observation 8

Our findings offer valuable insights for enhancing feature quality and interpretability in e-commerce ranking models, particularly in addressing worst-case predictive scenarios.

5.4 S4: DNN Robustness Improvement

The ongoing tug-of-war between adversarial attacks [5, 74] and their defenses [26, 65] has yielded many robustness improvement strategies, such as adversarial training [2]. Thus, we review the existing works to identify potential robustness improvement methodology that could enhance e-commerce ranking system robustness.

We categorize existing robustness improvement methods into two classes [4]: 1) model-level robustness improvements; and 2) input- or target-level robustness improvements. To the best of our knowledge, the majority of existing research focuses on 2), owing to its lightweight nature. These methods often include techniques such as automatic grammar checking [28], spam detection [7, 28], and frequency-guided word substitutions [37]. Notably, these methodologies are incorporated into commercial e-commerce systems. Our empirical analyses are based on evaluations conducted within a commercial environment, leading to the following observation.

Observation 9

While input- or target-level robustness improvement is commonly used in e-commerce systems, they fall short of effectively improving the robustness of the ranking models.

Due to the paucity of research on model-level robustness in ranking systems, we perform a comprehensive literature review in other domains such as image and NLP. Certified robustness [26] and adversarial training [14] are widely acknowledged as potent methodologies for enhancing model robustness. However, these approaches have not yet been explored within the sphere of ranking systems, especially those in e-commerce systems.

Adversarial training, which aims to improve model robustness by solving a min-max optimization problem, often adversely affects the performance of the normal or benign model [58, 59, 73]. Such a trade-off is typically unacceptable in commercial settings. Besides, the computational overhead associated with adversarial training is substantial [48], rendering it economically unfeasible for many organizations. Thus, improving the efficiency and performance of e-commerce ranking systems is an important future work.

Regarding certified robustness, although it offers theoretical guarantees of robustness, its current applications are limited to small models and simpler tasks, such as image classification [26]. This limitation is primarily due to the complexity of the methodology and the associated computational costs. Thereby, these potent robustness-improving methods are conspicuously absent in the current research on ranking models, which requires future efforts.

Observation 10

To the best of our knowledge, model-level robustness improvement in ranking models remains underexplored. Due to their potential for robustness enhancement and theoretical guarantees, these approaches offer a compelling future direction.

6 DISCUSSION

Other interesting problems. We observe two interesting types of issues: 1) incongruities in item ranking and 2) difficulties in handling negations. In the first category, the e-commerce ranking system displays a limited capacity for semantic discernment between queries.

Figure 6: Example of two semantically distinct queries sharing a same ranked item on the e-commerce retailer.

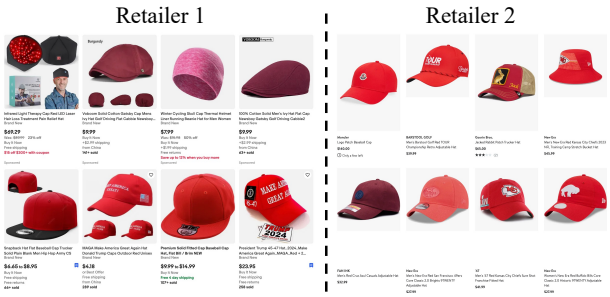

Figure 7: Search results for the query 'hat not red' on two commercialized e-commerce retailers.

This lack of semantic understanding leads to overlapping items in the ranking lists for semantically distinct queries. For example, queries like 'motor skate' and 'skate motor' yielded a shared ranked item depicted in Fig. 6. In the second category, *negation* poses a substantial challenge. A query such as 'hat not red' frequently results in the display of red hats, indicating a failure to understand the negation. Fig. 7 offers an example from two e-commerce retailers. These observations suggest that current e-commerce ranking systems have substantive issues, requiring further enhancement.

Limitations. First, its scope is confined to the U.S. market and English language on one e-commerce retailer. While its prominence and ubiquity lend a degree of generality to our findings, the results may not hold across different linguistic or cultural contexts. Second, there are potential biases that may affect the study, including data and ranking item bias. Despite efforts to mitigate these through data cleaning, residual biases may persist. Third, the ranking list is comprised of item hashing IDs, which are not semantically analyzed in this study. For instance, some items are very similar but use different IDs. Thus, future work could benefit from incorporating semantic understandings of these IDs into metrics and analyses.

7 CONCLUSION

In this paper, we perform a systematic exploration of the robustness of a leading commercialized e-commerce ranking system. We propose a novel metric for conducting large-scale measurement studies and presenting insights and solutions. Our investigation uncovers a general issue within e-commerce ranking systems: the lack of robustness in producing consistent ranking outcomes for semantically identical queries. We hope that our findings and insights can further inspire the research of e-commerce ranking systems to potently improve the robustness and accuracy.

8 ACKNOWLEDGMENTS

We would like to thank Qi Alfred Chen, Vamsi Salaka, Abbas Kazemipour, Hancao Li, Zhongruo Wang, Zhen Zhang, Junze Liu, Dan Luo, Ziwen Wan, Xinyang Zhang, Tong Wu, and the anonymous reviewers for their valuable and insightful feedback.

REFERENCES

[1] Eric Arrington. 2019. WHAT IS LATENCY AND HOW MUCH IS IT COSTING YOU. (2019). https://akfpartners.com/growth-blog/what-is-latency

[2] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356* (2021).

[3] Andrzej Białecki, Robert Muir, Grant Ingersoll, and Lucid Imagination. 2012. Apache lucene 4. In *SIGIR 2012 workshop on open source information retrieval*. 17.

[4] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 2021. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 176–194.

[5] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. Ieee, 39–57.

[6] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 621–630.

[7] Xuanang Chen, Ben He, Le Sun, and Yingfei Sun. 2023. Defense of Adversarial Ranking Attack in Text Retrieval: Benchmark and Baseline via Detection. *arXiv preprint arXiv:2307.16816* (2023).

[8] Xiaokai Chu, Jiashu Zhao, Lixin Zou, and Dawei Yin. 2022. H-ernie: A multi-granularity pre-trained language model for web search. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*. 1478–1489.

[9] Tukur Dahiru. 2008. P–VALUE, A TRUE TEST OF STATISTICAL SIGNIFICANCE? A CAUTIONARY NOTE. *Annals of Ibadan postgraduate medicine* 6, 1 (2008), 21–26.

[10] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 65–74.

[11] Tomasz Drabas and Denny Lee. 2017. *Learning PySpark*. Packt Publishing Ltd.

[12] Tommaso Dreossi, Daniel J Fremont, Shromona Ghosh, Edward Kim, Hadi Ravanbakhsh, Marcell Vazquez-Chanlatte, and Sanjit A Seshia. 2019. Verifai: A toolkit for the formal design and analysis of artificial intelligence-based systems. In *International Conference on Computer Aided Verification*. Springer, 432–442.

[13] M Rami Ghorab, Dong Zhou, Alexander O'connor, and Vincent Wade. 2013. Personalised information retrieval: survey and classification. *User Modeling and User-Adapted Interaction* 23 (2013), 381–443.

[14] Morgane Goibert and Elvis Dohmatob. 2019. Adversarial robustness via label-smoothing. *arXiv preprint arXiv:1906.11567* (2019).

[15] Prakhar Gupta and Yulia Tsvetkov. 2021. Synthesizing Adversarial Negative Responses for Robust Response Ranking and Evaluation. *Computational linguistics Association for Computational Linguistics* (2021).

[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 *Empirical Methods in Natural Language Processing (EMNLP)*. 4081–4091.

[17] Yupin Huang, Jiri Gesi, Xinyu Hong, Han Cheng, Kai Zhong, Vivek Mittal, Qingjun Cui, and Vamsi Salaka. 2023. Behavior-driven query similarity prediction based on pre-trained language models for e-commerce search. In *SIGIR 2023 Workshop on eCommerce*. https://www.amazon.science/publications/behavior-driven-query-similarity-prediction-based-on-pre-trained-language-models-for-e-commerce-search

[18] Maor Ivgi and Jonathan Berant. 2021. Achieving Model Robustness through Discrete Adversarial Training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1529–1544.

[19] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[20] Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified Robustness to Adversarial Word Substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4129–4142. https://doi.org/10.18653/v1/D19-1423

[21] Maurice G Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1/2 (1938), 81–93.

[22] Junho Kim, Byung-Kwan Lee, and Yong Man Ro. 2021. Distilling robust and non-robust features in adversarial examples by information bottleneck. *Advances in Neural Information Processing Systems* 34 (2021), 17148–17159.

[23] Ching-Yun Ko, Zhaoyang Lyu, Lily Weng, Luca Daniel, Ngai Wong, and Dahua Lin. 2019. POPQORN: Quantifying robustness of recurrent neural networks. In *International Conference on Machine Learning*. PMLR, 3468–3477.

[24] Mei Kobayashi and Koichi Takeda. 2000. Information retrieval on the web. *ACM computing surveys (CSUR)* 32, 2 (2000), 144–173.

[25] Anis Koubaa. 2023. GPT-4 vs. GPT-3.5: A Concise Showdown. (2023).

[26] Linyi Li, Tao Xie, and Bo Li. 2023. Sok: Certified robustness for deep neural networks. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1289–1310.

[27] Aiwei Liu, Honghai Yu, Xuming Hu, Shu'ang Li, Li Lin, Fukun Ma, Yawen Yang, and Lijie Wen. 2022. Character-level white-box adversarial attacks against transformers via attachable subwords substitution. *arXiv preprint arXiv:2210.17004* (2022).

[28] Jiawei Liu, Yangyang Kang, Di Tang, Kaisong Song, Changlong Sun, Xiaofeng Wang, Wei Lu, and Xiaozhong Liu. 2022. Order-Disorder: Imitation Adversarial Attacks for Black-box Neural Ranking Models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2025–2039.

[29] Kai Liu, Xin Liu, An Yang, Jing Liu, Jinsong Su, Sujian Li, and Qiaoqiao She. 2020. A robust adversarial training approach to machine reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8392–8400.

[30] 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*. 1647–1656.

[31] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Topic-oriented Adversarial Attacks against Black-box Neural Ranking Models. *arXiv preprint arXiv:2304.14867* (2023).

[32] Zhiye Liu. 2023. ChatGPT Will Command More Than 30,000 Nvidia GPUs: Report. (2023). https://www.tomshardware.com/news/chatgpt-nvidia-30000-gpus

[33] Chen Ma, Ningfei Wang, Qi Alfred Chen, and Chao Shen. 2024. SlowTrack: Increasing the Latency of Camera-Based Perception in Autonomous Driving Using Adversarial Examples. In *AAAI 2024*.

[34] Rishabh Maheshwary, Saket Maheshwary, and Vikram Pudi. 2021. A Strong Baseline for Query Efficient Attacks in a Black Box Setting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 8396–8409.

[35] C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[36] John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. *EMNLP 2020* (2020), 119.

[37] Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis D Griffin. 2020. Frequency-guided word substitutions for detecting textual adversarial examples. *arXiv preprint arXiv:2004.05887* (2020).

[38] Mavuto M Mukaka. 2012. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal* 24, 3 (2012), 69–71.

[39] Behrooz Parhami. 1994. Voting Algorithms. *IEEE transactions on reliability* 43, 4 (1994), 617–629.

[40] Prolific. 2019. Definitive human data to deliver world-leading research and AI. (2019). https://www.prolific.co/

[41] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. 2020. Understanding and Mitigating the Tradeoff between Robustness and Accuracy. In *International Conference on Machine Learning*. PMLR, 7909–7919.

[42] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer, 232–241.

[43] Philip Sedgwick. 2014. Spearman's rank correlation coefficient. *Bmj* 349 (2014).

[44] Junjie Shen, Ningfei Wang, Ziwen Wan, Yunpeng Luo, Takami Sato, Zhisheng Hu, Xinyang Zhang, Shengjian Guo, Zhenyu Zhong, Kang Li, et al. 2022. Sok: On the semantic ai security in autonomous driving. *arXiv preprint arXiv:2203.05314* (2022).

[45] Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. 2020. Robustness Verification for Transformers. In *International Conference on Learning Representations*.

[46] Walter Simoncini and Gerasimos Spanakis. 2021. SeqAttack: On adversarial attacks for named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 308–318.

[47] Congzheng Song, Alexander M Rush, and Vitaly Shmatikov. 2020. Adversarial Semantic Collisions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4198–4210.

[48] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, et al. 2021. Towards efficient and effective adversarial training. *Advances in Neural Information Processing Systems* 34 (2021), 11821–11833.

[49] Julián Urbano and Mónica Marrero. 2017. The treatment of ties in AP correlation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 321–324.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[51] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser,

Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[52] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[53] Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu, and Qi Alfred Chen. 2023. Does Physical Adversarial Example Really Matter to Autonomous Driving? Towards System-Level Effect of Adversarial Object Evasion Attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4412–4423.

[54] 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.

[55] Xiaojie Wang, Zhicheng Dou, Tetsuya Sakai, and Ji-Rong Wen. 2016. Evaluating search result diversity using intent hierarchies. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 415–424.

[56] Xiaojie Wang, Ruoyuan Gao, Anoop Jain, Graham Edge, and Sachin Ahuja. 2023. How well do offline metrics predict online performance of product ranking models?. In *Proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval*.

[57] Xiaojie Wang, Ji-Rong Wen, Zhicheng Dou, Tetsuya Sakai, and Rui Zhang. 2017. Search result diversity evaluation based on intent hierarchies. *IEEE Transactions on Knowledge and Data Engineering* 30, 1 (2017), 156–169.

[58] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2018. Kdgan: Knowledge distillation with generative adversarial networks. *Advances in neural information processing systems* 31 (2018).

[59] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Adversarial distillation for learning with privileged provisions. *IEEE transactions on pattern analysis and machine intelligence* 43, 3 (2019), 786–797.

[60] Y Wang, L Lyu, and A Anand. [n. d.]. BERT rankers are brittle: a study using adversarial document perturbations (2022). *DOI: https://doi. org/10.48550/ARXIV* 2206 ([n. d.]).

[61] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In *Conference on learning theory*. PMLR, 25–54.

[62] 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 (2023), 1–27.

[63] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. 2020. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing* 17 (2020), 151–178.

[64] Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. 2020. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems* 33, 1129–1141.

[65] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *NDSS*.

[66] Zhuolin Yang, Linyi Li, Xiaojun Xu, Bhavya Kailkhura, Tao Xie, and Bo Li. 2022. On the Certified Robustness for Ensemble Models and Beyond. In *International Conference on Learning Representations*.

[67] Mao Ye, Chengyue Gong, and Qiang Liu. 2020. SAFER: A Structure-free Approach for Certified Robustness to Adversarial Word Substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3465–3475. https://doi.org/10.18653/v1/2020.acl-main.317

[68] Muchao Ye, Chenglin Miao, Ting Wang, and Fenglong Ma. 2022. TextHoaxer: budgeted hard-label adversarial attacks on text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3877–3884.

[69] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 587–594.

[70] Lifan Yuan, Yichi Zhang, Yangyi Chen, and Wei Wei. 2021. Bridge the gap between cv and nlp! a gradient-based textual adversarial attack framework. *arXiv preprint arXiv:2110.15317* (2021).

[71] Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. 2018. Neural ranking models with multiple document fields. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 700–708.

[72] Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. OpenAttack: An Open-source Textual Adversarial Attack Toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. 363–371.

[73] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*. PMLR, 7472–7482.

[74] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*.

[75] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. FreeLB: Enhanced Adversarial Training for Natural Language Understanding. In *International Conference on Learning Representations*.