UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Robust Neural Information Retrieval

An Adversarial and Out-of-Distribution Perspective

Liu, Yu-An; Zhang, Ruqing; Guo, Jiafeng; de Rijke, Maarten; Fan, Yixing; Cheng, Xueqi

DOI
10.1145/3768153

Publication date
2026

Document Version
Final published version

Published in
ACM Transactions on Information Systems

License
CC BY

Link to publication

Citation for published version (APA):
Liu, Y.-A., Zhang, R., Guo, J., de Rijke, M., Fan, Y., & Cheng, X. (2026). Robust Neural Information Retrieval: An Adversarial and Out-of-Distribution Perspective. *ACM Transactions on Information Systems*, *44*(1), Article 17. https://doi.org/10.1145/3768153

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

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

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

Download date:08 Oct 2026

Robust Neural Information Retrieval: An Adversarial and Out-of-Distribution Perspective

YU-AN LIU, RUQING ZHANG, and JIAFENG GUO, CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China
MAARTEN DE RIJKE, University of Amsterdam, Amsterdam, The Netherlands
YIXING FAN and XUEQI CHENG, CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China

Recent advances in neural information retrieval models have significantly enhanced these models' effectiveness across information retrieval tasks. The robustness of these models, which is essential for ensuring their reliability in practice, has also garnered significant attention. With a wide array of research on robust information retrieval being published, we believe it is the opportune moment to consolidate the current status, glean insights from existing methodologies, and lay the groundwork for future development. Robustness of information retrieval is a multifaceted concept and we emphasize the importance of robustness against performance variance, out-of-distribution scenarios, and adversarial attacks. With a focus on out-of-distribution and adversarial robustness, we dissect robustness solutions for dense retrieval models and neural ranking models, respectively, recognizing them as pivotal components of the neural information retrieval pipeline. We provide an in-depth discussion of methods, datasets, and evaluation metrics, shedding light on challenges and future directions in the era of large language models. To accompany this survey, we release three additional resources: (1) a curated list of publications related to robust information retrieval, (2) a tutorial based on this

This work was funded by the Strategic Priority Research Program of the CAS under Grant No. XDB0680102, the National Natural Science Foundation of China (NSFC) under Grant Nos. 62472408, 62372431 and 62441229, the National Key Research and Development Program of China under Grant No. 2023YFA1011602, the Youth Innovation Promotion Association CAS under Grant No. 2021100, the Lenovo-CAS Joint Lab Youth Scientist Project, and the Strategic Priority Research Program of the CAS under Grant No. XDB0680301. This work was also (partially) funded by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union under grant agreement No. 101070212 (FINDHR) and No. 101201510 (UNITE). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of their respective employers, funders and/or granting authorities.
Authors' Contact Information: Yu-An Liu, CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China; e-mail: liuyuan21b@ict.ac.cn; Ruqing Zhang (corresponding author), CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China; e-mail: zhangruqing@ict.ac.cn; Jiafeng Guo(corresponding author), CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China; e-mail: guojiafeng@ict.ac.cn; Maarten de Rijke, University of Amsterdam, Amsterdam, The Netherlands; e-mail: m.derijke@uva.nl; Yixing Fan, CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China; e-mail: fanyixing@ict.ac.cn; Xueqi Cheng, CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China; e-mail: cxq@ict.ac.cn.

This work is licensed under Creative Commons Attribution International 4.0.

© 2025 Copyright held by the owner/author(s).
ACM 1558-2868/2025/11-ART17
https://doi.org/10.1145/3768153

survey, and (3) a heterogeneous benchmark for robust information retrieval, BestIR, that collects all known datasets for evaluating information retrieval systems for robustness. We hope that this study provides useful clues for future research on the robustness of information retrieval models and helps to develop trustworthy IR systems.

CCS Concepts: • **Information systems → Retrieval models and ranking**;

Additional Key Words and Phrases: Robustness, trustworthy systems

ACM Reference format:
Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2025. Robust Neural Information Retrieval: An Adversarial and Out-of-Distribution Perspective. *ACM Trans. Inf. Syst.* 44, 1, Article 17 (November 2025), 48 pages.
https://doi.org/10.1145/3768153

1 Introduction

Recently, with advances in deep learning, neural **information retrieval (IR)** models have witnessed significant progress [61, 63]. With the development of training methodologies such as pre-training [52, 118] and fine-tuning [84, 139, 193], neural IR models have demonstrated remarkable effectiveness in learning query-document relevance patterns. When deploying neural IR models, an aspect equally essential as their effectiveness is their robustness. A good IR system must not only exhibit high effectiveness under normal conditions but also demonstrate robustness in the face of abnormal conditions.

Why Is Robustness Important in IR? The natural openness of IR systems makes them vulnerable to intrusion, and the consequences can be severe. For example: (1) search engines are vulnerable to black-hat **search engine optimization (SEO)** attacks,[1] necessitating significant efforts to curb these infringements.[2] and (2) search engines are confronted with large amounts of unseen data on a daily basis. The working algorithm needs to be improved constantly to ensure that search effectiveness is maintained.[3]

Recently, research has begun to investigate the robustness of IR systems [29, 35, 100, 105, 167, 183]. As neural networks gain increasing popularity in IR, many studies have found that neural IR systems inherit a wide variety of problematic robustness issues from deep neural networks. In response, the field of robust neural IR is garnering increasing attention, as evidenced by the growing number of papers published on the topic annually, as depicted in Figure 1.[4] The robustness issues are differently represented in real IR scenarios and raise concerns about deploying neural IR systems in the real world. Therefore, the study of robust neural IR is crucial for building reliable IR systems.

How to Define Robustness in IR? User attention mainly focuses on the top-K results and increases with higher rankings [129]. Based on this, we argue that robustness in IR refers to the consistent performance and resilience on the top-K results of an IR system when faced with a variety of unexpected scenarios. Robustness is not a simple concept; it encompasses multiple dimensions, as illustrated by research within the **machine learning (ML)** community [154, 195]. In IR, we identify several facets of robustness:

[1] https://www.bleepingcomputer.com/news/security/15-000-sites-hacked-for-massive-google-seo-poisoning-campaign/.
[2] https://www.bbc.com/news/technology-28687513, https://developers.google.com/search/docs/essentials.
[3] https://developers.google.com/search/news.
[4] See Appendix A for a description of the protocol we followed to select the sources aggregated in Figure 1 and surveyed in this paper.

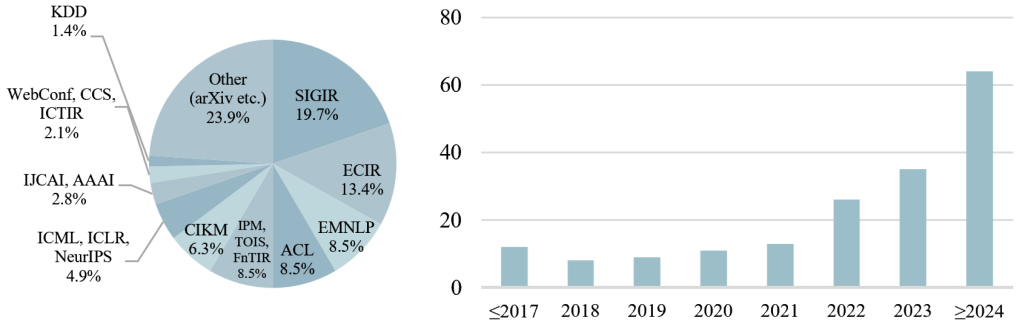

Fig. 1. Statistics of publications related to robust neural IR and covered in this survey. "Other" includes arXiv (mostly), TREC, ICDM, NAACL, and TACL.

(1) ***Independent and identically distributed (IID)*** *robustness* emphasizes the worst-case performance across different individual queries under the IID data assumption [183];
(2) ***Out-of-distribution (OOD)*** robustness refers to the generalizability of an IR model on unseen queries and documents from different distributions of the training dataset [167]; and
(3) *Adversarial robustness* refers to the ability of the IR model to defend against malicious adversarial attacks aimed at manipulating rankings [183].

In this survey, our focus is on adversarial robustness and OOD robustness, which have garnered significant attention. For adversarial robustness, studies primarily approach the topic from two angles, i.e., adversarial attacks and defense, to enhance the robustness of IR models. For OOD robustness, the emphasis is on improving the generalizability of IR models to both unseen documents and unseen queries. To study the above two aspects, we zoom in on two key components of the neural IR framework: first-stage retrieval and the subsequent ranking stage. We focus on **dense retrieval models (DRMs)** and **neural ranking models (NRMs)** to further explore the aforementioned research perspectives.

Relation to Other Surveys. There are several surveys on robustness in the fields of **natural language processing (NLP)** [174, 175] and **computer vision (CV)** [2, 43]. However, the field of IR presents its own unique characteristics: (1) unlike NLP, which often focuses on individual examples, IR concerns ranking a collection of documents, highlighting the need for robustness across the ranked lists, and (2) different from continuous image data in CV, IR deals with robustness related to discrete text documents. Consequently, the studies explored in these surveys are not directly transferrable as references within the IR field.

Surveys specific to the IR domain tend to concentrate on effectiveness in areas like pre-training [48, 198], ranking models [63], initial retrieval stages [61], and the explainability of IR systems [4]. Thus, there is a noticeable gap in the literature: a dedicated survey that consolidates and introduces research pertaining to robustness in IR is absent.

To complement this survey we release the following resources alongside it: (1) a curated list of publications related to robust IR, (2) a tutorial on robust IR, and (3) a benchmark that collects datasets for assessing the robustness of IR systems.

Contributions of This Survey. This article's contributions are as follows:

— A comprehensive overview and categorization: We define robustness in IR by summarizing the literature and further dividing it into distinct categories; we provide a curated list of publications to support this aspect of the survey.

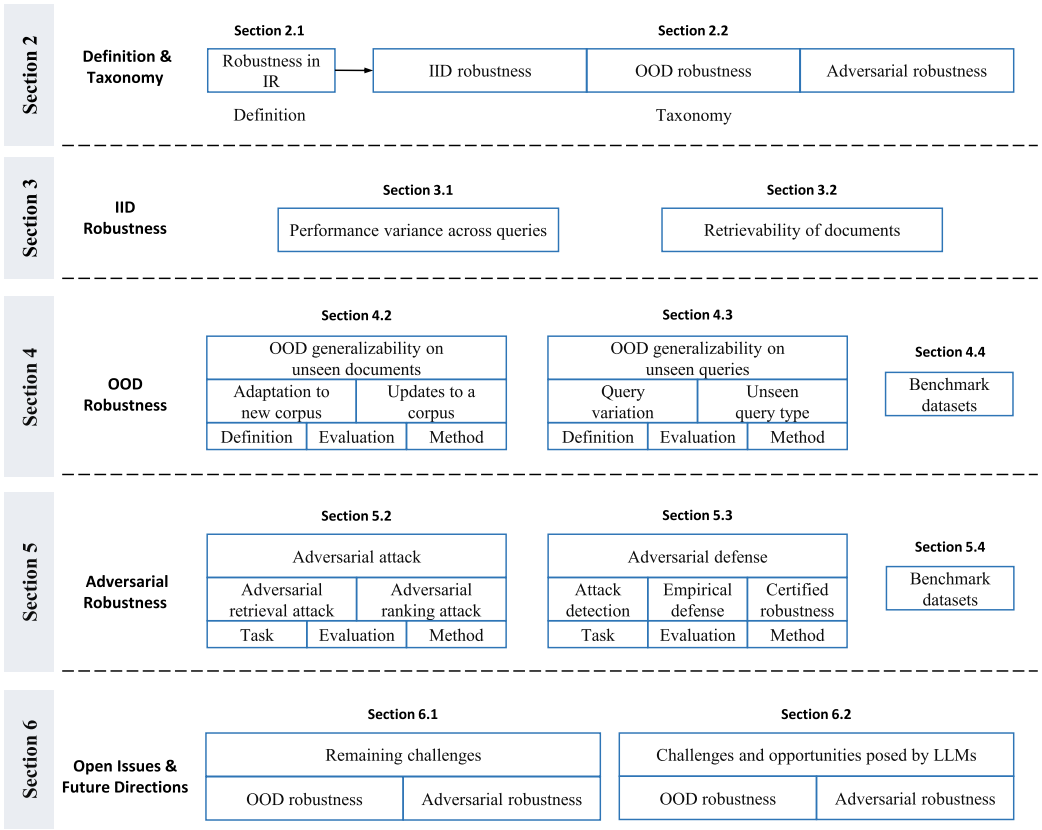

Fig. 2. Overview of the survey. Section 6 is only partially listed here because of space limitations.

—A detailed discussion of methodologies and datasets: We offer a detailed discussion of methodologies, datasets, and evaluation metrics pertinent to each aspect of robustness. We support this part of the survey by providing a tutorial and a benchmark, BestIR, which integrates the datasets mentioned in this survey to facilitate follow-up work.
—Identification of open issues and future trends: We highlight challenges and potential future trends, particularly in the age of **large language models (LLMs)**.

Organization. Figure 2 depicts the organization of our survey. In Section 2, we introduce the IR task, and give a definition and taxonomy of robustness in IR. In Section 3, we highlight two key challenges for IID robustness, i.e., performance variance across queries and retrievability of documents, and present specific methods for addressing these scenarios. In Section 5, we examine adversarial attack and defense tasks, alongside their respective datasets, evaluation criteria, and state-of-the-art methodologies. In Section 4, we show two key scenarios for OOD robustness, i.e., OOD generalizability on unseen documents and OOD generalizability to unseen queries, and present specific datasets, evaluation metrics, and methods for solving these scenarios. In Section 6, we describe remaining challenges and emerging opportunities for robustness of IR in the era of LLMs. Finally, Section 7 summarizes the survey and offers concluding remarks.

2 Definition and Taxonomy

In this section, we provide a formal definition of robustness in the context of IR and outline the taxonomy pertinent to this domain.

IR Task. To provide a clear understanding, we first formalize the ad-hoc IR task. Suppose that $R = \{r_1, r_2, \ldots, r_l\}$ is the set of relevance levels, where l denotes the number of levels. A total order exists among the relevance labels such that $r_l \succ r_{l-1} \succ \cdots \succ r_1$, where $\succ$ denotes the order relation. The minimum value of the relevance label is 0, which usually implies no relevance. Suppose that $Q = \{q_1, q_2, \ldots, q_m\}$ is the set of queries in a training dataset. Each query q_i is associated with a list of documents $D_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,N}\}$ and a list of relevance labels $Y_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,N}\}$, where $y_{i,j} \in R$ denotes the label of document $d_{i,j}$ and N is the document list size. Then we obtain the training dataset $\mathcal{D}_{\text{train}} = \{(q_i, D_i, Y_i)\}_{i=1}^{m}$.

We use f to denote the IR model; it predicts the relevance score $f(q, d)$ based on a given query q and document d. The IR model f is derived by learning from the following objective:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(q,D,Y)\sim\mathcal{D}_{\text{train}}} \mathcal{L}\left(f(q, d), Y; \theta\right), \tag{1}$$

where θ are the parameters of the IR model f, and $\mathcal{L}$ is a ranking loss function.

The ranking performance of an IR model is usually evaluated by a metric M that focuses on the top-K ranking results, e.g., Recall@K, **normalized discounted cumulative gain (NDCG)**@K and **mean reciprocal rank (MRR)**@K. Given a triple (q, D, Y) and an IR model f, the score of M on query q is calculated by:

$$M\left(f; (q, D, Y), K\right) = \sum_{(d, y_d)\in(D,Y)} y_d \cdot h\left(\pi_f\left(q, d\right)\right) \cdot \mathbb{I}\left\{\pi_f\left(q, d\right) \le K\right\}, \tag{2}$$

where $\pi_f(q, d)$ is the rank of document d under query q ranked by model f, h is the mapping function related to ranking, dependent on the specific metric, and $\mathbb{I}\{\cdot\}$ is an indicator function which is equal to 1 when its condition is satisfied and 0 otherwise. Notably, here we only consider meta-evaluation metrics. For composite evaluation metrics, such as average precision, M can be regarded as the meta metric, specifically precision, while the complete value needs to be derived.

Further, given a test dataset $\mathcal{D}_{\text{test}}$, the ranking performance $\mathcal{R}_M$ against metric M is calculated by:

$$\mathcal{R}_M\left(f; \mathcal{D}_{\text{test}}, K\right) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(q,D,Y)\in\mathcal{D}_{\text{test}}} M\left(f; (q, D, Y), K\right). \tag{3}$$

Typically, the ranking performance $\mathcal{R}_M$ of the IR model f on the test dataset $\mathcal{D}_{\text{test}}$ against metric M refers to the average evaluation score of M.

IR Models. In order to balance efficiency and effectiveness, the IR task is usually addressed as a pipeline consisting of a first-stage retrieval stage and re-ranking stage [25]:

(1) First-stage retrieval identifies a small set of candidate documents from millions of documents. Therefore, in the first-stage retrieval, D_i refers to the entire corpus. Considering efficiency, the neural IR model, represented by the DRM [61], usually adopts a dual-encoder architecture. In this architecture, the interaction function η is often null.
(2) The re-ranking stage generates the final ranked list for a query and a small set of candidate documents [63], referred to as D_i in this stage. To this end, the NRM with cross-encoder architecture is often modeled jointly by all the matching functions.

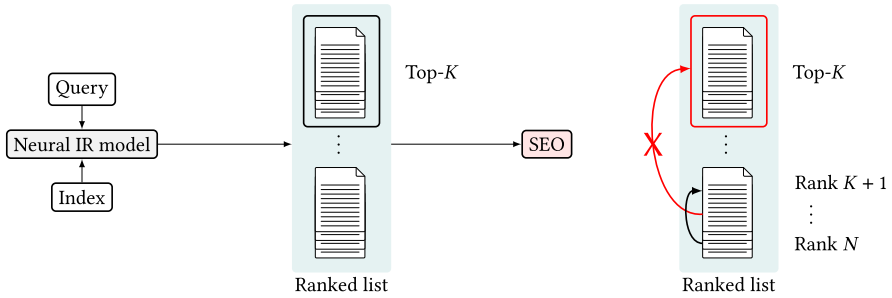

Fig. 3. The core of robust IR is to protect the stability of the Top-K results.

The Relationship between Robustness and Explainability. In IR, a number of studies have been devoted to exploring the explainability of models, including revealing a model's feature preferences for relevance judgments [69, 153, 169] and measuring the model's explainability [133, 171]. Robustness focuses on the evaluation of model performance, while explainability is concerned with explaining model performance from an internal perspective. Explaining a model's mechanism for arriving at relevance judgments may help to develop robustness enhancement methods, while studying robustness can provide a richer external evaluation perspective for explainability studies. In this article, we focus on the evaluation and enhancement methods of robustness, and supplement our understanding with the results of explainability research (if any) where appropriate.

2.1 Definition of Robustness in IR

Robustness refers to the ability to withstand disturbances or external factors that may cause a system to malfunction or provide inaccurate results [70]. It is important for practical applications, especially in safety-critical scenarios, e.g., medical retrieval [6], financial retrieval [68], patent retrieval [114], and sensitive retrieval [156]. If, for any reason, an IR system behaves abnormally, the service provider can lose time, manpower, opportunities, and even credibility. With the development of deep learning, robustness has received much attention in the fields of CV [13] and NLP [174]. Concerns about model robustness in these fields are mainly focused on the test phase. In this scenario, the model is trained on an unperturbed dataset but tested for its performance when exposed to adversarial examples or OOD data [56, 140].

In IR, the robustness of model in the test phase is also important due to the widespread availability of SEO [65] and the need for models to adapt to unseen data. Hence, in this survey, we follow prior work and only discuss the robustness of a model in the test phase.

In most deployed systems, when presented with a ranked list, users focus most of their attention on the top-K search results, as evidenced by a significant drop in traffic and click-through rates further down the list [129, 181] and by the prevalence of ranking metrics like MRR [36] and NDCG [74], which primarily evaluate the effectiveness of these top-ranking results. The relationship between top-K result stability and robust IR is shown in Figure 3. Taking SEO as an example, it aims to get a specific document displayed in the top-K-ranked results. A robust neural IR model protects its top-K results from being affected. Consequently, ensuring the integrity and robustness of the top-K-ranked results is crucial for deploying IR models in practical web search applications. Based on this, we present a formal definition of *top-K robustness in IR* by incorporating the original test dataset $\mathcal{D}_{\text{test}}$ from the initial dataset and the unseen test dataset $\mathcal{D}^*_{\text{test}}$.

Definition 2.1 (Top-K Robustness in IR). Let $\delta \geq 0$ denote an acceptable error threshold. Let $f_{\mathcal{D}_{\text{train}}}$ be an IR model trained on training dataset $\mathcal{D}_{\text{train}}$, with a corresponding test dataset $\mathcal{D}_{\text{test}}$, and an

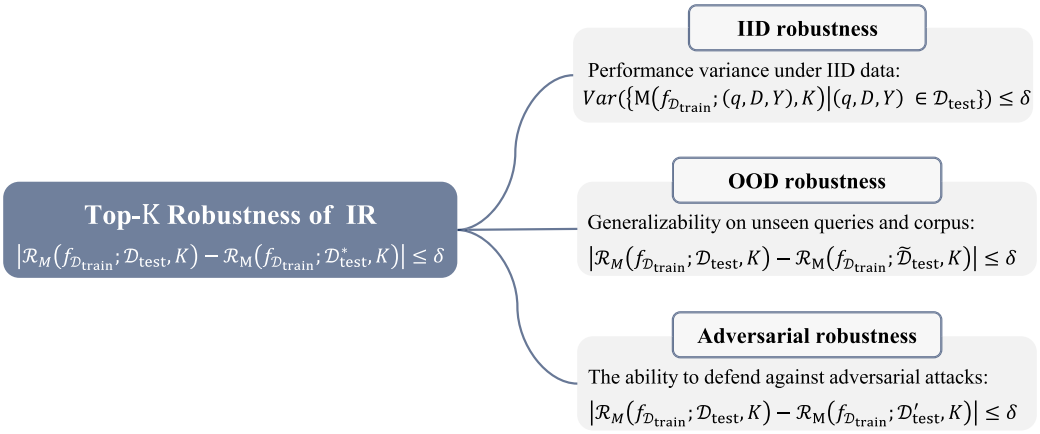

Fig. 4. A taxonomy of robustness in IR. In this survey, we pay special attention to adversarial robustness and OOD robustness.

unseen test dataset $\mathcal{D}^*_{\text{test}}$, for the top-$K$ ranking results. If:

$$\left|\mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}_{\text{test}}, K\right) - \mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}^*_{\text{test}}, K\right)\right| \leq \delta, \tag{4}$$

we consider the model $f_{\mathcal{D}_{\text{train}}}$ to be *δ-robust for metric M*.

The formal definition of top-K robustness in IR is inspired by differential privacy [45]. The difference is that in this article, we express robustness in terms of how drastically the performance of the IR model changes in different test environments. Therefore, robustness is not an absolute concept; the value of δ depends on the acceptable level of robustness of the IR model in the specific application environment. Differences between the unseen test dataset $\mathcal{D}^*_{\text{test}}$ and the original test dataset $\mathcal{D}_{\text{test}}$ in different robustness scenarios will be specifically analyzed below.

In Section 2.2, below, $\mathcal{D}^*_{\text{test}}$ refers to different test datasets depending on the context: $\mathcal{D}'_{\text{test}}$ for test datasets with adversarial examples in adversarial robustness, $\tilde{\mathcal{D}}_{\text{test}}$ for test datasets from new domains in OOD robustness.

2.2 Taxonomy of Robustness in IR

In IR, robustness threats come in different flavors, including IID robustness [183], OOD robustness [167], and adversarial robustness [183]. A high-level taxonomy of robustness in IR is shown in Figure 4.

2.2.1 IID Robustness. Typically, the performance of IR models is first represented by their overall performance on IID data. Recently, it has been recognized that performance stability across IID queries may be compromised when we try to improve the average retrieval effectiveness across all queries [196]. Therefore, a robust neural IR model should not only have good retrieval performance on the overall testing queries, but also ensure that the performance on individual queries is not too bad. Next, we give a formal definition of IID robustness in IR based on Definition 2.1.

Definition 2.2 (IID Robustness of IR). Let the following be given: an IR model $f_{\mathcal{D}_{\text{train}}}$ trained on training dataset $\mathcal{D}_{\text{train}}$ with a corresponding IID test dataset $\mathcal{D}_{\text{test}}$, and an acceptable error threshold δ, for the top-K ranking result. If:

$$\text{Var}\left(\{M\left(f_{\mathcal{D}_{\text{train}}}; (q, D, Y), K\right) \mid (q, D, Y) \in \mathcal{D}_{\text{test}}\}\right) \leq \delta, \tag{5}$$

where $\mathrm{Var}(\cdot)$ is the variance of the ranking performance of the IR model $f_{\mathcal{D}_{\text{train}}}$ on $\mathcal{D}_{\text{test}}$, then the model f is considered to be *δ-robust in terms of IID data for metric M*.

2.2.2 OOD Robustness. IR models need to cope with a constant stream of unseen data [183]. The key behind addressing this challenge is how to adapt the model to new data outside of the familiar distribution. There are a variety of OOD scenarios in IR, so the OOD robustness of the model is of broad interest. First, the query entered by the user may be unknown and of varying quality [103, 137]. Then, search engine application scenario migration and incremental new documents will likely bring in OOD data [18, 23, 167]. Manual labeling of unseen data as well as retraining IR models incurs significant resource overheads [127, 167]. Therefore, a crucial question is how to efficiently train IR models to achieve effective performance on unseen data.

OOD robustness measures the performance of an IR model on unseen queries and documents from distributions that differ from the training dataset. By introducing a test dataset $\tilde{\mathcal{D}}_{\text{test}}$ in a new domain, we give a formal definition of OOD robustness in IR based on Definition 2.1.

Definition 2.3 (OOD Robustness of IR). Let the following be given: an IR model $f_{\mathcal{D}_{\text{train}}}$, an original dataset with training and test data, $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, drawn from the original distribution $\mathcal{G}$, along with a new test dataset $\tilde{\mathcal{D}}_{\text{test}}$ drawn from the new distribution $\tilde{\mathcal{G}}$, and an acceptable error threshold δ, for the top-K ranking result. If:

$$\left|\mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}_{\text{test}}, K\right) - \mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \tilde{\mathcal{D}}_{\text{test}}, K\right)\right| \leq \delta \text{ where } \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}} \sim \mathcal{G}, \tilde{\mathcal{D}}_{\text{test}} \sim \tilde{\mathcal{G}}, \tag{6}$$

then the model f is considered to be *δ-robust against OOD data for metric M*. In OOD robustness, the new test dataset $\tilde{\mathcal{D}}_{\text{test}}$ consists of different document sources (e.g., a web crawl or patent library, new or old documents), or different query forms (variants, new query types) than the original test dataset $\mathcal{D}_{\text{test}}$.

2.2.3 Adversarial Robustness. In a competitive scenario, content providers may aim to promote their products or documents in rankings for specific queries [87]. This has provided a market for SEO and has led to the development of attack techniques against search engines. Traditional attacks against search engines are generally called *term spamming*. They usually resort to stacking keywords to achieve a boost in ranking. In recent years, with the development of deep learning, a number of neural approaches have emerged that attack through more imperceptible perturbations. As search engines evolve, defense methods against term spamming are maturing as well. Adversarial robustness focuses on the stability of an IR model's performance when imperceptible malicious perturbations are added to documents. By introducing a test dataset $\mathcal{D}'_{\text{test}}$ with adversarial examples, we give a formal definition of adversarial robustness in IR based on Definition 2.1.

Definition 2.4 (Adversarial Robustness in IR). Let the following be given: an IR model $f_{\mathcal{D}_{\text{train}}}$ trained on training dataset $\mathcal{D}_{\text{train}}$ with a corresponding testing dataset $\mathcal{D}_{\text{test}}$, a new document set D_{adv} containing adversarial examples, and an acceptable error threshold δ, for the top-K ranking result. If:

$$\left|\mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}_{\text{test}}, K\right) - \mathcal{R}_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}'_{\text{test}}, K\right)\right| \leq \delta \text{ such that } \mathcal{D}'_{\text{test}} \leftarrow \mathcal{D}_{\text{test}} \cup D_{\text{adv}}, \tag{7}$$

where $\mathcal{D}' \leftarrow \mathcal{D}_{\text{test}} \cup D_{\text{adv}}$ denotes injecting the set of all generated adversarial examples D_{adv} into the original test dataset, then model f is considered to be *δ-robust against adversarial examples for metric M*. In adversarial robustness, according to existing research [100, 104, 182, 199], the number of adversarial documents D_{adv} injected into the original test dataset $\mathcal{D}_{\text{test}}$ is generally within 10% of the original number of documents to simulate a scenario where the corpus is poisoned.

Relatively little work has been done on IID robustness. In the following, we will first briefly introduce IID robustness and its related improvement methods. Then, we pay special attention to the two other notions of robustness depicted in Figure 4, i.e., OOD robustness and adversarial robustness. Depending on the types of OOD generalizability in IR models, existing work can be categorized into OOD generalizability on unseen documents and OOD generalizability on unseen queries; we will discuss these directions in detail in Section 4. Work on adversarial robustness usually proceeds along two lines: adversarial attacks and adversarial defenses; we will discuss these lines in detail in Section 5.

3 IID Robustness

Most IR models are designed under the assumption that observations are IID random variables, focusing primarily on improving the average effectiveness of retrieval results. However, prior work [196] has pointed out that when attempting to enhance the average retrieval effectiveness across all queries, the stability of performance among individual queries may be compromised. Moreover, some of the documents in the corpus may be hard to retrieve and also affect the performance of the IR system [7]. Therefore, this section analyzes the IID robustness of IR models by emphasizing the variance in query performance and the retrievability of documents.

3.1 Variance in Query Performance

The performance variance of IR models refers to the variance in effectiveness across different individual queries. When an IR model achieves improvements in average retrieval effectiveness (e.g., Mean Average Precision, MAP [151]), the performance of certain individual queries may deteriorate. Although failures in a small number of queries may not significantly impact the average performance, users interested in these queries are unlikely to tolerate such deficiencies. Therefore, an ideal IR model should achieve high average effectiveness while maintaining low performance variance [197].

Here, we propose using the **variance of normalized average precision (VNAP)** to measure performance variance, defined as follows:

$$VNAP = E\left[\left(NAP(q_t) - E\left[NAP(q_t)\right]\right)^2\right], \tag{8}$$

where $E[\cdot]$ denotes the expectation over a set of queries assumed to be uniformly distributed and $NAP(q_t)$ represents the normalized average precision for query q_t, defined as:

$$NAP(q_t) = \frac{AP(q_t)}{E[AP(q_t)]}, \tag{9}$$

where $AP(q_t)$ represents the average precision for query q_t, defined as:

$$AP(q_t) = \frac{1}{R_{q_t}} \sum_{k=1}^{R_{q_t}} \frac{1}{o_k} \sum_{n=1}^{o_k} \mathbb{I}\{y_{tn} > 0\}, \tag{10}$$

where R_{q_t} denotes the number of relevant documents associated with query q_t, o_k represents the rank position of the kth relevant document predicted by the IR model (ranging from 1 to the size of the document list), y_{tn} indicates the true label of document d_{tn}, and $\mathbb{I}\{\cdot\}$ is an indicator function used to count the number of relevant documents.

It is worth noting that VNAP is similar to VAP [196], except that we normalize the average precision to eliminate the influence of average performance. Since different models may have varying levels of average performance, it is necessary to eliminate this influence to better measure the variance of IR models. Models with lower variance are considered more robust.

3.2 Retrievability of Documents

While variance in query performance focuses on ensuring that poorly performing queries in a retrieval system are not excessively bad, this perspective considers robustness from the query angle. From the document perspective, difficulty in retrieving relevant documents for a particular query may be the cause of the variance in performance. Since users typically only view the top-K results returned by the retrieval system, we define a document as retrievable if it appears in the top-K results for a given query. Furthermore, a document is considered *retrievable* if there exists a query that can retrieve it.

For certain tasks, it is necessary to ensure that all documents in a collection are retrievable. For example, in patent retrieval [114], patent searchers need to ensure that the retrieval system can locate all documents in the collection relevant to their information needs. To measure whether a document d is retrievable with respect to a model f, we introduce the notion of retrievability.

Definition 3.1 (Retrievability of Documents). Azzopardi and Vinay [7] define document retrievability as the likelihood of a document being retrieved by an IR model in the top-K result, expressed as:

$$R(d)@K = \sum_{q \in Q} p(q) \cdot \mathbb{I}\{\pi_f(q, d) \leq K\}, \tag{11}$$

where $p(q)$ denotes the probability of users issuing query q (often assumed to be 1), Q is the set of all possible queries, $\pi_f(q, d)$ represents the rank position of document d given by IR model f for query q, and $I\{\cdot\}$ is an indicator function that equals 1 if the condition is satisfied and 0 otherwise.

In practical settings, since the set Q is vast, calculating retrievability as defined above is infeasible. Therefore, an estimation is needed. One approach [7] is to use a subset of all possible queries Q, which is sufficiently large and contains relatively likely or feasible queries. For instance, historical queries from query logs [53] or synthetically generated queries using deep-learning models [1] can be used. Another method involves estimating retrievability based on document features [8], which allows for quick estimation of retrievability levels.

Retrievability Bias. Given the retrievability score $R(d)@K$ for each document, the inequality in retrievability across all documents can be measured. To this end, we use the Gini coefficient to evaluate the bias in retrievability:

$$G = \frac{\sum_{i=1}^{N} (2 \cdot i - N - 1) \cdot R(d_i)@K}{N \cdot \sum_{j=1}^{N} R(d_j)@K}, \tag{12}$$

where N is the number of documents. A Gini coefficient of 0 indicates equal retrievability for all documents (complete equality), while a coefficient of 1 indicates that one document has all the retrievability while others have none (complete inequality). Thus, a lower Gini coefficient implies less bias in the model.

Although a system with lower retrievability bias does not necessarily indicate higher effectiveness, studies have found correlations between the two [10, 179, 180], suggesting that retrievability bias metrics may be useful for selecting better retrieval systems.

To reduce retrievability bias in systems, Bashir and Rauber [9] propose a new document selection process combining query expansion techniques for pseudo-relevance feedback in the patent search domain. Bogers and Petras [15] suggest using document retrievability across a set of query variants to determine which documents to use for relevance feedback. For DRMs, Penha et al. [138] propose a method to enhance document retrievability and reduce retrievability bias through controllable

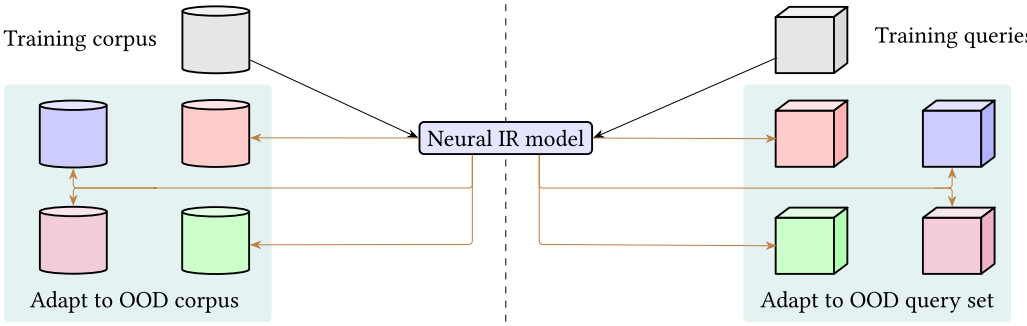

Fig. 5. OOD generalizability on unseen documents vs. queries in neural IR.

query generation, which involves both augmenting the training set and expanding queries during user searches.

Overall, the IID robustness of an IR model consists of the performance variance from the query perspective and the retrievability from the document perspective. With respect to each of these two perspectives, we have briefly described the limited existing work, leaving more exploration for the future. In the next two sections, we will introduce OOD robustness and adversarial robustness in detail.

4 OOD Robustness

In addition to IID robustness, deep neural networks lack generalizability to OOD data. When faced with data that differs from the distribution of the training data, neural networks struggle to maintain performance. In IR, this problem has begun to be exposed and attract attention as neural IR models are now widely being used [167, 183]. Prior work focuses on the OOD generalizability of DRMs, since OOD data has a direct impact on the retrieval stage. We refer back to Section 2.2 for a definition of OOD robustness in IR. In this section, we present specific work on OOD generalizability on unseen documents and OOD generalizability on unseen queries, respectively.

4.1 Overview

For a long time, research on neural IR models has been carried out in a narrow IID setting. In this setting, the model faces homogeneous data during training and testing. But IR systems are widely used in fields such as search engines [72], digital libraries [33], medical search [111], legal search [121], and at the same time, the scenarios that IR systems need to cope with are becoming more complex.

OOD Generalizability Requirements in IR. With deep neural network models being applied to IR, neural IR models have demonstrated excellent results on many tasks [61, 63]. But IR systems need to face more than just a single task or scenario [167]. In order to deploy an IR system, it is often necessary to construct a training dataset for the neural IR model in it. This process is usually time-consuming and expensive and hence many IR systems are expected to be able to cope with a wide variety of data not seen during training [103, 137, 167]. Therefore, OOD generalizability is a key requirement for contemporary IR systems, given the dynamic nature of user needs and evolving data landscapes. In IR, OOD generalizability is focused on unseen documents and unseen queries as illustrated in Figure 5.

Why Should IR Models Be Able to Generalize to Unseen Documents? In real-world scenarios, the data landscape is constantly evolving, with new documents and information being generated regularly. It is expensive to annotate each new corpus and retrain the IR models. Therefore, a neural IR model that can generalize well to unseen corpora ensures its relevance and usefulness over time, without requiring constant retraining or fine-tuning. Moreover, in complex real-life scenarios, generalizability to a new corpus helps IR models against distributional shifts or domain-specific biases. This helps to ensure that IR models deliver reliable ranking results irrespective of diverse contexts, qualities, and domains.

Why Should IR Models Be Able to Generalize to Unseen Queries? The set of possible queries that users may input is vast and constantly evolving [60, 137, 206]. E.g., 15% of daily Google searches are brand new.[5] The nature of information needs is dynamic and diverse [62, 92]. Users often express their information needs in varied ways, using different vocabulary, language styles, or even typos [31, 205]. This challenge becomes particularly pronounced in the context of ever-changing user interests and the introduction of new vocabularies. Therefore, IR models must possess the ability to handle queries that were not encountered during training. Without generalizability to unseen queries, IR models risk providing inadequate or irrelevant results, ultimately diminishing user satisfaction and trust in the system [20, 137]. A robust neural IR model should be able to understand and accommodate these variations, effectively retrieving relevant information regardless of how the query is formulated.

Furthermore, in reality, there is a wide variety of query types that are often not fully or adequately accessible during IR model training [206, 207]. But, a robust IR system should perform consistently in response to a wide range of query types.

4.2 OOD Generalizability to Unseen Documents

As argued above, IR systems need to adapt to different environments and variations in the corpus. However, retraining the neural IR models in each new environment is costly. Previous work has only analyzed the generalizability of IR models across different domains [147, 167, 183]. In this work, we summarize work on adaptation to a new corpus and updates to a corpus. Figure 6 illustrates how we organize the discussion of different methodologies.

4.2.1 Definition. Generalizability to unseen documents implies the ability of an IR model to maintain retrieval performance when encountering a new and unfamiliar corpus. In IR, improving the model's OOD generalizability under unseen documents is mainly reflected in enhancing the retrieval performance of the IR model under various new corpus. Without loss of generality, given a test set $\tilde{\mathcal{D}}_{\text{test}}$ with new corpus $\tilde{C}$, they draw the new distribution $\tilde{\mathcal{G}}$, the goal of improving the OOD generalizability of a neural IR model f on unseen documents under top-K ranked results can usually be formalized as:

$$\max \mathcal{R}_M \left(f_{\mathcal{D}_{\text{train}}}; \tilde{\mathcal{D}}_{\text{test}}, K \right) \text{ such that } C_{\sim \mathcal{G}} \in \mathcal{D}_{\text{train}},\ \tilde{C}_{\sim \tilde{\mathcal{G}}} \in \tilde{\mathcal{D}}_{\text{test}}. \tag{13}$$

Specifically, the new corpus $\tilde{C}$ may result from two main scenarios with respect to an unseen new corpus and updates to a corpus:

Adaptation to New Corpora. Adaptation to a new corpus refers to the trained IR models that may be hard to adapt to a corpus of new domains in the absence of supervised data [167]. The goal of improving generalizability on the adaptation to a new corpus for a neural IR model f under top-K

[5] https://blog.google/products/search/our-latest-quality-improvements-search/.

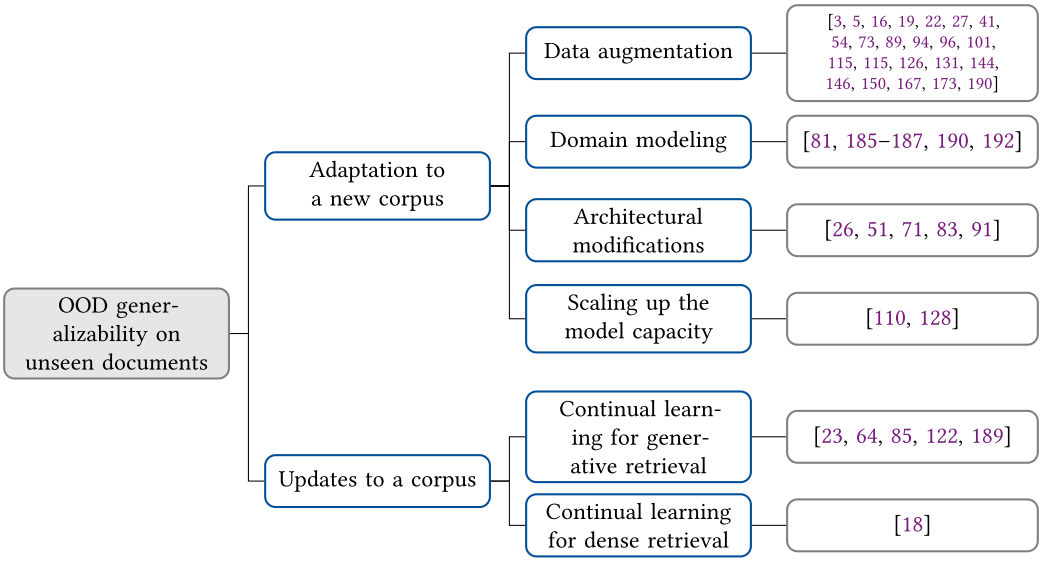

Fig. 6. Classification of OOD generalizability on unseen documents.

ranked results can usually be formalized as:

$$\max \mathcal{R}_M \left(f_{\mathcal{D}^o_{\text{train}}}; \mathcal{D}^n_{\text{test}}, K \right) \text{ such that } C^o_{\sim \mathcal{G}} \in \mathcal{D}^o_{\text{train}}, \; C^n_{\sim \tilde{\mathcal{G}}} \in \mathcal{D}^n_{\text{test}}, \tag{14}$$

where o is the domain of the original corpus and n is the new corpus domain. Among the main solutions for adaptation to a new corpus are data augmentation [16, 73, 173], domain modeling [190], architectural modifications [83], and scaling up the model capacity [128].

Updates to a Corpus. Updates to a corpus refer to the problem for the trained IR model to maintain its ranking performance under continuously arriving new documents [23]. The goal of improving generalizability on the updates to a corpus for a neural IR model f under top-K ranked results can be formalized as:

$$\max \mathcal{R}_M \left(f_{\mathcal{D}^{\Sigma_0^t}_{\text{train}}}; \mathcal{D}^{t+1}_{\text{test}} \right) \text{ such that } C^{\Sigma_0^t}_{\sim \mathcal{G}} \in \mathcal{D}^{\Sigma_0^t}_{\text{train}}, \; C^{t+1}_{\sim \tilde{\mathcal{G}}} \in \mathcal{D}^{t+1}_{\text{test}}, \tag{15}$$

where t is the time session of each corpus update. The main solution for maintaining the ranking performance is continual learning, but the different paradigms of **generative retrieval (GR)** [124] and **dense retrieval (DR)** [61] lead to different solutions in the two settings.

Below, we first introduce the evaluation metrics widely used for OOD generalizability on unseen documents. Then, we detail solutions for adaptation to a new corpus and updates to a corpus, respectively.

4.2.2 Evaluation. OOD generalizability of IR models on unseen documents is mainly measured by the ranking performance under the new corpus. For both adaptation to a new corpus and updates to a corpus, ranking performance is the common evaluation. For updates to a corpus, previous work also evaluates the degree to which the old corpus is forgotten.

Metrics for Ranking Performance. For adaptation to a new corpus and updates to a corpus, the ranking performance of IR models under unseen documents is evaluated by common metrics:

— *NDCG* [74] evaluates the quality of ranked results by measuring the gain of a document based on its position in the ranked list;
— *MRR* [36] evaluates the performance of a ranking result by calculating the average of the reciprocal ranks of the first relevant document answer;
— *HIT* [23] evaluates the proportion of times a relevant document is found within a set of top-N ranked results; and
— *AP* [123] evaluates the average performance of the ranking performance metrics, overall new domains in adaptation to new corpus, and sessions in updates to a corpus; the ranking performance metric could be any of the above.

Metrics for the Degree of Forgetting the Old Corpus. Updates to a corpus are an ongoing process with many sessions; they require that the model memorizes new data without forgetting the old. Therefore, some metrics have been proposed to evaluate the model performance from a time-series perspective.

— *Training time* [18] evaluates the total time it takes for the IR model to learn new data while recalling old data;
— $Forget_t$ [18] evaluates how much the model forgets at session t:

$$\text{Forget}_t = \frac{1}{t}\sum_{j=0}^{t-1} \max_{l\in\{0,\ldots,t-1\}} \left(p_{l,j} - p_{t,j}\right), \tag{16}$$

where p is the ranking performance under any common metrics; and
— *FWT* [18] evaluates how well the model transfers knowledge from one session to future sessions:

$$\text{FWT} = \frac{\sum_{i=1}^{j-1}\sum_{j=2}^{T} p_{i,j}}{\frac{T(T-1)}{2}}, \tag{17}$$

where T is the total number of sessions.

4.2.3 Adaptation to New Corpora. Proposed solutions to the adaptation to a new corpus involve data augmentation, distributionally robust optimization, and domain-invariant projection.

Data Augmentation. Data augmentation involves generating or modifying data in such a way that it bridges the gap between the source domain (the domain where the model was originally trained) and the target domain (the new domain where the model is to be applied). This can include techniques like synthesizing new data examples through transformations that maintain the integrity of the underlying patterns, translating examples from one domain to another, or creating semi-synthetic samples. GPL [173] uses an unsupervised domain adaptation method generative pseudo labeling, which combines a query generator with pseudo labeling from a cross-encoder. HyperR [19] performs a hyper-prompted training mechanism to enable uniform retrieval across tasks of different domains.

There is other work that conducts unsupervised pre-training by using large-scale positive and negative pairs with different data augmentation methods such as, query generation [22, 96, 115, 115, 150, 167], synthetic pre-training [73, 126, 146, 190], or synthetic relevance labels [41, 54, 94, 101, 144]. Overall, data augmentation can enrich the training set to include more domain-relevant variations, thereby enhancing the model's ability to generalize across domains.

Very recently, LLMs for data augmentation have significantly enhanced IR models by enabling effective corpus adaptation [3, 16, 89]. Anaya-Isaza and Mera-Jiménez [5] explore various data augmentation strategies combined with transfer learning to improve MRI-based brain tumor

detection accuracy. Chen et al. [27] develop a cross-domain augmentation network to enhance click-through rate prediction by transferring knowledge between domains with different input features. Oh et al. [131] propose a prompt-based data augmentation method using generative language models for creating synthetic parallel corpora, improving neural machine translation performance.

Domain Modeling. Domain modeling seeks to model the data from both the source and target domains into a common feature space where the differences between the domains are minimized. The idea is to learn a representation of the data that retains the essential information for the task at hand while discarding domain-specific features that might lead to bias or overfitting. By doing so, the model learns to focus on the underlying task without being distracted by differences between the domains. COCO-DR [190] use implicit distributionally robust optimization to reweight samples from different source query clusters for improving model robustness over rare queries during fine-tuning. Together with contrastive learning, this approach significantly improves the generalization of DRM over different corpora. There have been many successive efforts to optimize for this problem, including MoDIR [186], and ToTER [81]. Xu et al. [187] address the domain OOD challenge by modeling a single passage as multiple units with two objectives. One is the semantic balance between units and the other is the extractability of essential units. Distributionally robust optimization helps in reducing the sensitivity of the model to changes in data distribution, thereby improving its adaptability.

Another way to deal with new domain data is domain-invariant projection. Zhan et al. [192] have been the first to use a relevance estimation module for modeling domain-invariant matching patterns and several domain adaption modules for modeling domain-specific features of multiple target corpora. Xian et al. [185] propose a list-level alignment method, which aligns the distributions of the lists and preserves their list structure. They also demonstrate the superiority of their method on theoretical grounds. The domain-invariant feature space enables the model to perform well on the target domain using knowledge acquired from the source domain, thereby facilitating effective domain adaptation.

Architectural Modifications. By optimizing the architecture of an IR model, the model can be made to have good domain adaptability. For instance, hybrid retrieval models have been employed to integrate out-of-domain semantics, enhancing zero-shot capabilities and using core strengths of foundational model features [26, 91]. Additionally, DESIRE-ME [83] uses a mixture-of-experts to tailor retrieval strategies effectively across various domains.

Moreover, methods like employing search agents in hybrid environments or using knowledge distillation with hard negative sampling further support the development of IR systems that maintain high performance in unseen domains [51, 71]. These strategies collectively enhance the adaptability and effectiveness of retrieval systems across a range of out-of-domain scenarios.

Scaling Up the Model Capacity. Scaling up the model capacity has been identified as a crucial approach that significantly boosts a model's ability to handle diverse data types and improve retrieval effectiveness across unfamiliar domains.

Ni et al. [128] explore the impact of enlarging dual-encoder architectures. They demonstrate that larger models are not only more capable of handling complex queries but also exhibit enhanced generalization across different domains. In a similar vein, Lu et al. [110] introduce "Ernie-search" which uses a novel method of self on-the-fly distillation to bridge the gap between cross-encoder and dual-encoder architectures. This technique enhances the dual-encoder's performance by distilling knowledge from a more powerful cross-encoder, effectively scaling up the retrieval capacity without the direct computational cost typically associated with larger models.

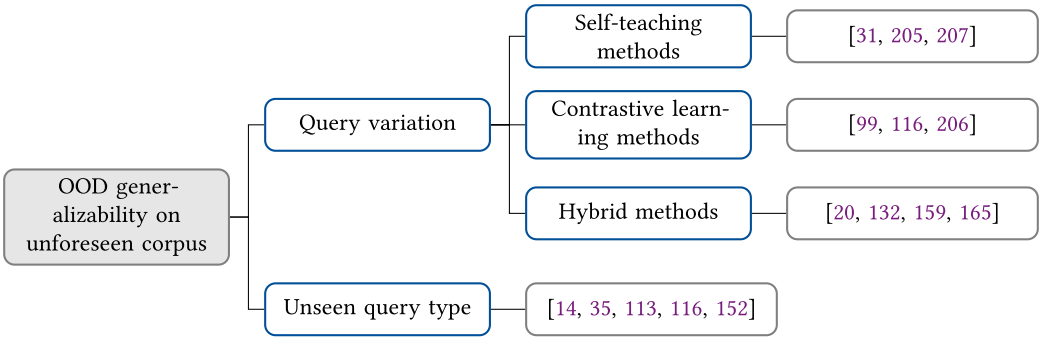

Fig. 7. Classification of OOD generalizability on unseen queries.

4.2.4 Updates to a Corpus. In this scenario, IR models need to be compatible with newly added documents, however, this can lead to catastrophic forgetting problems with old documents. Therefore, continuous learning [42] has become a dominant approach, which aims to adapt the model to the newly added data without losing the ability to understand the old data by quickly adapting to the new unlabeled (little labeled) data.

Continual Learning for GR. In GR, a sequence-to-sequence model is adopted to unify both the indexing and retrieval stages. All document information is encoded into the model parameters. The tight binding of the index to the retrieval module makes updating the index costly. To tackle this challenge, DSI++ [122] adapts a continual learning method for DSI [166] to incrementally index new documents while maintaining the ability to answer user queries related to both previously and newly indexed documents. After that, CorpusBrain++ [64] uses a continual learning method on another GR model called CorpusBrain [24]. Chen et al. [23] propose CLEVER to incrementally index new documents while supporting the ability to query both newly encountered documents and previously learned documents. CLEVER performs incremental product quantization [75] to update a partial quantization codebook, and use a memory-augmented learning mechanism to form meaningful connections between old and new documents. Subsequent work on continuous learning has been devoted to the problem of guaranteeing to updates to a corpus for different GR models, for example, DynamicIR [189] and IncDSI [85].

Continual Learning for DR. In DR, the model needs to learn the representation space of the entire corpus and encode each document into an embedding to serve as the index. Therefore, continual learning for DR should effectively adapt to the evolving distribution with the unlabeled new-coming documents, and avoid re-inferring all embeddings of old documents to efficiently update the index each time the model is updated. L^2R [18] uses backward-compatible representations to deal with this problem. It first selects diverse support negatives for model training, and then uses a ranking alignment objective to ensure the backward-compatibility of representations.

4.3 OOD Generalizability to Unseen Queries

Deep-learning-based models, constrained by their training data, often falter when faced with novel query formulations. Previous work has analyzed the generalizability of IR models under query variants and different query types [103, 137, 183, 204], respectively. Next, we summarize prior work and how to improve the generalizability of IR models under unseen queries. The specific methodology categorization is shown in Figure 7.

4.3.1 Definition. Generalizability to unseen queries indicates the capacity of an IR model to sustain its retrieval performance when confronted with new and unfamiliar query formulations. Enhancing a model's OOD generalizability with respect to unseen queries mainly involves improving the retrieval accuracy of the IR model across a variety of novel queries. Without loss of generality, given a test set $\tilde{\mathcal{D}}_{\text{test}}$ comprising new queries $\tilde{\boldsymbol{Q}}$, which introduce a new distribution $\tilde{\mathcal{G}}$, the objective of augmenting the OOD generalizability of a neural IR model f for unseen queries under top-K ranked results can be formalized as:

$$\max \mathcal{R}_M \left(f_{\mathcal{D}_{\text{train}}}; \tilde{\mathcal{D}}_{\text{test}}, K \right) \text{ such that } \boldsymbol{Q}_{\sim \mathcal{G}} \in \mathcal{D}_{\text{train}}, \tilde{\boldsymbol{Q}}_{\sim \tilde{\mathcal{G}}} \in \tilde{\mathcal{D}}_{\text{test}}. \tag{18}$$

Specifically, the new queries $\tilde{\boldsymbol{Q}}$ may result from two main scenarios with respect to query variation and unseen query type:

Query Variation. Query variations refer to different expressions of the same information need. The way in which information is expressed may impact the effectiveness of IR models [137]. Some query variants, which introduce additional information, e.g., through query expansion, tend to enhance the retrieval performance [11, 12]. Some introduce noise, such as typos, grammatical errors, and variations in word order, which often challenges the robustness of the IR model [20, 137, 206]. In this article, we mainly focus on the latter one.

The goal of improving generalizability on the query variation for a neural IR model f under top-K ranked results can usually be formalized as:

$$\max \mathcal{R}_M \left(f_{\mathcal{D}_{\text{train}}}; G\left(\mathcal{D}_{\text{test}}\right), K \right) \text{ such that } \boldsymbol{Q}_{\sim \mathcal{G}} \in \mathcal{D}_{\text{train}}, G\left(\boldsymbol{Q}\right)_{\sim \tilde{\mathcal{G}}} \in G\left(\mathcal{D}_{\text{test}}\right), \tag{19}$$

where $G(\cdot)$ is a query variation generator that can generate the query variant based on each query in $\boldsymbol{Q}$. Among the main solutions for maintaining consistent performance when conformed with query variation are (1) self-teaching method, (2) contrastive learning method, and (3) hybrid method.

Unseen Query Type. Unseen query type refers to unfamiliar query types with new query intents that have not been seen during model training. The main solution for unseen query types is cross-domain regularization. The goal of improving generalizability on an unseen query type for a neural IR model f under top-K ranked results can be formalized as:

$$\max \mathcal{R}_M \left(f_{\mathcal{D}^{\tau_i}_{\text{train}}}; \mathcal{D}^{\tau_j}_{\text{test}}, K \right) \text{ such that } \boldsymbol{Q}^{\tau_i}_{\sim \mathcal{G}} \in \mathcal{D}^{\tau_i}_{\text{train}}, \boldsymbol{Q}^{\tau_j}_{\sim \mathcal{G}} \in \mathcal{D}^{\tau_j}_{\text{test}}, \tag{20}$$

where τ_i and τ_j are the different types of queries.

Next, we first introduce evaluation metrics that are widely used for OOD generalizability on unseen queries. Then, we detail existing solutions for query variation and unseen query type, respectively.

4.3.2 Evaluation. OOD generalizability of IR models on unseen queries is mainly measured by the ranking performance under the new query set. In addition to the metrics we mentioned in Section 4.2.2 for measuring ranking performance, there are two other metrics for ranking. There are also specific metrics for unseen query types to evaluate differences in the performance of IR models across query types:

Metrics for Ranking Performance. In addition to MRR and NDCG, the ranking performance of the IR model under unseen queries is evaluated by other common metrics for query variation and unseen query type:

— *Recall* [31] measures the proportion of relevant documents that are successfully retrieved from the total amount of relevant documents available.

— *MAP* [90] quantifies the average precision of retrieval across different recall levels, effectively summarizing the precision at each point where a relevant document is retrieved.

Specific Metrics for Unseen Query Type. DR_{OOD} evaluates the drop rate between the ranking performance on the original type of queries and the ranking performance on the unseen type of queries [183]:

$$DR_{OOD} = \frac{p_{OOD} - p_{IID}}{p_{IID}}, \tag{21}$$

where p_{IID} is the ranking performance on original type of queries and p_{OOD} is the ranking performance on unseen type of queries.

4.3.3 Query Variation. Solutions to the query variation challenge include (1) a self-teaching method, (2) a contrastive learning method, and (3) a hybrid method.

Self-Teaching Methods. The self-teaching approach to query variations focuses on distilling the matching capabilities of the IR model on the original clean query to the case of query variants. These methods often align the model output for distillation. Chen et al. [31] argue that the drift between query variations and original queries in model representation space affects the subsequent effectiveness of IR models. Based on this, they propose RoDR, which calibrates the in-batch local ranking of query variants to that of the original query for the representation space alignment. Zhuang and Zuccon [207] also notice this issue and employ CharacterBERT [47] as the backbone encoder to perform a character-level self-teaching method. This method distills knowledge from queries without typos into the queries with typos in a character embedding space. ToRoDer [205] uses a pre-training method that uses bottlenecked information to recover the query variation.

Contrastive Learning Methods. Contrastive learning-based approaches to query variation typically make the model robust to query variants by enhancing the supervision of query variants against the original query as well as positive and negative samples. Zhuang and Zuccon [206] propose a simple typos-aware training method for BERT-based DRMs and NRMs. During training, this method randomly selects query variants and migrated the supervised signals of the original queries directly for training. By comparing the similarity between a query and its variations and other distinct queries with contrastive learning, Sidiropoulos and Kanoulas [158] improve the robustness of IR models when encountering typos. MIRS [99] uses a robust contrastive method by injecting [MASK] tokens into query variations and encouraging the representation similarity between the original query and the variation.

Hybrid Methods. Some work enables IR models to perform better than previously when dealing with query variants through a combination of self-teaching and contrastive learning. Tasawong et al. [165] propose a typo-robust representation learning method that combines contrastive learning with dual self-teaching achieving competitive performance. CAPOT [20] introduces a notion of an anchoring loss between the unaltered model and the aligned model and designs a contrastive alignment post-training method to learn a robust model. Sidiropoulos and Kanoulas [159] argue that previous work does not make full use of positive samples and employ contrastive learning with self-teaching that supports multiple positives. There is also work that uses LLMs to enhance the ability to deal with typos [132].

4.3.4 Unseen Query Type. Part of the unseen query type problem may be caused by unseen documents. Solutions to such problems are presented in listed in Section 4.2. In this subsection, we present additional solutions for unseen query types.

Table 1. Benchmark Datasets for Unseen Documents

Type	Dataset	#Corpora	Publications
Adaptation to a new corpus	BEIR [167]	18	[3, 5, 16, 19, 22, 26, 27, 41, 51, 54, 71, 73, 81, 83, 89, 91, 94, 96, 101, 110, 115, 126, 128, 131, 144, 146, 150, 167, 173, 185–187, 190, 190, 192]

Type	Dataset	#Documents	$\#Q_{train}$	$\#Q_{dev}$	$\#Q_{eval}$	Publications
Updates to a corpus	CDI-MS [23]	3.2M	370K	5,193	5,793	[23]
	CDI-NQ [23]	8.8M	500K	6,980	6,837	[23]
	LL-LoTTE [18]	5.5M	16K	8.5K	8.6K	[18]
	LL-MultiCPR [18]	3.0M	136K	15K	15K	[18]

#Documents denotes the number of documents in corpus; $\#Q_{train}$ denotes the number of queries available for training; $\#Q_{dev}$ denotes the number of queries available for development; $\#Q_{eval}$ denotes the number of queries available for evaluation; #Corpora denotes the number of corpora.

Wu et al. [183] analyze the robustness of ranking models in the face of unseen query types with five types of queries. They find that most NRMs do not generalize well to unseen query types. Even after training on multiple types of queries, NRMs still perform poorly when faced with a new kind of query. Among all ranking models, traditional probabilistic ranking models, such as BM25 [149], have the strongest generalizability to OOD query types, while NRMs are the worst. Liu et al. [103] analyze the robustness of GR models and DRMs under different query types and find that both models are sensitive to query types.

To improve the generalizability to unseen query types for IR models, Cohen et al. [35] explore the use of adversarial learning as a regularization technique across different domains within the ranking task framework. By employing an adversarial discriminator and training a NRM across a limited number of domains, the discriminator acts to give negative feedback, thereby preventing the model from adopting domain-specific representations. Bigdeli et al. [14] propose to integrate two kinds of triplet loss functions into neural rankers such that they ensure that each query is moved along the embedding space, through the transformation of its embedding representation, in order to be placed close to its relevant document. In this way, they provide the opportunity to jointly rank documents and difficult queries. There has been some work focusing on the challenges of unseen query types in different scenarios [113, 116, 152].

4.4 Benchmark Datasets

In this section, we present commonly used datasets for studying OOD robustness in IR. All datasets can be found in the BestIR benchmark; details about BestIR can be found in Appendix B.

4.4.1 Datasets for Unseen Documents. The datasets for unseen documents mainly involve adaptation to new corpus datasets and updates to corpus datasets. Datasets on unseen documents in IR are listed in Table 1.

Adaptation to a New Corpus. In order to model the migration of models between corpora of different domains, datasets that cater for adaptation to a new corpus typically aggregate multiple domain-specific IR datasets. Of these, BEIR [167] is the best-known example; it includes nine retrieval tasks, such as fact-checking, news retrieval, question answering, entity retrieval. It also has 18 datasets, across diverse tasks, diverse domains, task difficulties, and diverse annotation strategies.

Table 2. Benchmark Datasets for Unseen Queries

Type	Dataset	#Q_{eval}	Publications
Query variation	DL-Typo [207]	60	[207]
	noisy-MS MARCO [20]	5.6k	[20]
	rewrite-MS MARCO [20]	5.6k	[20]
	noisy-NQ [20]	2k	[20]
	noisy-TQA [20]	3k	[20]
	noisy-ORCAS [20]	20k	[20]
	variations-ANTIQUE [137]	2k	[137]
	variations-TREC19 [137]	430	[137]
	Zhuang and Zuccon [206]	41k	[206]
Unseen query type	MS MARCO [127]	15k	[183]
	L4 [163]	10k	[35]

#Q_{eval} denotes the number of queries available for evaluation.

Updates to a Corpus. Updates to a corpus focus on the performance of an IR model when updates to a corpus occur. In order to follow updates to a corpus over time, the available datasets are mainly constructed by slicing or expanding the existing dataset. For example, to mimic the arrival of new documents in MS MARCO [127], CDI-MS [23] first randomly samples 60% documents from the whole corpus as the base documents. Then, it randomly samples 10% documents from the remaining corpus as the new document set, which is repeated four times.

4.4.2 Datasets for Unseen Queries. The datasets for unseen queries mainly involve query variation datasets and unseen query type datasets. Existing datasets on unseen queries in IR are shown in Table 2.

Query Variation Datasets. The importance of query variation datasets lies in their ability to simulate real-world search scenarios, where users often have unique ways of expressing their information needs. Query variation datasets contain sets of queries that target the same information need but are expressed in alternative ways, reflecting the natural diversity in how different users might phrase their search queries. Such datasets can include paraphrased queries, queries with typos, order-swapped queries, and queries without stop words. For example, Penha et al. [137] construct query variation datasets by turning queries from TREC DL19 [39] and ANTIQUE [66] into different variants using four categories in 10 ways.

Unseen Query Type Datasets. Unseen query type datasets have queries that are not represented in the training data, either by virtue of their topic or the nature of the information being sought. For example, the MS MARCO dataset [127] only contains five types of queries, i.e., location, numeric, person, description, and entity. The primary purpose of these datasets is to test the generalization ability of IR models to novel, real-world query scenarios that users may present.

5 Adversarial Robustness

Deep neural networks have been found to be vulnerable to adversarial examples that can produce misdirection with human-imperceptible perturbations [46, 57]. In IR, deep learning-based models are also likely to inherit these adversarial vulnerabilities [164], which raises concerns about the robustness of neural IR systems. Building on the definitions of adversarial robustness in IR provided

in Section 2.2, we survey work on adversarial robustness, focusing specifically on adversarial attacks and adversarial defenses.

5.1 Overview

In IR, adversarial robustness has gained significant attention in competitive ranking scenarios, such as web search, product search, and pharmaceutical search. In these scenarios, content publishers often want their content to achieve a higher position on the search results page to gain more exposure, which ultimately translates into more benefits [87].

Gaming and SEO in Competitive Ranking. When content publishers aim to improve their rankings, the competitive ranking scenario becomes a gaming environment. In such scenarios, each content publisher observes the current ranking of their documents and takes actions to improve their rankings [87]. SEO is a typical representative of this type of activity; it has been around since the dawn of the world wide web [65]. It includes white-hat SEO [59], which modifies documents in good faith and within the rules and expectations of search engines to optimize the quality of web pages. In contrast, black-hat SEO [21], maliciously exploiting loopholes of search engines, is used to get a site ranking higher in search results.

Apart from the ranking competition among content publishers, there is also a game-theoretic relationship between IR system owners and content publishers: IR system owners guide the optimization of the entire corpus toward higher-quality content by establishing content manipulation rules and ranking mechanisms, while content publishers seek to maximize their rankings based on the current rules [87]. This gaming relationship provides the context and platform for studying adversarial attacks and defenses in IR systems.

Traditional Web Spamming. Black-hat SEO creates a poor experience for the audience and is a common concern among site owners. One adversarial search environment brought about by black-hat SEO is primarily manifested in web spamming [21]. Web spamming refers to the manipulation of Web page content to make spam pages appear relevant for certain queries. Its specific methods, including term spamming and link spamming, have been described in detail by Gyöngyi and Garcia-Molina [65]. Since web spamming methods are often straightforward and simple, they are usually easy to restrict. A representative approach is online spam detection [200, 201], which can effectively identify term spamming based on TF-IDF features in the corpus. In the age of neural networks, this traditional attack and defense is translated to neural methods.

Why Study Adversarial Attacks in IR? Building on developments in deep learning, neural networks are now widely being used in IR models and have achieved excellent performance. Although traditional web spamming can also have a significant attack effect on neural IR models, this method does not really pose a threat due to its ease of detection. However, in the context of black-hat SEO, neural IR models are at risk of being attacked due to the inherent vulnerability inherited from neural networks. Therefore, adversarial attacks are being studied to expose vulnerability flaws in neural IR models in advance.

Why Study Adversarial Defense in IR? To mitigate adversarial attacks, there is a growing body of work that is focusing on adversarial defenses. Adversarial defense focuses on the early hardening of model vulnerabilities discovered by adversarial attacks. Its goal is to obtain robust neural retrieval models to build reliable IR systems.

The relationship between adversarial attacks and defenses is shown in Figure 8. Adversarial robustness has begun to gain widespread attention. Below, we describe these efforts from several perspectives: adversarial attacks, adversarial defenses, and benchmark datasets.

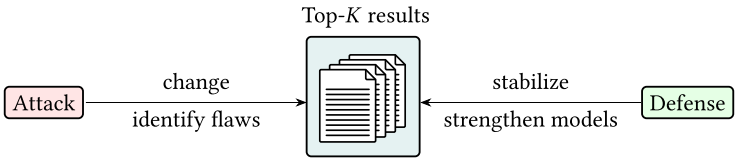

Fig. 8. Purpose and relationship between adversarial attacks and defenses.

5.2 Adversarial Attacks

As neural networks have become increasingly prevalent in IR systems, they have also become a target for adversarial attacks. Studying adversarial attacks can help understand the vulnerability of neural IR models before deploying them in real-world applications, and it can also be used as a surrogate evaluation and support the development of appropriate countermeasures.

5.2.1 What Are the Differences between IR Attacks and CV/NLP Attacks? Adversarial attacks have undergone significant development in the fields of CV and NLP [46, 57], where attacks are typically directed at image retrieval and text classification tasks, respectively. However, the landscape of adversarial attacks differs in IR: (1) compared with image retrieval attacks, IR attacks need to maintain semantic consistency of the perturbed document with the original document by considering the textual semantic similarity, rather than pixel-level perturbations within a fixed range in continuous space; and (2) inspired by black-hat SEO, the goal of IR attacks is to inject imperceptible perturbations within a document to improve its ranking for one or multiple specific queries within the entire candidate set or corpus, not inducing classification errors by the model.

Without loss of generality, given a query q and target document d, the goal of generating imperceptible perturbations p to attack against a neural IR model f under top-K ranked results can usually be formalized as:

$$\max_{p} \ \left(K - \pi_f\left(q, d \oplus p\right) + \lambda \cdot \mathrm{Sim}\left(d, d \oplus p\right)\right), \tag{22}$$

where $\pi_f\left(q, d \oplus p\right)$ denotes the ranking position of the perturbed document $d \oplus p$ in the ranked list generated by f with respect to query q. The $\mathrm{Sim}\left(\cdot\right)$ function measures the similarity between the adversarial example and the original document, both textually as well as semantically. λ is a regularization parameter used to balance two goals: keeping the adversarial samples as close as possible to the original document, while allowing adversarial samples to be ranked as high as possible. Ideally, the adversarial sample $d \oplus p$ preserves the original semantics of d and is imperceptible to human judges yet misleading to the neural IR models.

5.2.2 What Is the Attack Setting? Depending on whether the attacker has access to the knowledge of the parameters of the target model, the attack setup can be categorized into two main types [100, 182]:

(1) *White-box attacks*: Here, the attacker can fully access the model parameters and use the model gradient to directly generate perturbations.
(2) *Black-box attacks*: Here, the model parameters cannot be obtained; the attacker usually adopts a transfer-based black-box attack paradigm [32]. They construct a surrogate white-box model by continuously querying the model and getting the output. Specifically, a surrogate model is trained to simulate the performance of the target model, and then the surrogate model is attacked to the transferability of the adversarial samples to indirectly attack the target model. In IR, attackers can query the target model to obtain a ranked list with rich information.

Algorithm 1: Adversarial Sample Generation for Neural IR Models

Require:
A target query q, a target document d, a query collection Q, a corpus C, a neural IR model f, and a ranking loss function $\tilde{\mathcal{L}}$

Ensure: An adversarial document d^{adv}

if f is a black-box model **then**
  Procedure Surrogate model imitation
  Query the model f with Q, and collect ranked lists.
  Train the surrogate model $\tilde{f}$ with Q and the collected ranked lists.
  $f = \tilde{f}$
end if
Procedure Adversarial attack
Initialize the adversarial example d^{adv} as a copy of the target document d.
for $t \leftarrow 1$ to η **do**
  Query the model f with the target query q, and collect the ranked list D.
  Calculate the gradient of $\tilde{\mathcal{L}}$ with respect to the target query q and target document d:
  gradient $\leftarrow \nabla_d \tilde{\mathcal{L}}\left(f, q, d^{adv}, D\right)$
  Generate the higher dimensional adversarial perturbation ρ:
  $\rho \leftarrow$ normalize(gradient)
  Mapping high-dimensional perturbations to text space:
  $p \leftarrow \rho$
  Add textual perturbations:
  $d^{adv} \leftarrow d \oplus p$
end for
return d^{adv}

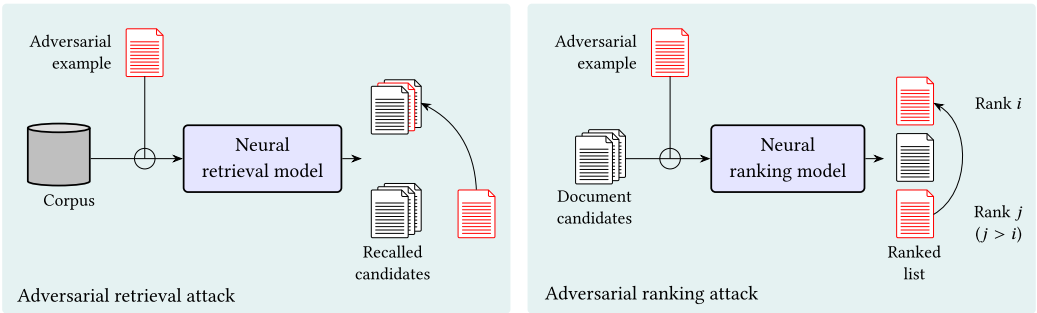

Fig. 9. Adversarial retrieval attacks vs. adversarial ranking attacks. AREA, Adversarial retrieval attacks.

Therefore, the surrogate model often has access to sufficient training data, making this attack effective [100, 182].

There are various attack methods and target models in IR. We present pseudo-code to illustrate a fundamental IR attack in Algorithm 1.

According to the type of target model, attack efforts against IR models can be categorized into two types, which are visualized in Figure 9:

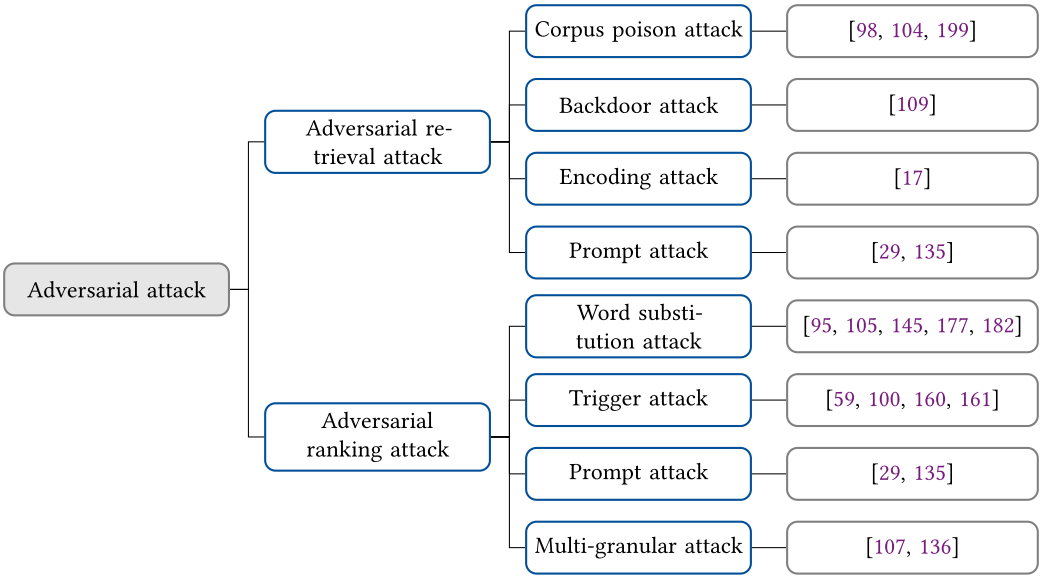

Fig. 10. Classification of adversarial attacks against neural IR models.

(1) **Adversarial retrieval attacks (AREA)** target the first-stage retrieval, mainly against DRMs; and
(2) *Adversarial ranking attacks* target the re-ranking stage, mainly against NRMs.

The methodology for conducting adversarial attacks in IR may differ based on the chosen attack strategy and the target model. A categorization of adversarial attacks in IR is shown in Figure 10.

It is worth noting that in adversarial attacks, we primarily focus on scenarios targeting the inference phase. In CV and NLP, there are cases where attackers manipulate the model by influencing the training phase [102, 188]. In IR, this assumption seems to differ from typical competitive ranking scenarios, such as SEO. Backdoor attacks are one such example [109], where attackers can inject documents with specific features into the training dataset (which is also part of the corpus). In IR attacks, there is only limited related work, which we will briefly introduce below. We look forward to seeing more research being pursued in this area in the future.

Next, we introduce evaluation, AREAs, and adversarial ranking attacks in IR.

5.2.3 Evaluation. The evaluation of adversarial attacks includes both attack performance and naturalness performance.

Attack Performance. Attack performance mainly refers to the degree of ranking improvement after a target document has been attacked. In general, the following automatic metrics for attack performance are widely adopted:

— **Attack success rate (ASR)**/SR [100, 105, 182], which evaluates the percentage of target documents successfully boosted under the corresponding target query;
— *Average boosted ranks* (Boost/Avg.boost) [100, 104, 105], which evaluates the average improved rankings for each target document under the corresponding target query;
— *Boosted top-K rate* (TKR) [100, 105], which evaluates the percentage of target documents that are boosted into top-K w.r.t. the corresponding target query; and

— **Normalized ranking shifts (NRS)** *rate* [104, 177], which evaluates the relative ranking improvement of adversarial examples that are successfully recalled into the initial set with K candidates:

$$\text{NRS} = (\Pi_d - \Pi_{d^{adv}})/\Pi_d \times 100\%, \tag{23}$$

where Π_d and $\Pi_{d^{adv}}$ are the rankings of d and d^{adv}, respectively, produced by the target IR model.

The effectiveness of an adversary is better with a higher value for all these metrics.

Naturalness Performance. Naturalness performance refers primarily to the degree to which a target document is imperceptible to humans after it has been attacked. In general, the following automatic metrics for naturalness performance and human evaluation are widely adopted:

— *Automatic spamicity detection*, which can detect whether target pages are spam or not; the utility-based term spamicity method OSD [200] is usually used to detect the adversarial examples;
— *Automatic grammar checkers*, i.e., Grammarly[6] and Chegg Writing,[7] which calculate the average number of errors in the adversarial examples;
— *Language model perplexity*, which measures the fluency using the average perplexity calculated using a pre-trained GPT-2 model [141]; and
— *Human evaluation*, which measures the quality of the attacked documents w.r.t. aspects of imperceptibility, fluency, and semantic similarity [100, 105, 183].

5.2.4 AREA. In this subsection, we introduce the task definition of AREAs and methods to achieve such attacks.

Task Definition. The AREA task is designed for attacks against DRMs. The objective of AREAs is centered around the manipulation of a document that initially fails to be recalled. By integrating adversarial perturbations, the aim is to ensure that this document is subsequently retrieved by the first-stage neural retrieval model, thereby securing its presence within the candidate set. This approach not only challenges the robustness and reliability of neural retrieval systems but also raises significant questions regarding the integrity of information authenticity [104, 199]. The goal of AREAs under top-K ranked results can be formalized as:

$$\max_{p} \ \left(K - \text{Recall}_f\left(q, d \oplus p\right) + \lambda \cdot \text{Sim}\left(d, d \oplus p\right)\right), \tag{24}$$

where $\text{Recall}_f\left(q, d \oplus p\right)$ denotes the recalled position of the perturbed document $d \oplus p$ produced by the first-stage retrieval model f with respect to query q given the entire corpus. A smaller value of Recall denotes a higher ranking.

Method. Current retrieval attack methods mainly include corpus poison attacks, backdoor attacks, and encoding attacks.

— *Corpus poison attack*: Corpus poisoning attacks construct adversarial samples against a specific query and inject them into the corpus in the inference phase so that they are recalled. MCARA [104] addresses this issue and attempts to mine the vulnerability of DRMs. It introduces the AREA task, which intends to deceive a DRM into retrieving a target document that is outside the initial set of candidate documents. Zhong et al. [199] adopt the HotFlip method [46] from NLP to iteratively add perturbations in the discrete token space to maximize its

[6] https://app.grammarly.com/.
[7] https://writing.chegg.com/.

similarity to a set of queries. In this way, they maximize the contamination of the corpus by a limited number of documents. MAWSEO [98] employs adversarial revisions to achieve real-world cybercriminal objectives, including rank boosting, vandalism detection evasion, topic relevancy, semantic consistency, user awareness (but not alarming) of promotional content, and so on.
In addition, there has been some work that has uncovered special sensitivities of retrieval models. For example, MacAvaney et al. [119] find that the high sensitivity of some models to word and sentence order is also biased towards recalling factually correct text (rather than just relevant text), and Weller et al. [178] find that denser retrieval models with dual-encoder architectures are weaker at discriminating the relevance of content that contains negative sentences.

— *Backdoor attack*: Backdoor attacks inject a small proportion of ungrammatical documents into the corpus. When user queries contain grammatical errors, the model will recall the learned triggering pattern and assign high relevance scores to those documents. Long et al. [109] introduces a novel scenario where the attackers aim to covertly disseminate targeted misinformation, such as hate speech or advertisements, through a retrieval system. To achieve this, they propose a backdoor attack triggered by grammatical errors and ensure that attack models can function normally for standard queries but are manipulated to return passages specified by the attacker when users unintentionally make grammatical mistakes in their queries.
— *Encoding attack*: By imperceptibly perturbing documents using uncommon encoded representations, encoding attacks control results across search engines for specific search queries. Boucher et al. [17] make words look the same as they originally do by adding an offset encoding to them, while the search engines are deceived. The experiment on a mirror of Simple Wikipedia shows that the proposed method can successfully deceive search engines in realistic scenarios.

5.2.5 Adversarial Ranking Attack. In this subsection, we introduce the task definition of adversarial ranking attacks and methods to achieve adversarial ranking attacks.

Task Definition. The adversarial ranking attack task is designed to attack against NRMs. Adversarial ranking attacks typically involve introducing adversarial perturbations to a document already present in the candidate set, to manipulate its ranking position either elevating or diminishing it as determined by a NRM. The goal of adversarial ranking attacks under top-K ranked results can usually be formalized as:

$$\max_{p} \left(K - \text{Rank}_f\left(q, d \oplus p\right) + \lambda \cdot \text{Sim}\left(d, d \oplus p\right)\right), \tag{25}$$

where $\text{Rank}_f\left(q, d \oplus p\right)$ denotes the position of the perturbed document $d \oplus p$ in the final ranked list generated by the NRM f with respect to query q. A smaller value of Rank denotes a higher ranking.

Method. Adversarial ranking attacks against NRMs include word substitution attacks, trigger attacks, and prompt attacks.

— *Word substitution attack*: Word substitution attacks typically boost a document's ranking by replacing a small number of words in the document with synonyms. A common method of *white-box* word substitution attack is the gradient-based attack, where the attacker uses the gradient of the loss function with respect to the input data to create adversarial examples. These examples are designed to cause the model to make incorrect relevance predictions or

rankings. Raval and Verma [145] present a systematic approach of using adversarial examples to measure the robustness of popular ranking models. They follow an approach that is similar to one used in text classification tasks [95] and perturb a limited number of tokens (with a minimum of one) in documents, replacing them with semantically similar tokens such that the rank of the document changes. Brittle-BERT [177] adds/replaces a small number of tokens to a highly relevant or non-relevant document to cause a large rank demotion or promotion. The authors find a small set of recurring adversarial words that, when added to documents, result in successful rank demotion/promotion of any relevant/non-relevant document, respectively. As for *black-box* word substitution attacks, in the field of ML, Szegedy et al. [164] find that adversarial examples have the property of cross-model transferability, i.e., the adversarial example generated by a surrogate model can also fool a target model. Black-box attacks in IR usually adopt this transfer-based paradigm due to the excellent performance of imitation of the target model. Wu et al. [182] propose the first black-box adversarial attack task against NRMs, the **word substitution ranking attack (WSRA)** task. The WSRA task aims to fool NRMs into promoting a target document in rankings by replacing important words in its text with synonyms in a semantic-preserving way. Based on this task, the authors propose a novel pseudo-relevance-based adversarial ranking attack method, which outperforms web spamming methods by 3.9% in ASR. The WSRA task focuses only on attacks on single query-document pairs and does not take into account the dynamic nature of search engines. Based on this, Liu et al. [105] introduce the topic-oriented adversarial ranking attack task, which aims to find an imperceptible perturbation that can promote a target document in ranking for a group of queries with the same topic.

— *Trigger attack*: Jiang et al. [79] find that NRMs are more sensitive to the text at the front of the position through the analysis of information bottleneck in the ranking models. Trigger attacks boost document rankings by injecting a generated trigger sentence into a specific location in the document (e.g., the beginning). Song et al. [160] propose using semantically irrelevant sentences (semantic collisions) as perturbations. They develop gradient-based approaches for generating collisions given white-box access to an NRM. Goren et al. [59] propose a document manipulation strategy to improve document quality for the purpose of improving document ranking. Liu et al. [100] propose a trigger attack method, PAT, empowered by a pairwise objective function, to generate adversarial triggers, which cause premeditated disorderliness with very few tokens. TRAttack [161] uses rewriting existing sentences in the text to improve document ranking with learning ability from the multi-armed bandit mechanism.

— *Prompt attack*: Prompt attacks use prompts to guide a language model to generate perturbations based on existing documents to improve document ranking. Chen et al. [29] propose a framework called **imperceptible document manipulation (IDEM)** to produce adversarial documents that are less noticeable to both algorithms and humans. IDEM finds the optimal connect sentence to insert into the document through a language model. Parry et al. [135] analyze the injection of query-independent prompts, such as "true" into documents and find that the prompt perturbation method is valid for several sequence-to-sequence relevance models like monoT5 [130].

— *Multi-granular attack*: Multi-granular attacks focus on generating high-quality adversarial examples by incorporating multi-granular perturbations, i.e., word level, phase level, and sentence level. Liu et al. [107] propose RL-MARA, a reinforcement learning framework, to navigate an appropriate sequential multi-granular ranking attack path. By incorporating word-level, phrase-level, and sentence-level perturbations to generate imperceptible adversarial examples, RL-MARA is able to increase the flexibility of creating adversarial examples, thereby improving the potential threat of the attack. In addition to gradient-based attacks, Parry

et al. [136] propose the use of LLMs to generate entity-specific promotional text for query-agnostic ranking attacks. By crafting text with strategic token placements and leveraging the transformer's tendency to prioritize certain positions, the method achieves query-agnostic content injection. Generated confrontation content may include specific words, phrases, or even descriptions. Experiments demonstrate that the attack successfully manipulates search rankings to promote target content, even when unrelated to user queries.

5.3 Adversarial Defenses

With the advent of SEO, many defenses have been created to counter malicious attacks. In the field of adversarial defenses, much work has been devoted to training robust neural IR models or identifying malicious adversarial examples in advance.

5.3.1 IR Defense Task. The primary objective of defenses in IR is to maintain, or even enhance, the performance of IR models when the test dataset includes adversarial examples. This involves the implementation of strategies during the training or inference phases: the goal is to ensure that the model's ability to accurately retrieve relevant documents remains uncompromised, even in the presence of manipulative adversarial perturbations.

Without loss of generality, given a test set $\mathcal{D}_{\text{test}}$ and an adversarial document set D_{adv}, the goal of adversarial defense against an neural IR model f under top-K ranked results can usually be formalized as:

$$\max \mathcal{R}_M \left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}'_{\text{test}}, K \right) \text{ such that } \mathcal{D}'_{\text{test}} \leftarrow \mathcal{D}_{\text{test}} \cup D_{\text{adv}}. \tag{26}$$

The adversarial defense task is in the training or testing phase. In the testing phase, it is usually in the form of attack detection. The training phase is usually in the form of both empirical defense and certified robustness. We present pseudo-code to illustrate a fundamental IR defense, as shown in Algorithm 2. A detailed categorization of adversarial defenses in IR is shown in Figure 11.

Next, we introduce the evaluation, attack detection, empirical defense, and certified robustness of adversarial defense in IR.

5.3.2 Evaluation. Adversarial defense assessment includes metrics for the training phase and metrics for the inference phase. Specifically, defenses for the training phase mainly comprise of empirical defenses and certified robustness; and defenses for the inference phase mainly concern the detection of adversarial samples.

Metrics Used in the Training Phase. Metrics in the training phase are mainly for evaluating the ability of empirical defenses and certified robustness methods to maintain the original ranked list in the presence of adversarial samples:

— *CleanMRR@k* evaluates MRR [36] performance on a clean dataset;
— *RobustMRR@k* [108] evaluates the MRR performance on the attacked test dataset;
— *ASR* [108] evaluates the percentage of the after-attack documents that are ranked higher than the original documents; and
— *Location square deviation* [108] evaluates the consistency between the original and perturbed ranked list for a query, by calculating the average deviation between the document positions in the two lists.

Metrics Used in the Inference Phase. Metrics in the inference phase are mainly used for evaluating the ability of attack detection methods to accurately recognize adversarial samples:

— *Point-wise detection accuracy* [28] evaluates the correctness of the detection of whether a single document has been perturbed or not;

Algorithm 2: Adversarial Defense in Neural IR Models

Require:
A query collection Q, a corpus C, a neural IR model f, potentially adversarial documents
Ensure: Safe and robust document rankings
Procedure Attack Detection
Train a detector model g using benign and adversarial examples from Q and C
for each document $d \in C$ **do**
  Compute the probability $g(d)$ of d being adversarial
  if $g(d) >$ threshold **then**
    Flag document d as adversarial
  end if
end for
Procedure Empirical Defense
Fine-tune f on adversarial examples using augmented training set Q', C'
for each training iteration **do**
  Apply random transformations to documents in C' to simulate adversarial perturbations
  Update f to minimize loss on transformed documents
end for
Procedure Certified Robustness
Incorporate certified defense methods into f (e.g., randomized smoothing)
for each query $q \in Q$ **do**
  Compute a robustness certificate for the ranking produced by $f(q)$
  if certificate fails **then**
    Adjust f to enhance robustness for q
  end if
end for
return Updated and defended neural IR model f

— *#DD* [28] denotes the average number of discarded documents ranked before the relevant document; and
— *#DR* [28] denotes the average number of discarded relevant documents.

5.3.3 Attack Detection. While progress in empirical defenses and certified robustness aids in training NRMs to be more robust in their defense against potential attacks, the detection of adversarial documents has also been explored.

Perplexity-Based Detection. Perplexity-based detection mainly uses the difference in the distribution of perplexity between the adversarial samples and the original document under the language model for recognition. Adversarial perturbations applied to original documents can significantly impact the semantic fluency of their content [29, 100, 104]. Song et al. [160] have developed a perplexity-based detection to counter ranking attacks. Detection involves using a pre-trained language model to assess the perplexity of documents, where higher perplexity values indicate less fluent text. Consequently, any document surpassing a certain perplexity threshold is filtered out from consideration.

Language-Based Detection. Language-based detection primarily uses "unnatural" language to identify adversarial samples. Adversarially generated or modified documents often exhibit grammatical inconsistencies or lack context coherence [29, 155], since documents that have been maliciously

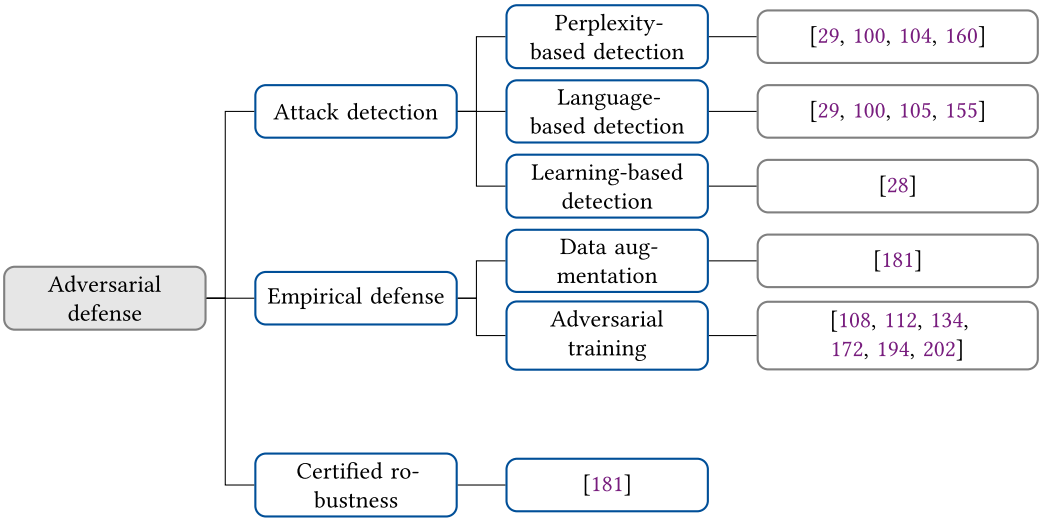

Fig. 11. Classification of adversarial defenses in neural IR.

perturbed are usually grammatically incorrect and incoherent. Some work uses grammar checkers, i.e., Grammarly and Chegg Writing, to detect adversarial examples [100, 105]. The number of grammatical errors in the target document according to the grammar checkers and the quality scoring are used as indicators of linguistic acceptability. Any document deemed to have poor linguistic acceptability is subsequently discarded.

Learning-Based Detection. Learning-based detection uses neural networks to model the characteristics of adversarial samples, and empirically learns to distinguish adversarial samples from clean samples. As mentioned earlier, spamming, perplexity-based, and language-based detectors lack knowledge of the adversarial documents, potentially leading to sub-optimal performance. Consequently, Chen et al. [28] introduce two kinds of detection task, namely point-wise and list-wise detection, to standardize evaluation processes of the efficacy of adversarial ranking detection methods. They fine-tune BERT and RoBERTa models using the original and adversarial document pairs present in the training set of the generated dataset. Experimental results demonstrate that a supervised classifier can effectively mitigate known attacks, the detection accuracy can be up to 99.5%, but it performs poorly against unseen attacks.

5.3.4 Empirical Defense. Empirical defenses attempt to make models empirically robust to known adversarial attacks; this has been extensively explored in image retrieval [120] and text classification [176]. The aim of an empirical defense is to find adversarial examples during training and use them to augment the training set.

Data Augmentation. Data augmentation often employs randomized or heuristic methods to transform the training samples, thereby expanding the training data. In the training phase, models that have seen augmented data will have some defense against adversarial samples. There are many data augmentation methods in NLP that achieve good defense performance [77, 148].

In IR, Wu et al. [181] are the first to apply data augmentation to the defense of NRMs. They find that data augmentation can reduce the ASR to some extent. However, it performs worse than

customized defenses for IR. Hence, simply augmenting the training documents (as in NLP) is not a robust defense against attacks in IR.

Adversarial Training. Adversarial training is one of the most effective defenses against specific seen attacks. By integrating pre-constructed adversarial samples into training data, adversarial training has demonstrated a strong defense in both CV and NLP [120, 203].

In IR, there is a body of work that implements adversarial training by means of adversarial optimization [172, 194, 202], such as GAN [55]. In this context, the goal is often to simply improve the effectiveness of the IR model. Meanwhile, there is work by Lupart and Clinchant [112] and Park and Chang [134] who attempt to use adversarial training for improving robustness. By incorporating adversarial examples into training data, adversarial training has become the de facto defense approach to adversarial attacks against NRMs. However, this defense mechanism is subject to a tradeoff between effectiveness and adversarial robustness.

To tackle this issue, Liu et al. [108] define the perturbation invariance of a ranking model and design a **perturbation-invariant adversarial training (PIAT)** method for ranking models to achieve a better effectiveness-robustness tradeoff. Experimental results on several ranking models demonstrate the superiority of PIAT compared to earlier adversarial defenses.

5.3.5 Certified Robustness. Since empirical defenses are only effective for certain attacks rather than all attacks, competition emerges between adversarial attacks and defense methods. To solve the attack-defense dilemma, researchers resort to certified defenses to make models provably robust to certain kinds of adversarial perturbations. In NLP, Jia et al. [78] have been the first to propose to certify the robustness of adversarial word substitutions by using interval bound propagation [44].

In the field of IR, Wu et al. [181] propose a rigorous and provable certified defense method for NRMs. They define certified Top-K robustness for ranking models since users mainly care about the top-ranked results in real-world scenarios. Then, they propose CertDR to achieve certified top-K robustness, based on the idea of randomized smoothing. Their experiments demonstrate that CertDR can significantly outperform state-of-the-art empirical defense methods for ranking models.

5.4 Benchmark Datasets

In this section, we present the datasets commonly used for studying adversarial robustness. Prior work on attacks and defenses against robustness has focused on experiments on previously published IR datasets, as shown in Table 3. All datasets can be found in the BestIR benchmark; see Appendix B for an overview of the resources collected in BestIR.

Basic Datasets. Some datasets in IR are adapted for reuse by attack and defense methods as basic datasets. These include MS MARCO document/passage [127] and Clueweb-09B [34]. Some work [104, 107, 182] performs experiments against attacks and defenses directly on these datasets. For example, prior work usually uses the training set of the basic dataset to train an NRM as the target model [107, 182]; the queries in the development set are used as target queries, and a portion of the documents in the ranked list of each query is sampled as target documents. Attacks on these target documents measure the performance of adversarial attack methods. We will discuss the specific evaluation methods in Section 5.2.

Expansion of Datasets. Some collaborative benchmarks, such as TREC DL19 [39] and TREC DL20 [38], have provided additional query collections for evaluation against the base dataset. Similarly, these query collections can be used to evaluate the effectiveness of attack methods. Queries in these datasets are often attacked as additional sets of targeted queries. For example, existing work

Table 3. Benchmark Datasets for Studying Adversarial Robustness

Dataset	#Documents	#Q_{train}	#Q_{dev}	#Q_{eval}	References
MS MARCO document [127]	3.2M	370K	5,193	5,793	[100, 104, 107, 181, 182]
MS MARCO passage [127]	8.8M	500K	6,980	6,837	[29, 104, 108, 177, 181, 182, 199]
Clueweb09-B [34]	50M	150	–	–	[107, 177]
Natural Questions [88]	21M	60K	8.8K	3.6K	[100, 109, 199]
TriviaQA [80]	21M	60K	8.8K	11.3K	[109]
TREC DL19 [39]	–	–	43	–	[100, 135, 177]
TREC DL20 [38]	–	–	54	–	[135]
TREC MB14 [97]	–	–	50	–	[100]
ASRC [142]	1,279	–	31	–	[58, 183]
Q-MS MARCO [105]	–	–	4,000	–	[105]
Q-Clueweb09 [105]	–	–	292	–	[105]
DARA [28]	164k	50k	3,490	3,489	[28]

#Documents denotes the number of documents in the corpus; #Q_{train} denotes the number of queries available for training; #Q_{dev} denotes the number of queries available for development; #Q_{eval} denotes the number of queries available for evaluation.

usually trains an NRM on the MS MARCO passage dataset and uses the 43 queries in TREC DL19 as the target queries to perform attacks [29, 100].

Off-the-Shelf Datasets. Some research has adapted the above datasets to construct new datasets that can be used directly to perform attacks or evaluate defenses. For example, ASRC [142] is based on documents in Clueweb that are manually modified to generate new adversarial samples for evaluating the model's adversarial defense abilities. Existing work has also used it to study the effects of manual manipulation of documents on IR systems [58, 168]. To perform a topic-oriented attack, Liu et al. [105] construct query groups on the same topic based on ORCAS [37] and the TREC 2012 Web Track [34] as a complement to MS MARCO document and Clueweb-09b, respectively. DARA [28] is a dataset for detecting adversarial ranking attacks and includes two types of detection task for adversarial documents.

6 Open Issues and Future Directions

In addition to the significant progress documented in Sections 4 and 5 above, robust neural IR presents several remaining challenges and opportunities for future research.

6.1 Remaining Challenges and Issues

In this subsection, we identify challenges and issues in robust neural IR with a special focus on adversarial robustness and OOD robustness, respectively.

6.1.1 Challenges on Adversarial Robustness in IR. Many key issues in adversarial robustness in IR have not received much attention yet.

Penetration Attacks against the Whole "Retrieval-Ranking" Pipeline. As explained in Sections 4 and 5, there is a considerable body of work that focuses on attacking the first-stage retrieval [104, 199] and re-rank stage [100, 182] separately. Penetration attacks focus on exploiting vulnerabilities within the retrieval re-rank pipeline, a critical component in IR systems that ranks results based on relevance to the query. These attacks aim to manipulate rankings by identifying and exploiting weaknesses in the pipeline's design or its underlying algorithms. This manipulation can result

in irrelevant or malicious content being ranked higher than genuine content, compromising the integrity and reliability of the IR system.

Universal Attacks. Universal attacks represent a form of adversarial threat that is particularly challenging due to its generalizability across different models and instances [170]. Unlike targeted attacks that aim at specific vulnerabilities within a system, universal attacks exploit common weaknesses that are present across a wide range of systems. This makes it difficult to defend against them, as they require solutions that are not just effective for a single model or instance but across the entire spectrum of possible configurations. The development of robust defenses against universal attacks is therefore a significant challenge that demands innovative approaches and a deep understanding of the underlying vulnerabilities.

Dynamic Attack Scenarios. Most prior work on adversarial attacks is based on static assumptions about search engines [29, 100, 182]. However, search engines operate within a dynamic landscape, which may include changes such as the expansion or reduction of the corpus, and the demotion of documents suspected of spamming in their rankings. In search engines, the search engine results page for a query is constantly changing. Research indicates that, in the dynamic environment of search engines, current attack methods struggle to maintain a consistent ranking advantage [105]. Therefore, designing attack methods that fully consider the dynamic nature of search engines is both practically significant and challenging. At the same time, it is worthwhile to explore the development of defense methods that evolve in tandem with updates to the search engine.

Gaming in Search Engines. The phenomenon of "gaming" in search engines, where individuals or entities manipulate search results for competitive advantage, poses a significant challenge to maintaining the integrity and relevance of search outcomes [87]. This competitive manipulation not only undermines the quality of information presented to users but also erodes trust in the search engine's ability to deliver unbiased and accurate results. As search engines evolve into more sophisticated platforms, so too do the methods employed by those looking to exploit their algorithms for personal gain [142]. This ongoing battle between search engines and gamers necessitates the development of more advanced detection and mitigation techniques that can adapt to new gaming strategies, preserving the search engine's role as a reliable source of information.

Defense against Unseen Attacks. In IR counter defenses, it is often the case that empirical defenses where the attack method is known, can yield good results [108]. Whereas in real scenarios, the attack methods are multiple and potentially unknown. Defending against unseen attacks is a paramount challenge in enhancing adversarial robustness in IR systems. These attacks are particularly daunting because they exploit new or unknown vulnerabilities, making traditional defense mechanisms, which are often designed to combat known threats, ineffective. The key to overcoming this challenge lies in the development of adaptive, intelligent systems capable of anticipating potential threats and dynamically adjusting their defense mechanisms in real-time. Achieving this level of adaptability and foresight requires a profound shift in the current paradigms of security in IR, embracing more proactive and predictive approaches.

Defense in Practice. Implementing effective defense mechanisms in practice is a balancing act between effectiveness, efficiency, and cost. Effective defense strategies are those that can accurately detect and neutralize threats without significantly impacting the user experience or the relevance of search results. However, the computational resources required for these strategies often come with high costs and can affect the efficiency of the search engine, leading to slower response times and decreased user satisfaction. To address these challenges, search engines are increasingly turning to ML and artificial intelligence technologies that can provide scalable and cost-effective solutions [93,

117]. These technologies enable the development of adaptive defense mechanisms that can learn from attack patterns and evolve over time, offering a dynamic approach to security that maintains the delicate balance between protecting the search engine and preserving its performance.

6.1.2 Challenges on OOD Robustness in IR. Several key issues in OOD robustness in IR have not received much attention yet.

Reliance on Large-Scale Data. The reliance on large-scale datasets for training and evaluating IR systems poses significant challenges in ensuring OOD robustness. Large datasets often contain biases and do not necessarily represent the diversity of real-world scenarios [82], leading to models that perform well on seen data but poorly on unseen, OOD examples [167, 190]. Addressing this challenge requires innovative approaches to data collection and model training that prioritize diversity and real-world applicability, ensuring that IR systems remain reliable and effective across a broad range of OOD scenarios.

Lack of Harmonized Benchmarks for Multiple OOD Scenarios. A major burden in enhancing OOD robustness in IR is the lack of harmonized benchmarks that accurately reflect the multitude of real-world, OOD scenarios. Without standardized benchmarks, it is difficult to assess the true robustness of IR systems across different contexts and to identify areas for improvement. Developing these benchmarks involves not only capturing a wide range of OOD scenarios but also ensuring that they are representative of the actual challenges faced by IR systems in practice. This effort is crucial for advancing the state of OOD robustness in IR.

Enhancing Continuous Corpus Adaptation for DRMs. As we have seen in Sections 4 and 5, prior work mainly considers the problem of one-shot adaptation of DRMs to new corpora [35, 167]. However, in real search engines, new documents are constantly added to the search engine and bring in a variety of new domains, which poses a challenge to the continuous adaptation ability of DRMs. To address this, it is essential to develop DRMs that not only quickly adapt to new corpora in a one-shot learning scenario but also continuously learn and adjust as new data is introduced. This requires innovative approaches that can dynamically update the models in an incremental fashion without the need for frequent retraining from scratch. Techniques such as online learning, transfer learning, and meta-learning can play pivotal roles in enhancing the continuous corpus adaptation of DRMs, ensuring they remain effective and relevant in the ever-changing search landscape.

Improving OOD Generalization of NRMs. Approaches to improve OOD generalization ability mainly target neural retrieval models [167, 190, 192]. More generally, the OOD robustness of NRMs should be optimized along with retrieval models. Enhancing the OOD generalization of NRMs involves developing models that can effectively handle queries that deviate significantly from the training distribution, thereby ensuring the retrieval of relevant and accurate results under a wide range of search scenarios. This challenge requires a multifaceted approach, incorporating advanced ML techniques such as robust representation learning, anomaly detection, and domain adaptation strategies. By prioritizing the OOD robustness of NRMs alongside DRMs, search engines can significantly improve their ability to serve high-quality, relevant content to users, even when faced with novel or unexpected queries.

Improving OOD Robustness in Practice. Enhancing OOD robustness in practice is essential for maintaining the effectiveness, efficiency, and cost-effectiveness of search engines. To achieve this, search engines must be able to accurately identify and process queries that fall outside the typical distribution of observed data, ensuring that even uncommon or unseen queries return relevant and useful results. Implementing robust OOD handling mechanisms can significantly improve the quality of search results [103, 167], but this often requires sophisticated algorithms that can detect

and adapt to OOD queries in real-time [137, 207]. While these algorithms can be computationally intensive, leading to higher operational costs, the investment in OOD robustness can ultimately enhance user satisfaction and trust in the search engine [103, 205]. Moreover, optimizing these algorithms for efficiency can help mitigate additional costs, ensuring that improvements in OOD robustness also align with the search engine's operational goals.

6.2 Challenges and Opportunities Posed by LLMs

LLMs have gained attention for their generative abilities. There have been several works applying them to IR with good results. The introduction of LLMs may bring new robustness problems and also provide new solutions to known robustness problems.

6.2.1 New Challenges to IR Robustness from LLMs. Recently, there has been a lot of exploratory work on using LLMs for IR tasks. However, these attempts may pose new robustness challenges to LLMs-based IR methods due to the robustness issues of LLMs themselves.

New Challenges to OOD Robustness. LLMs have shown biases and input sensitivities [50, 76], and these will affect the OOD robustness of IR systems: (1) the training process of LLMs often introduces domain bias due to the limited representation of real-world diversity in the training data, which can result in degraded performance when handling out-of-domain queries or documents; existing work also reveals that neural IR models may prefer documents generated by LLMs in corpora [40]; and (2) LLMs are highly sensitive to slight variations in input [30, 50, 76], potentially leading to inconsistent IR outcomes. Addressing both the domain bias and input sensitivity is crucial for developing robust and generalizable IR systems with LLMs.

New Challenges to Adversarial Robustness. When applied to IR systems, the adversarial vulnerability of the LLMs themselves is imported at the same time: (1) LLMs used as ranking models [162] are susceptible to hallucinations, generating plausible yet factually incorrect or irrelevant information, which can lead to the retrieval of misleading data and undermine the reliability of IR systems [107]; and (2) the large scale and opacity of LLMs complicate the diagnosis and mitigation of vulnerabilities, making defensive measures technically challenging and resource-intensive. Addressing hallucinations and managing defense costs are critical for ensuring the integrity and credibility of LLM-based IR systems [157, 184].

6.2.2 New Opportunities for IR Robustness via LLMs. While the use of LLMs may introduce new robustness risks, the power of LLMs also provides new ideas for improving robustness. In the field of NLP, several publications enhance the robustness of NLP models using LLMs [e.g., 67, 125, 191], but not so much yet in IR.

New Opportunities to OOD Robustness. The powerful generation and language understanding ability of LLMs can help to improve the OOD robustness of IR systems. (1) *Synthesizing OOD training data with LLMs.* Some preliminary attempts have revealed that LLMs perform well in generating relevant queries for unannotated documents [16, 49, 143]. LLMs may also generate diverse and complex data that mirror OOD scenarios, providing IR systems with the training needed to better handle unfamiliar or novel situations. Such synthetic data can help improve the generalizability and robustness of IR models against OOD inputs. (2) *LLMs for OOD detection.* Using LLMs' abilities in language understanding, to detect and manage OOD queries has not been carefully explored. By identifying queries that deviate from the training distribution, LLMs may trigger specialized handling for such cases, thereby enhancing the robustness and reliability of IR systems [157, 184].

New Opportunities to Adversarial Robustness. LLMs hold promise for improving the adversarial robustness of IR systems through their ability to generate and predict adversarial examples. (1) *Generating adversarial examples with LLMs.* LLMs have been shown to achieve preliminary success in generating adversarial examples against IR systems [106, 136]. By exposing systems to a wider array of adversarial tactics during training, LLMs can help develop more robust IR models. (2) *IR model defense assisted with LLMs.* LLMs hold promising potential in assisting the development of defense mechanisms by predicting and countering adversarial strategies, which is worth further exploration. Leveraging LLMs in simulation environments to anticipate potential attacks could also enable the proactive enhancement of IR systems.

7 Conclusion

The landscape of IR has evolved significantly with the advent of neural methods. This evolution has brought with it new challenges in the robustness of IR systems. These robustness issues undermine the trust of users in search engines, making it a critical concern for both researchers and practitioners in the field.

In this survey, we have organized the IR literature on various forms of robustness, focusing particularly on IID robustness, OOD robustness, and adversarial robustness. We have also discussed the remaining challenges of these fields, as well as potential future directions for research. This survey contributes to that journey by providing a structured overview of the current state-of-the-art, offering a roadmap for future research directions, and inspiring continued exploration and innovation in the field. While significant progress has been made in understanding and robustness of IR, there is still much work to be done. As the field continues to evolve, it will be crucial to develop robust defenses against these attacks, to ensure the integrity of search results and maintain the trust of users.

In conclusion, robust neural IR represents a complex and multifaceted problem space, but also an opportunity for innovative research and development. As we look forward, the goal of developing IR systems that are not only robust but also adaptable, trustworthy, and user-centric is essential, promising to redefine the boundaries of what is possible in IR.

Acknowledgments

We thank our reviewers and associate editor for their constructive feedback and suggestions.

References

[1] Amin Abolghasemi, Suzan Verberne, Arian Askari, and Leif Azzopardi. 2023. Retrievability bias estimation using synthetically generated queries. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3712–3716.

[2] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. 2021. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access* 9 (2021), 155161–155196.

[3] Abhijit Anand, Jurek Leonhardt, Jaspreet Singh, Koustav Rudra, and Avishek Anand. 2024. Data augmentation for sample efficient and robust document ranking. *ACM Transactions on Information Systems* 42, 5, Article 119 (September 2024), 29 pages. DOI: https://doi.org/10.1145/3634911

[4] Avishek Anand, Procheta Sen, Sourav Saha, Manisha Verma, and Mandar Mitra. 2023. Explainable information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3448–3451.

[5] A. Anaya-Isaza and Leonel Mera-Jiménez. 2022. Data augmentation and transfer learning for brain tumor detection in magnetic resonance imaging. *IEEE Access* 10 (2022), 23217–23233. DOI: https://doi.org/10.1109/ACCESS.2022.3154061

[6] Mordechai Averbuch, Tom H. Karson, Benjamin Ben-Ami, Oded Maimon, and Lior Rokach. 2004. Context-sensitive medical information retrieval. In *Proceedings of the 11th World Congress on Medical Informatics (Medinfo '04)*. IOS Press, 282–286.

[7] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 561–570.

[8] Shariq Bashir. 2014. Estimating retrievability ranks of documents using document features. *Neurocomputing* 123 (2014), 216–232.
[9] Shariq Bashir and Andreas Rauber. 2010. Improving retrievability of patents in prior-art search. In *Proceedings of the European Conference on Information Retrieval*. Springer, 457–470.
[10] Shariq Bashir and Andreas Rauber. 2017. Retrieval models versus retrievability. In *Current Challenges in Patent Information Retrieval*. M. Lupu, K. Mayer, N. Kando, and A. Trippe (Eds.). The Information Retrieval Series, Vol. 37, Springer, Berlin, 185–212. DOI: https://doi.org/10.1007/978-3-662-53817-3_7
[11] Nicholas J. Belkin, Paul Kantor, Edward A. Fox, and Joseph A. Shaw. 1995. Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management* 31, 3 (1995), 431–448.
[12] Rodger Benham, J. Shane Culpepper, Luke Gallagher, Xiaolu Lu, and Joel M. Mackenzie. 2018. Towards efficient and effective query variant generation. In *Proceedings of the Conference on Design of Experimental Search Information REtrieval Systems (DESIRES)*, 62–67.
[13] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. 2021. Understanding robustness of transformers for image classification. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 10211–10221.
[14] Amin Bigdeli, Negar Arabzadeh, and Ebrahim Bagheri. 2024. Learning to jointly transform and rank difficult queries. In *Proceedings of the European Conference on Information Retrieval*. Springer, 40–48.
[15] Toine Bogers and Vivien Petras. 2017. Supporting book search: A comprehensive comparison of tags vs. controlled vocabulary metadata. *Data and Information Management* 1, 1 (2017), 17–34.
[16] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Data augmentation for information retrieval using large language models. arXiv:2202.05144. Retrieved from https://arxiv.org/abs/2202.05144
[17] Nicholas Boucher, Luca Pajola, Ilia Shumailov, Ross Anderson, and Mauro Conti. 2023. Boosting big brother: Attacking search engines with encodings. arXiv:2304.14031. Retrieved from https://arxiv.org/abs/2304.14031
[18] Yinqiong Cai, Keping Bi, Yixing Fan, Jiafeng Guo, Wei Chen, and Xueqi Cheng. 2023. L2R: Lifelong learning for first-stage retrieval with backward-compatible representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 183–192.
[19] ZeFeng Cai, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, Xin Alex Lin, Liang He, and Daxin Jiang. 2022. HypeR: Multitask hyper-prompted training enables large-scale retrieval generalization. In *Proceedings of the 11th International Conference on Learning Representations*.
[20] Daniel Campos, ChengXiang Zhai, and Alessandro Magnani. 2023. Noise-robust dense retrieval via contrastive alignment post training. arXiv:2304.03401. Retrieved from https://arxiv.org/abs/2304.03401
[21] Carlos Castillo and Brian D. Davison. 2011. Adversarial web search. *Foundations and Trends in Information Retrieval* 4, 5 (2011), 377–486.
[22] Ramraj Chandradevan, Kaustubh D. Dhole, and Eugene Agichtein. 2024. DUQGen: Effective unsupervised domain adaptation of neural rankers by diversifying synthetic query generation. arXiv:2404.02489. Retrieved from https://arxiv.org/abs/2404.02489
[23] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Continual learning for generative retrieval over dynamic corpora. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 306–315.
[24] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2022. CorpusBrain: Pre-train a generative retrieval model for knowledge-intensive language tasks. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 191–200.
[25] Ruey-Cheng Chen, Luke Gallagher, Roi Blanco, and J. Shane Culpepper. 2017. Efficient cost-aware cascade ranking in multi-stage retrieval. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 445–454.
[26] Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. 2022. Out-of-domain semantics to the rescue! Zero-shot hybrid retrieval models. In *Proceedings of the European Conference on Information Retrieval*. Springer, 95–110.
[27] Xu Chen, Zida Cheng, Shuai Xiao, Xiaoyi Zeng, and Weilin Huang. 2023. Cross-domain augmentation networks for click-through rate prediction. arXiv:2305.03953. Retrieved from https://arxiv.org/abs/2305.03953
[28] Xuanang Chen, Ben He, Le Sun, and Yingfei Sun. 2023. Defense of adversarial ranking attack in text retrieval: Benchmark and baseline via detection. arXiv:2307.16816. Retrieved from https://arxiv.org/abs/2307.16816
[29] Xuanang Chen, Ben He, Zheng Ye, Le Sun, and Yingfei Sun. 2023. Towards imperceptible document manipulations against neural ranking models. In *Findings of the Association for Computational Linguistics: ACL '23*, 6648–6664.
[30] Xinyi Chen, Baohao Liao, Jirui Qi, Panagiotis Eustratiadis, Christof Monz, Arianna Bisazza, and Maarten de Rijke. 2024. The SIFo benchmark: Investigating the sequential instruction following ability of large language models. arXiv:2406.19999. Retrieved from https://arxiv.org/abs/2406.19999

[31] Xuanang Chen, Jian Luo, Ben He, Le Sun, and Yingfei Sun. 2022. Towards robust dense retrieval via local ranking alignment. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 1980–1986.
[32] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. 2019. Improving black-box adversarial attacks with a transfer-based prior. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 10934–10944.
[33] Gobinda G. Chowdhury and Sudatta Chowdhury. 2003. *Introduction to Digital Libraries*. Facet publishing.
[34] Charles L. Clarke, Nick Craswell, and Ian Soboroff. 2009. *Overview of the TREC 2009 Web Track*. Technical Report. Waterloo University.
[35] Daniel Cohen, Bhaskar Mitra, Katja Hofmann, and W. Bruce Croft. 2018. Cross domain regularization for neural ranking models using adversarial learning. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1025–1028.
[36] Nick Craswell. 2009. Mean reciprocal rank. In *Encyclopedia of Database Systems*. L. Liu and M. T. Özsu (Eds.), Vol. 1703, Springer, Boston. DOI: https://doi.org/10.1007/978-0-387-39940-9_488
[37] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 20 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*.
[38] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. In *Proceedings of the Text REtrieval Conference*.
[39] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. In *Proceedings of the Text REtrieval Conference*.
[40] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the LLM era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6437–6447.
[41] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. arXiv:2209.11755. Retrieved from https://arxiv.org/abs/2209.11755
[42] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7 (2021), 3366–3385.
[43] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. 2021. A Systematic review of robustness in deep learning for computer vision: Mind the gap? arXiv:2112.00639. Retrieved from https://arxiv.org/abs/2112.00639
[44] Krishnamurthy Dvijotham, Sven Gowal, Robert Stanforth, Relja Arandjelovic, Brendan O'Donoghue, Jonathan Uesato, and Pushmeet Kohli. 2018. Training verified learners with learned verifiers. arXiv:1805.10265. Retrieved from https://arxiv.org/abs/1805.10265
[45] Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the International Colloquium on Automata, Languages, and Programming*. Springer, 1–12.
[46] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 31–36.
[47] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6903–6915.
[48] Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang, and Jiafeng Guo. 2022. Pre-training methods in information retrieval. *Foundations and Trends in Information Retrieval* 16, 3 (2022), 178–317.
[49] Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. 2024. Scaling laws for dense retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1339–1349.
[50] Emilio Ferrara. 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday* 28 (2023).
[51] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural IR models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2353–2359.
[52] Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 2843–2853.
[53] Joseph L. Gastwirth. 1972. The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics* 54, 3 (1972), 306–316.

[54] Suyu Ge, Chenyan Xiong, Corby Rosset, Arnold Overwijk, Jiawei Han, and Paul Bennett. 2023. Augmenting zero-shot dense retrievers with plug-in mixture-of-memories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1796–1812.

[55] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM* 63, 11 (2020), 139–144.

[56] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR '15)*. Yoshua Bengio and Yann LeCun (Eds.), Conference Track Proceedings. Retrieved from http://arxiv.org/abs/1412.6572

[57] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[58] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2018. Ranking robustness under adversarial document manipulations. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 395–404.

[59] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2020. Ranking-incentivized quality preserving content modification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 259–268.

[60] David Graus, Daan Odijk, and Maarten de Rijke. 2018. The birth of collective memories: Analyzing emerging entities in text streams. *Journal of the Association for Information Science and Technology* 69, 6 (June 2018), 773–786.

[61] Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems* 40, 4 (2022), 1–42.

[62] Jiafeng Guo, Xueqi Cheng, Gu Xu, and Xiaofei Zhu. 2011. Intent-aware query similarity. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 259–268.

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

[64] Jiafeng Guo, Changjiang Zhou, Ruqing Zhang, Jiangui Chen, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. CorpusBrain++: A continual generative pre-training framework for knowledge-intensive language tasks. arXiv:2402.16767. Retrieved from https://arxiv.org/abs/2402.16767

[65] Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Web spam taxonomy. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Vol. 5. Citeseer, 39–47.

[66] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. 2020. Antique: A non-factoid question answering benchmark. In *Proceedings of the Advances in Information Retrieval: 42nd European Conference on IR Research (ECIR '20)*, Part II. Springer, 166–173.

[67] Alec Helbling, Mansi Phute, Matthew Hull, and Duen Horng Chau. 2023. LLM self defense: By self examination, LLMs know they are being tricked. arXiv:2308.07308. Retrieved from https://arxiv.org/abs/2308.07308

[68] Jorg Hering. 2017. The annual report algorithm: Retrieval of financial statements and extraction of textual information. In *Proceedings of the CS and IT Conference*, Vol. 7. CS and IT Conference Proceedings.

[69] Maria Heuss, Maarten de Rijke, and Avishek Anand. 2024. RankingSHAP—Listwise feature attribution explanations for ranking models. arXiv:2403.16085. Retrieved from https://arxiv.org/abs/2403.16085

[70] Peter J. Huber. 1981. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics.

[71] Michelle Chen Huebscher, Christian Buck, Massimiliano Ciaramita, and Sascha Rothe. 2022. Zero-shot retrieval with search agents and hybrid environments. arXiv:2209.15469. Retrieved from https://arxiv.org/abs/2209.15469

[72] Judit Bar Ilan. 1998. Search engine results over time: A case study on search engine stability. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics* 2 (1998), Paper 1.

[73] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. arXiv:2112.09118. Retrieved from https://arxiv.org/abs/2112.09118

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

[75] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 1 (2010), 117–128.

[76] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, Pascale Fung, et al. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys* 55, 12 (2023), 1–38.

[77] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2021–2031.

[78] 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)*, 4129–4142.

[79] Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. 2021. How does BERT rerank passages? An attribution analysis with information bottlenecks. In *Proceedings of the 4th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 496–509.

[80] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611.

[81] SeongKu Kang, Shivam Agarwal, Bowen Jin, Dongha Lee, Hwanjo Yu, and Jiawei Han. 2024. Improving retrieval in theme-specific applications using a corpus topical taxonomy. arXiv:2403.04160. Retrieved from https://arxiv.org/abs/2403.04160

[82] Robert M. Kaplan, David A. Chambers, and Russell E. Glasgow. 2014. Big data and large sample size: A cautionary note on the potential for bias. *Clinical and Translational Science* 7, 4 (2014), 342–346.

[83] Pranav Kasela, Gabriella Pasi, Raffaele Perego, and Nicola Tonellotto. 2024. DESIRE-ME: Domain-enhanced supervised information retrieval using mixture-of-experts. In *Proceedings of the European Conference on Information Retrieval*. Springer, 111–125.

[84] Omar Khattab and Matei A. Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48.

[85] Varsha Kishore, Chao Wan, Justin Lovelace, Yoav Artzi, and Kilian Q. Weinberger. 2023. IncDSI: Incrementally updatable document retrieval. In *Proceedings of the International Conference on Machine Learning*. PMLR, 17122–17134.

[86] Barbara Ann Kitchenham and Stuart Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report. Guidelines for Performing Systematic Literature Reviews in Software Engineering.

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

[88] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[89] Thiago Laitz, Konstantinos Papakostas, Roberto Lotufo, and Rodrigo Nogueira. 2024. InRanker: Distilled rankers for zero-shot information retrieval. arXiv:2401.06910. Retrieved from https://arxiv.org/abs/2401.06910

[90] Ray R. Larson. 2010. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology* 61 (Apr. 2010), 852–853. DOI: https://doi.org/10.1002/asi.21234

[91] Hyunji Lee, Luca Soldaini, Arman Cohan, Minjoon Seo, and Kyle Lo. 2023. Back to basics: A simple recipe for improving out-of-domain retrieval in dense encoders. arXiv:2311.09765. Retrieved from https://arxiv.org/abs/2311.09765

[92] Damien Lefortier, Pavel Serdyukov, and Maarten de Rijke. 2014. Online exploration for detecting shifts in fresh intent. In *Proceedings of the 23rd ACM Conference on Information and Knowledge Management (CIKM '14)*. ACM, 589–598.

[93] Jurek Leonhardt, Henrik Müller, Koustav Rudra, Megha Khosla, Abhijit Anand, and Avishek Anand. 2024. Efficient neural ranking using forward indexes and lightweight encoders. *ACM Transactions on Information Systems* 42, 5 (2024), 1–34.

[94] Minghan Li and Eric Gaussier. 2024. Domain adaptation for dense retrieval and conversational dense retrieval through self-supervision by meticulous pseudo-relevance labeling. arXiv:2403.08970. Retrieved from https://arxiv.org/abs/2403.08970

[95] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4208–4215.

[96] Davis Liang, Peng Xu, Siamak Shakeri, Cicero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Embedding-based zero-shot retrieval through query generation. arXiv:2009.10270. Retrieved from https://arxiv.org/abs/2009.10270

[97] Jimmy Lin, Miles Efron, Garrick Sherman, Yulu Wang, and Ellen M. Voorhees. 2013. Overview of the TREC-2013 microblog track. In *Proceedings of the Text REtrieval Conference (TREC)*, Vol. 2013, 21.

[98] Zilong Lin, Zhengyi Li, Xiaojing Liao, XiaoFeng Wang, and Xiaozhong Liu. 2023. MAWSEO: Adversarial Wiki search poisoning for illicit online promotion. arXiv:2304.11300. Retrieved from https://arxiv.org/abs/2304.11300

[99] Junping Liu, Mingkang Gong, Xinrong Hu, Jie Yang, and Yi Guo. 2023. MIRS:[MASK] insertion based retrieval stabilizer for query variations. In *Proceedings of the International Conference on Database and Expert Systems Applications*. Springer, 392–407.

[100] 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.

[101] Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2022. Challenges in generalization in open domain question answering. In *Findings of the Association for Computational Linguistics: NAACL '22*, 2014–2029.

[102] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proceedings of the 16th European Conference on Computer Vision (ECCV '20)*, Part X. Springer, 182–199.

[103] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Wei Chen, and Xueqi Cheng. 2023. On the robustness of generative retrieval models: An out-of-distribution perspective. arXiv:2306.12756. Retrieved from https://arxiv.org/abs/2306.12756

[104] 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.

[105] 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. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, 1700–1709.

[106] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Attack-in-the-chain: Bootstrapping large language models for attacks against black-box neural ranking models. arXiv:2412.18770. Retrieved from https://arxiv.org/abs/2412.18770

[107] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Multi-granular adversarial attacks against black-box neural ranking models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, 1391–1400.

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

[109] Quanyu Long, Yue Deng, LeiLei Gan, Wenya Wang, and Sinno Jialin Pan. 2024. Backdoor attacks on dense passage retrievers for disseminating misinformation. arXiv:2402.13532. Retrieved from https://arxiv.org/abs/2402.13532

[110] Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, and Haifeng Wang. 2022. ERNIE-Search: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval. arXiv:2205.09153. Retrieved from https://arxiv.org/abs/2205.09153

[111] Gang Luo, Chunqiang Tang, Hao Yang, and Xing Wei. 2008. MedSearch: A specialized search engine for medical information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 143–152.

[112] Simon Lupart and Stéphane Clinchant. 2023. A study on FGSM adversarial training for neural retrieval. In *Proceedings of the European Conference on Information Retrieval*. Springer, 484–492.

[113] Simon Lupart, Thibault Formal, and Stéphane Clinchant. 2023. MS-Shift: An analysis of MS MARCO distribution shifts on neural retrieval. In *Proceedings of the European Conference on Information Retrieval*. Springer, 636–652.

[114] Mihai Lupu and Allan Hanbury. 2013. Patent retrieval. *Foundations and Trends® in Information Retrieval* 7, 1 (2013), 1–97.

[115] Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1075–1088.

[116] Xiaofei Ma, Cicero dos Santos, and Andrew O. Arnold. 2021. Contrastive fine-tuning improves robustness for neural rankers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP '21*, 570–582.

[117] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Scattered or connected? An optimized parameter-efficient tuning approach for information retrieval. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 1471–1480.

[118] Xinyu Ma, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. A contrastive pre-training approach to discriminative autoencoder for dense retrieval. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM)*, 4314–4318.

[119] Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the behavior of neural IR models. *Transactions of the Association for Computational Linguistics* 10 (2022), 224–239.

[120] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*.

[121] K. Tamsin Maxwell and Burkhard Schafer. 2008. Concept and context in legal information retrieval. In *Proceedings of the 2008 conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*. IOS Press, 63–72.

[122] Sanket Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. 2023. DSI++: Updating transformer memory with new documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8198–8213.

[123] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. 2023. An empirical investigation of the role of pre-training in lifelong learning. *Journal of Machine Learning Research* 24, 214 (2023), 1–50. Retrieved from http://jmlr.org/papers/v24/22-0496.html

[124] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: Making domain experts out of dilettantes. *ACM SIGIR Forum* 55 (2021), 1–27.

[125] Raha Moraffah, Shubh Khandelwal, Amrita Bhattacharjee, and Huan Liu. 2024. Adversarial text purification: A large language model approach for defense. arXiv:2402.06655. Retrieved from https://arxiv.org/abs/2402.06655

[126] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. arXiv:2201.10005. Retrieved from https://arxiv.org/abs/2201.10005

[127] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 Co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, December 9, 2016. CEUR Workshop Proceedings, Vol. 1773. CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[128] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9844–9855.

[129] Shuzi Niu, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2012. Top-k learning to rank: Labeling, ranking and evaluation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 751–760.

[130] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP '20*, 708–718.

[131] Seokjin Oh, Su Ah Lee, and Woohwan Jung. 2023. Data augmentation for neural machine translation using generative language model. arXiv:2307.16833. Retrieved from https://arxiv.org/abs/2307.16833

[132] Ziyang Pan, Kangjia Fan, Rongyu Liu, and Daifeng Li. 2023. Towards robust neural rankers with large language model: A contrastive training approach. *Applied Sciences* 13, 18 (2023), 10148.

[133] Saran Pandian, Debasis Ganguly, and Sean MacAvaney. 2024. Evaluating the explainability of neural rankers. In *Proceedings of the European Conference on Information Retrieval*. Springer, 369–383.

[134] Dae Hoon Park and Yi Chang. 2019. Adversarial sampling and training for semi-supervised information retrieval. In *Proceedings of the World Wide Web Conference*, 1443–1453.

[135] Andrew Parry, Maik Fröbe, Sean MacAvaney, Martin Potthast, and Matthias Hagen. 2024. Analyzing adversarial attacks on sequence-to-sequence relevance models. In *Proceedings of the European Conference on Information Retrieval*. Springer, 286–302.

[136] Andrew Parry, Sean MacAvaney, and Debasis Ganguly. 2024. Exploiting positional bias for query-agnostic generative content in search. In *Findings of the Association for Computational Linguistics: ACL '24*, 11030–11047.

[137] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *Proceedings of the European Conference on Information Retrieval*. Springer, 397–412.

[138] Gustavo Penha, Enrico Palumbo, Maryam Aziz, Alice Wang, and Hugues Bouchard. 2023. Improving content retrievability in search with controllable query generation. In *Proceedings of the ACM Web Conference 2023*, 3182–3192.

[139] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. Association for Computational Linguistics.

[140] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2009. *Dataset Shift in Machine Learning*. The MIT Press.

[141] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[142] Nimrod Raifer, Fiana Raiber, Moshe Tennenholtz, and Oren Kurland. 2017. Information retrieval meets game theory: The ranking competition between documents' authors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 465–474.

[143] Thilina Rajapakse and Maarten de Rijke. 2023. Improving the generalizability of the dense passage retriever using generated datasets. In *Proceedings of the 45th European Conference on Information Retrieval (ECIR '23)*. Springer, 94–109.

[144] Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2022. Learning to retrieve passages without supervision. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2687–2700.

[145] Nisarg Raval and Manisha Verma. 2020. One word at a time: Adversarial attacks on retrieval models. arXiv:2008.02197. Retrieved from https://arxiv.org/abs/2008.02197

[146] Revanth Gangi Reddy, Vikas Yadav, Md Arafat Sultan, Martin Franz, Vittorio Castelli, Heng Ji, and Avirup Sil. 2021. Towards robust neural retrieval models with synthetic pre-training. arXiv:2104.07800. Retrieved from https://arxiv.org/abs/2104.07800

[147] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2023. A thorough examination on zero-shot dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP '23*, 15783–15796.

[148] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 856–865.

[149] Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. Springer, 232–241.

[150] Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer. 2023. Questions are all you need to train a dense passage retriever. *Transactions of the Association for Computational Linguistics* 11 (2023), 600–616.

[151] Hinrich Schütze, Christopher D. Manning, and Prabhakar Raghavan. 2008. *Introduction to Information Retrieval*, Vol. 39. Cambridge University Press.

[152] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6138–6148.

[153] Procheta Sen, Debasis Ganguly, Manisha Verma, and Gareth J. F. Jones. 2020. The curious case of IR explainability: Explaining document scores within and across ranking models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2069–2072.

[154] Muhammad Shafique, Mahum Naseer, Theocharis Theocharides, Christos Kyrkou, Onur Mutlu, Lois Orosa, and Jungwook Choi. 2020. Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead. *IEEE Design & Test* 37, 2 (2020), 30–57.

[155] Lujia Shen, Xuhong Zhang, Shouling Ji, Yuwen Pu, Chunpeng Ge, Xing Yang, and Yanghe Feng. 2023. TextDefense: Adversarial text detection based on word importance entropy. arXiv:2302.05892. Retrieved from https://arxiv.org/abs/2302.05892

[156] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 43–50.

[157] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. arXiv:2104.07567. Retrieved from https://arxiv.org/abs/2104.07567

[158] Georgios Sidiropoulos and Evangelos Kanoulas. 2022. Analysing the robustness of dual encoders for dense retrieval against misspellings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2132–2136.

[159] Georgios Sidiropoulos and Evangelos Kanoulas. 2024. Improving the robustness of dense retrievers against typos via multi-positive contrastive learning. In *Proceedings of the European Conference on Information Retrieval*. Springer, 297–305.

[160] 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.

[161] Junshuai Song, Jiangshan Zhang, Jifeng Zhu, Mengyun Tang, and Yong Yang. 2022. TRAttack: Text rewriting attack against text retrieval. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, 191–203.

[162] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? Investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14918–14937.

[163] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL '08) with the Human Language Technology Conference (HLT)*, 719–727.

[164] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR '14)*. Conference Track Proceedings.

[165] Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2023. Typo-robust representation learning for dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1106–1115.

[166] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 35, 21831–21843.

[167] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[168] Ziv Vasilisky, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2023. Content-based relevance estimation in retrieval settings with ranking-incentivized document manipulations. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, 205–214.

[169] Manisha Verma and Debasis Ganguly. 2019. LIRME: Locally interpretable ranking model explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1281–1284.

[170] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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)*, 2153–2162.

[171] Jonas Wallat, Fabian Beringer, Abhijit Anand, and Avishek Anand. 2023. Probing BERT for ranking abilities. In *Proceedings of the European Conference on Information Retrieval*. Springer, 255–273.

[172] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A minimax game for unifying generative and discriminative information retrieval models. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 515–524.

[173] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2345–2360.

[174] Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. 2023. Towards a robust deep neural network against adversarial texts: A survey. *IEEE Transactions on Knowledge and Data Engineering* 35 (2023), 3159–3179.

[175] Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4569–4586.

[176] Xiaosen Wang, Yichen Yang, Yihe Deng, and Kun He. 2021. Adversarial training with fast gradient projection method against synonym substitution based text attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 13997–14005.

[177] Yumeng Wang, Lijun Lyu, and Avishek Anand. 2022. BERT rankers are brittle: A study using adversarial document perturbations. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, 115–120.

[178] Orion Weller, Dawn Lawrie, and Benjamin Van Durme. 2024. NevIR: Negation in neural information retrieval. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2274–2287.

[179] Colin Wilkie and Leif Azzopardi. 2013. Relating retrievability, performance and length. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 937–940.

[180] Colin Wilkie and Leif Azzopardi. 2014. A retrievability analysis: Exploring the relationship between retrieval bias and retrieval performance. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 81–90.

[181] Chen Wu, Ruqing Zhang, Jiafeng Guo, Wei Chen, Yixing Fan, Maarten de Rijke, and Xueqi Cheng. 2022. Certified robustness to word substitution ranking attack for neural ranking models. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 2128–2137.

[182] 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.

[183] Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. Are neural ranking models robust? *ACM Transactions on Information Systems* 41, 2 (2022), 1–36.

[184] Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. 2023. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. arXiv:2401.00396. Retrieved from https://arxiv.org/abs/2401.00396

[185] Ruicheng Xian, Honglei Zhuang, Zhen Qin, Hamed Zamani, Jing Lu, Ji Ma, Kai Hui, Han Zhao, Xuanhui Wang, and Michael Bendersky. 2023. Learning list-level domain-invariant representations for ranking. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 36.

[186] Ji Xin, Chenyan Xiong, Ashwin Srinivasan, Ankita Sharma, Damien Jose, and Paul Bennett. 2022. Zero-shot dense retrieval with momentum adversarial domain invariant representations. In *Findings of the Association for Computational Linguistics: ACL '22*, 4008–4020.

[187] Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. BERM: Training the balanced and extractable representation for matching to improve generalization ability of dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

[188] Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021. Rethinking stealthiness of backdoor attack against NLP models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5543–5557.

[189] Soyoung Yoon, Chaeeun Kim, Hyunji Lee, Joel Jang, and Minjoon Seo. 2023. Continually updating generative retrieval on dynamic corpora. arXiv:2305.18952. Retrieved from https://arxiv.org/abs/2305.18952

[190] Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. COCO-DR: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. arXiv:2210.15212. Retrieved from https://arxiv.org/abs/2210.15212

[191] Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. 2024. AutoDefense: Multi-agent LLM defense against jailbreak attacks. arXiv:2403.04783. Retrieved from https://arxiv.org/abs/2403.04783

[192] Jingtao Zhan, Qingyao Ai, Yiqun Liu, Jiaxin Mao, Xiaohui Xie, Min Zhang, and Shaoping Ma. 2022. Disentangled modeling of domain and relevance for adaptable dense retrieval. arXiv:2208.05753. Retrieved from https://arxiv.org/abs/2208.05753

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

[194] Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Adversarial retriever-ranker for dense text retrieval. In *Proceedings of the International Conference on Learning Representations*.

[195] Jie M. Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* 48, 1 (2020), 1–36.

[196] Peng Zhang, Dawei Song, Jun Wang, and Yuexian Hou. 2013. Bias-variance decomposition of IR evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1021–1024.

[197] Peng Zhang, Dawei Song, Jun Wang, and Yuexian Hou. 2014. Bias–variance analysis in estimating true query model for information retrieval. *Information Processing & Management* 50, 1 (2014), 199–217.

[198] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems* 42, 4 (2024), 1–60.

[199] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*.

[200] Bin Zhou and Jian Pei. 2009. OSD: An online web spam detection system. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Vol. 9.

[201] Bin Zhou, Jian Pei, and Zhaohui Tang. 2008. A spamicity approach to web spam detection. In *Proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 277–288.

[202] Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Guodong Long, Binxing Jiao, and Daxin Jiang. 2023. Towards robust ranker for text retrieval. In *Findings of the Association for Computational Linguistics: ACL '23*, 5387–5401.

[203] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. FreeLB: Enhanced adversarial training for natural language understanding. arXiv:1909.11764. Retrieved from https://arxiv.org/abs/1909.11764

[204] Shengyao Zhuang, Xinyu Mao, and Guido Zuccon. 2022. Robustness of neural rankers to typos: A comparative study. In *Proceedings of the 26th Australasian Document Computing Symposium*, 1–6.

[205] Shengyao Zhuang, Linjun Shou, Jian Pei, Ming Gong, Houxing Ren, Guido Zuccon, and Daxin Jiang. 2023. Typos-aware bottlenecked pre-training for robust dense retrieval. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 212–222.

[206] Shengyao Zhuang and Guido Zuccon. 2021. Dealing with typos for BERT-based passage retrieval and ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2836–2842.
[207] Shengyao Zhuang and Guido Zuccon. 2022. CharacterBERT and self-teaching for improving the robustness of dense retrievers on queries with typos. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1444–1454.

Appendices

A Source Selection

We loosely followed the guidelines by Kitchenham and Charters [86] for the selection of publications used in our survey.

A.1 Sources

We used the sources listed in Table A1 for our survey.

Table A1. Venues, Journals, and Repositories Used for the Survey

Source	Acronym or URL
AAAI Conference on Artificial Intelligence	AAAI
Annual Meeting of the Association for Computational Linguistics	ACL
arXiv	https://arxiv.org
ACM Conference on Computer and Communications Security	CCS
ACM International Conference on Information and Knowledge Management	CIKM
European Conference on Information Retrieval	ECIR
Conference on Empirical Methods in Natural Language Processing	EMNLP
Foundations and Trends in Information Retrieval	FnTIR
International Conference on Learning Representations	ICLR
International Conference on Data Mining	ICDM
International Conference on Machine Learning	ICML
International Conference on the Theory of Information Retrieval	ICTIR
International Joint Conference on Artificial Intelligence	IJCAI
Information Processing and Management	IPM
ACM SIGKDD Conference on Knowledge Discovery and Data Mining	KDD
Annual Conference of the North American Chapter of the Association for Computational Linguistics	NAACL
Conference on Neural Information Processing Systems	NeurIPS
SIGIR Conference on Research and Development in Information Retrieval	SIGIR
Transactions of the Association for Computational Linguistics	TACL
ACM Transactions on Information Retrieval	TOIS
Text Retrieval Conference	TREC
International World Wide Web Conference	WebConf

A.2 Inclusion Criteria

For papers in the sources listed in Table A1 we used the following criteria to include them in our survey:

(IC1) The paper proposes a definition of one or several robustness notions in the context of IR.
(IC2) The paper proposes an approach to improving the robustness of an IR model.
(IC3) The paper proposes a method to evaluate one or several robustness notions in the context of IR.
(IC4) The paper presents one or several benchmarks for assessing the robustness of IR models.
(IC5) The paper presents a study that investigates the foundations of robustness of IR models.

A.3 Exclusion Criteria

For papers in the sources listed in Table A1 we used the following criteria to exclude them from our survey:

(EC1) The paper is not written in English.
(EC2) The paper is not in the date range of January 2012 to July 2024.
(EC3) An extended version of the paper has been published, which subsumes its contents.

B The BestIR Benchmark

BestIR aims to provide a robustness evaluation benchmark for neural IR models. In order to address the comprehensive robustness challenge, we construct benchmarks mainly in terms of seven types of two aspects of robustness, i.e., adversarial robustness and OOD robustness.

Table B1. Statistics of Datasets in BestIR Benchmark

Robustness	Type	Dataset	#Doc	#Q_{train}	#Q_{dev}	#Q_{eval}
Adversarial robustness	Basic datasets	MS MARCO document [127]	3.2M	370K	5,193	5,793
		MS MARCO passage [127]	8.8M	500K	6,980	6,837
		Clueweb09-B [34]	50M	150	–	–
		Natural Questions [88]	21M	60K	8.8K	3.6K
		TriviaQA [80]	21M	60K	8.8K	11.3K
	Expansion of datasets	TREC DL19[a] [39]	–	–	43	–
		TREC DL20[a] [38]	–	–	54	–
		TREC MB14[a] [97]	–	–	50	–
	Off-the-shelf datasets	ASRC [142]	1,279	–	31	–
		Q-MS MARCO [105]	–	–	4,000	–
		Q-Clueweb09 [105]	–	–	292	–
		DARA [28]	164k	50k	3,490	3,489
OOD robustness	Adaptation to a new corpus	BEIR[a] [167]	18 corpora from datasets in BEIR			
	Updates to a corpus	CDI-MS [23]	3.2M	370K	5,193	5,793
		CDI-NQ [23]	8.8M	500K	6,980	6,837
		LL-LoTTE [18]	5.5M	16K	8.5K	8.6K
		LL-MultiCPR [18]	3.0M	136K	15K	15K
	Query variation	DL-Typo [207]	–	–	–	60
		noisy-MS MARCO [20]	–	–	–	5.6K
		rewrite-MS MARCO [20]	–	–	–	5.6K
		noisy-NQ [20]	–	–	–	2K
		noisy-TQA [20]	–	–	–	3K
		noisy-ORCAS [20]	–	–	–	20K
		variations-ANTIQUE [137]	–	–	–	2K
		variations-TREC19 [137]	–	–	–	430
		Zhuang and Zuccon [206]	–	–	–	41K
	Unseen query type	MS MARCO [127]	–	–	–	15K
		L4 [163]	–	–	–	10K

#Doc denotes the number of documents in corpus; #Q_{train} denotes the number of queries available for training; #Q_{dev} denotes the number of queries available for development; #Q_{eval} denotes the number of queries available for evaluation.
[a]Under current implicit assumptions, one may attempt to use BEIR test queries over MSMARCO.

For adversarial robustness, the datasets are usually used both for adversarial defense and adversarial attack tasks. There are three construction methodologies: (i) basic datasets, which are the original IR datasets that are directly performed attacks and defenses on; (ii) expansion of datasets, which are extensions of the original dataset to model unknown queries or new documents; and (iii) off-the-shelf datasets, which are datasets customized for the task of adversarial attack or defense that can be used directly for evaluation.

OOD robustness, datasets are used to evaluate the performance of the model under unseen documents and unseen queries, respectively. For each scenario, there are two types of evaluation perspectives. For unseen documents: (i) adaptation to a new corpus consists mainly of IR datasets from different domains; and (ii) updates to a corpus consist mainly of the same dataset sliced and diced based on factors such as time or randomly. For unseen queries: (i) query variation consists of different variants of the same query intent, such as typos, changes in word order, and changes in the form of expression; and (ii) unseen query type consists mainly of the different types of queries in a dataset.

Table B1 summarizes the statistics of the datasets provided in BestIR. BestIR is publicly available at https://github.com/Davion-Liu/BestIR. There are lots of datasets available within each aspect for robustness challenges and continue growing. We try to balance the assessment of each aspect of robustness to fully evaluate the model's abilities. In the future, we will consider the issue of robustness in a broader sense and introduce datasets into BestIR. Meanwhile, researchers can also focus on observing specific aspects of robustness performance according to this categorization according to their concerns.

Received 25 July 2024; revised 27 April 2025; accepted 26 August 2025