



Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey

Journal:	<i>Computing Surveys</i>
Manuscript ID	CSUR-2025-0546
Paper:	Long Survey Paper
Date Submitted by the Author:	14-May-2025
Complete List of Authors:	Ni, Bo; Vanderbilt University, Liu, Zheyuan; University of Notre Dame, Lei, Yongjia; University of Oregon Wang, Leyao; Vanderbilt University Zhao, Yuying; Vanderbilt University Cheng, Xueqi; Vanderbilt University Zeng, Qingkai ; University of Notre Dame Dong, Xin; Meta Reality Labs Xia, Yinglong; Meta Reality Labs Kenthapadi, Krishnaram; Oracle Health AI Rossi, Ryan; Adobe Research, Derroncourt, Franck; US Tanjim, Mehrab; Adobe Ahmed, Nesreen; Cisco Research Liu, Xiaorui; NCSU Fan, Wenqi; The Hong Kong Polytechnic University blasch, erik; Air Force Research Laboratory, Information Directorate Wang, Yu; University of Oregon Jiang, Meng; University of Notre Dame, Derr, Tyler; Vanderbilt University, CS
Computing Classification Systems:	Natural Language Processing
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
2024_Trustworthy_RAG_survey_csur.zip	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56

Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey

BO NI, Vanderbilt University, USA
ZHEYUAN LIU[†], University of Notre Dame, USA
YONGJIA LEI[†], University of Oregon, USA
LEYAO WANG[†], Vanderbilt University, USA
YUYING ZHAO, Vanderbilt University, USA
XUEQI CHENG, Vanderbilt University, USA
QINGKAI ZENG, University of Notre Dame, USA
LUNA DONG, Meta, USA
YINGLONG XIA, Meta, USA
KRISHNARAM KENTHAPADI, Oracle Health AI, USA
RYAN ROSSI, Adobe Research, USA
FRANCK DERNONCOURT, Adobe Research, USA
MD MEHRAB TANJIM, Adobe Research, USA
NESREEN AHMED, Cisco Research, USA
XIAORUI LIU, North Carolina State University, USA
WENQI FAN, The Hong Kong Polytechnic University, China
ERIK BLASCH, Air Force Research Lab, USA
YU WANG^{*}, University of Oregon, USA
MENG JIANG^{*}, University of Notre Dame, USA
TYLER DERR^{*}, Vanderbilt University, USA

[†]Significant Contribution.
^{*}Corresponding Authors.

Authors' addresses: Bo Ni, bo.ni@vanderbilt.edu, Vanderbilt University, Nashville, USA; Zheyuan Liu[†], zliu29@nd.edu, University of Notre Dame, Notre Dame, USA; Yongjia Lei[†], yongjia@uoregon.edu, University of Oregon, Eugene, USA; Leyao Wang[†], leyao.wang@vanderbilt.edu, Vanderbilt University, Nashville, USA; Yuying Zhao, yuying.zhao@vanderbilt.edu, Vanderbilt University, Nashville, USA; Xueqi Cheng, xueqi.cheng@vanderbilt.edu, Vanderbilt University, Nashville, USA; Qingkai Zeng, qzeng@nd.edu, University of Notre Dame, Notre Dame, USA; Luna Dong, lunadong@meta.com, Meta, Menlo Park, USA; Yinglong Xia, yxia@meta.com, Meta, Menlo Park, USA; Krishnaram Kenthapadi, krishnaram.kenthapadi@oracle.com, Oracle Health AI, Redwood Shores, USA; Ryan Rossi, ryarossi@gmail.com, Adobe Research, San Jose, USA; Franck Dernoncourt, dernonco@adobe.com, Adobe Research, San Jose, USA; Md Mehrab Tanjim, tanjim@adobe.com, Adobe Research, San Jose, USA; Nesreen Ahmed, n.kamel@gmail.com, Cisco Research, San Jose, USA; Xiaorui Liu, xliu96@ncsu.edu, North Carolina State University, Raleigh, USA; Wenqi Fan, wenqi.fan@polyu.edu.hk, The Hong Kong Polytechnic University, Hong Kong, China; Erik Blasch, erik.blasch.1@us.af.mil, Air Force Research Lab, Rome, USA; Yu Wang^{*}, yu.wang.1@vanderbilt.edu, University of Oregon, Eugene, USA; Meng Jiang^{*}, mjiang2@nd.edu, University of Notre Dame, Notre Dame, USA; Tyler Derr^{*}, tyler.derr@vanderbilt.edu, Vanderbilt University, Nashville, USA.

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

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 0004-5411/2025/8-ART111
<https://doi.org/XXXXXXX.XXXXXXX>

111:2

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

Retrieval-Augmented Generation (RAG) enhances AI-generated content by integrating external knowledge, improving relevance, and reducing hallucinations. However, RAG also introduces risks related to reliability, safety, privacy, fairness, explainability, and accountability, which impact trustworthiness. While various methods aim to address these concerns, a unified framework is lacking. This paper bridges that gap by presenting a comprehensive roadmap for trustworthy RAG systems. We provide a structured analysis of key challenges, existing solutions, and future directions across these aspects. Additionally, we highlight downstream applications where trustworthy RAG can make a significant impact, encouraging further research and adoption in real-world AI systems.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; • **Information systems** → **Information retrieval**.

Additional Key Words and Phrases: Retrieval-Augmented Generation, Trustworthy AI, Large Language Models

ACM Reference Format:

Bo Ni, Zheyuan Liu[†], Yongjia Lei[†], Leyao Wang[†], Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, Ryan Rossi, Franck Dernoncourt, Md Mehrab Tanjim, Nesreen Ahmed, Xiaorui Liu, Wenqi Fan, Erik Blasch, Yu Wang^{*}, Meng Jiang^{*}, and Tyler Derr^{*}. 2025. Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey. *J. ACM* 37, 4, Article 111 (August 2025), 35 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address key limitations of Large Language Models (LLMs), such as hallucinations, outdated knowledge, and limited explainability [49, 196]. By incorporating external information into the generation context, RAG improves accuracy, reliability, and keeps models up-to-date with minimal retraining. These benefits have profound implications for real-world applications. For example, RAG has been effectively applied in medical question answering [144, 174, 190], legal document drafting [118, 170], and educational chatbots [154].

Trustworthiness is context-dependent [31, 85, 171], but in Artificial Intelligence, it often refers to characteristics that make a system *worthy of trust*. In 2022, the National Institute of Standards and Technology (NIST) published guidelines for trustworthy AI, defining trustworthiness from several perspectives: Reliability, Privacy, Safety, Fairness, Explainability, and Accountability [151].

Reliability ensures that the system consistently performs as expected and produces accurate results under various conditions. It includes addressing challenges such as uncertainty quantification and generalization, which are critical for enhancing system reliability. For instance, in a legal analysis system, reliability involves balancing uncertainty quantification (e.g., confidence of retrieved legal citations) and generalization (e.g., applying precedents to new cases).

Privacy focuses on safeguarding user data, ensuring control over personal information. Since RAG has been applied to sensitive domains like the medical field, protecting user information is important. For example, when a healthcare assistant retrieves medical records or generates treatment suggestions, the system must prevent data breaches and ensure sensitive patient details embedded in the language model remain secure.

Safety addresses the system's capacity to prevent and mitigate harm, with a focus on defending against adversarial attacks and reducing risks from malicious actors. Current AI systems often interact with high-risk users, such as teenagers, who may be exposed to harmful or inappropriate content. Adversarial attacks and jailbreaking attempts that alter the chatbot's behavior could lead to misinformation, inappropriate responses, or even dangerous suggestions. Thus, building robust safeguards is crucial for ensuring safety and preventing harm in such interactions.

Fairness focuses on minimizing biases introduced during both retrieval and generation stages, as these biases can significantly affect outcomes in high-stakes domains. Recent advancements include

the use of re-ranking methods to mitigate societal biases in retrieval and fine-tuning techniques to balance demographic fairness with system performance. For example, the admissions assistant must ensure fair treatment of applicants by addressing potential biases.

Explainability emphasizes the need for transparent decision-making processes, enabling users to understand how outputs are generated. For example, a university admissions assistant powered by RAG should offer clear explanations of how student profiles are matched with program requirements, providing insights that users can readily understand and verify.

Accountability pertains to AI governance, including policymaking and law enactment, but also extends to technical aspects like tracing the origins and processes behind AI-generated content. For example, a news system tracing its sources improves accountability and reduces misinformation. Content watermarking helps track provenance and provides an audit trail for verification.

Despite their recent success, concerns about the trustworthiness of RAG-based systems have become an increasingly interest subject. First, RAG systems are susceptible to reliability issues since developers must ensure the output is accurately grounded on the retrieved content [49, 88]. Second, the integration of external databases introduces additional leakage channels. It is imperative to ensure that the RAG systems do not expose private information from both the external databases and the training data of the underlying LLM during the generation process. Third, the reliance on an external database introduces a new attack surface, exposing the systems to a range of adversarial threats [40, 176, 189, 200]. As a result, safety improvements are needed to safeguard the systems. RAG systems pose new challenges regarding data privacy [140]. Additionally, RAG could be susceptible to fairness issues [141] from both the retrieval process and the generation process. How the retrieved data is selected and utilized can significantly affect the fairness of the generated content. The implicit bias during the generation could also be affected by the retrieved content due to the increased confidence [63]. Moreover, explainability remains a significant challenge, as RAG systems often lack transparency in how retrieved information influences generated responses. Lastly, with the rise and potential use of LLMs, accountability is a subject for policymakers on the use of RAG systems. Although progress has been made, these challenges significantly restrict the wide adoption of RAG systems in real-world scenarios, especially in high-stakes scenarios such as medication, legal consulting, and education [170, 174, 190]. Thus, it is essential to incorporate the trustworthy perspective while advancing the RAG systems.

Due to the importance of trustworthiness, a plethora of research has been developed to advance the application of RAG in Large Language Models. However, there is no systematic review of this area's current advancements and challenges. To organize the various perspectives, this paper formulates a systematic discussion on the state of trustworthy RAG in Large Language Models. The list of papers discussed is provided in the GitHub repository*.

2 PRELIMINARIES

This section provides the preliminaries of the RAG framework for LLMs. We will introduce the concept of RAG and the common downstream tasks. We acknowledge the wide range of applications of RAG in domains other than LLMs (e.g., Image Generation), but this survey limits the scope to the applications of RAG in LLMs, sometimes referred to as Retrieval Augmented Language Models (RA-LLMs) [39]. As a simplification of the terminology, in the rest of this survey, we use RAG, RA-LLM, and RAG-based systems interchangeably.

*<https://github.com/Arstanley/Awesome-Trustworthy-Retrieval-Augmented-Generation>

111:4

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

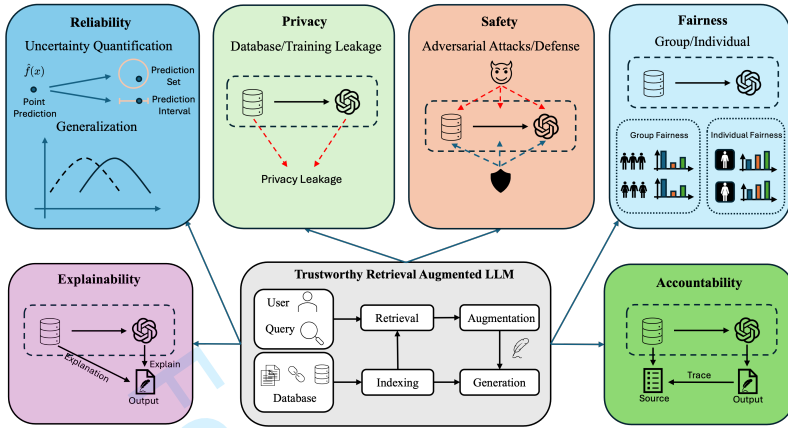


Fig. 1. An overview of the key components and dimensions of Trustworthy Retrieval Augmented Generation (RAG) for Large Language Models (LLMs) that are covered in this survey.

2.1 Retrieval Augmented Generation

As shown in Figure 1, a typical RAG framework comprises three stages: retrieval, knowledge augmentation, and generation. Given a query, the system first retrieves relevant context to support reasoning. Following the classification in [49], retrieval involves two steps: *indexing* and *retrieving*. In *indexing*, heterogeneous sources (e.g., PDFs, HTML, Markdown) are chunked and embedded into vectors, then stored in a vector database. However, symbolic or relational databases may be used instead for some tasks, such as those involving knowledge graphs (e.g., GraphRAG). In *retrieving*, the query is vectorized (or aligned to symbolic structures) and used to fetch relevant chunks or nodes. *Knowledge augmentation* incorporates retrieved content into the generation process, often via prompt injection or fine-tuning. Finally, the *generation* stage produces responses informed by the augmented knowledge.

2.2 Common Tasks in RAG

The RAG paradigm has enabled a wide range of applications. The following section briefly introduces some common tasks—such as question answering and chatbots—along with commonly used datasets associated with these tasks.

Question Answering (QA). One of the primary downstream tasks for RAG-based language models is QA, where the system generates accurate answers to user queries using external retrieved context. Depending on the QA sub-task—such as multi-hop, multiple-choice, or open-domain QA—different evaluation metrics like Hits@ n , F1, Exact Match, or lexical overlap are used to assess answer quality [101, 148, 193]. These metrics offer a practical gauge of performance across varying answer formats and reasoning requirements. Commonly used QA datasets include:

- **MMLU** [59]: A commonly used dataset for multiple-choice question answering (MCQA). It contains MCQA questions from 57 domains, including STEM (science, technology, engineering, and math), humanities, and medicine. In their experiments, existing uncertainty quantification research chooses to use a subset of the dataset for evaluation [78, 185].
- **TriviaQA** [69]: A widely-used large-scale dataset for open-domain question answering, designed to test models on questions from Wikipedia and web search engines. TriviaQA includes question-answer pairs along with evidence documents for context [88, 123, 145].

- **WebQSP** [186]: A popular dataset for multi-hop knowledge base question answering, focusing on the task of answering questions by traversing multiple entities and relations within a knowledge graph. WebQSP provides questions labeled with their corresponding semantic parses, enabling models to learn complex query structures for effective knowledge graph traversal and reasoning [101, 110, 148].

Chatbots. RAG-based language models are widely used in chatbots, enhancing both task-oriented and open-domain dialogue systems by integrating external knowledge for more informed and up-to-date responses [5, 41, 147]. This is especially valuable in domain-specific contexts like healthcare or finance. Evaluation typically centers on two goals: dialogue quality—measured by coherence, utility, and human-likeness [2, 19]—and task effectiveness in real-world settings. Common NLG metrics such as BLEU and ROUGE are also used, though no standardized evaluation exists. Commercial chatbots are often assessed through subtasks (e.g., code generation), while domain-specific systems rely on tailored benchmarks. We outline common datasets below.

- **HellaSwag** [191]: A commonly used dataset for commonsense reasoning and story completion tasks. This dataset contains scenarios requiring contextually appropriate completions, testing model ability to reason beyond surface-level semantics. It has been widely adopted for benchmarking commonsense reasoning capabilities in large language models [37].
- **HumanEval** [21]: A widely used dataset for evaluating code generation capabilities of language models. HumanEval includes programming problems of varying difficulty levels. It is a standard benchmark for assessing the coding performance of generative models [37].
- **MedicationQA** [14]: A popular dataset for question answering in the medical domain, focusing on patient-generated questions about medication. It includes complex medical queries paired with evidence-based answers, making it a crucial benchmark for evaluating the applicability of language models in healthcare and patient communication [37, 82].

Others. Beyond language-based tasks, RAG-based language models can be applied to a diverse range of downstream tasks, including recommendation systems [87], software engineering [62], and AI for scientific discovery [4]. However, these applications are often overlooked in discussions of trustworthy rag frameworks. Recognizing their importance, we highlight the need for further exploration of trustworthiness in these domains and propose to address them in future directions.

Trustworthy Evaluation. Comprehensive trustworthiness assessment requires metrics that capture bias, fairness, and reliability in generated answers. Since AI trustworthiness spans multiple dimensions, evaluation must also consider the given context. We will introduce these metrics in detail throughout the survey in their corresponding sections.

2.3 Motivation

Although trustworthiness has been studied in deep learning and standalone LLMs, RAG-based LLMs introduce unique challenges that require dedicated attention. Their use in high-stakes applications amplifies the risks of errors and biases. Moreover, the multi-stage RAG pipeline—indexing, retrieval, and generation—can compound issues like hallucinations and bias, making them harder to detect and mitigate. Unlike static LLMs, RAG systems rely on external sources, increasing susceptibility to unreliable or biased information. These distinctions limit the applicability of existing trustworthiness frameworks in RAG-based applications. While prior works have addressed individual concerns [35, 140, 178, 203], a unified overview is missing. This survey addresses that gap by systematically reviewing current efforts and outlining future directions for trustworthy RAG-based LLMs.

Table 1. Comparison with Existing Surveys on RAG and Trustworthy LLMs.

Surveys	Pillars of Trustworthiness											
	Reliability		Privacy		Safety		Fairness		Explainability		Accountability	
	Uncertainty	Generalizability	External	Training Data	Jailbreaking	Defense	Retrieval	Generation	Retrieval	Generation	Retrieval	Generation
LLM [94]	✓	✓	✗	✓	✓	✓	✗	✓	✗	✓	✓	✓
[64]	✗	✗	✗	✓	✓	✓	✗	✓	✗	✓	✗	✓
RAG [39]	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
[49]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗
[203]	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

We mark some of the items ✗ for LLM related survey as they do not focus on RAG.

2.3.1 Related Surveys and Differences. As shown in Table 1, existing surveys on RAG [39, 49] and trustworthy LLMs [64, 94] touch on aspects of trustworthiness. However, the RAG surveys [39, 49] offer limited analysis across trustworthiness dimensions, while the LLM trust surveys focus primarily on generation and overlook retrieval-specific issues [64, 94]. Despite their limitations, existing surveys consistently highlight trustworthiness as a key future direction [39, 49], underscoring the growing interest within the community.

A recent survey [203] focuses on trustworthy RAG but emphasizes empirical evaluation over comprehensive review. In comparison, our survey offers a broader literature analysis, categorizing challenges and solutions across the six trustworthiness dimensions: Reliability, Privacy, Safety, Fairness, Explainability, and Accountability. We aim to provide a unifying framework that inspires future research and development.

3 RELIABILITY OF RETRIEVAL AUGMENTED GENERATION

While RAG improves factual consistency and adaptability, it also introduces unique reliability challenges. Unlike standalone generative models, RAG reliability depends not only on the underlying LLM but also the alignment of retrieved information. Ensuring reliability in RAG therefore requires evaluating both the retrieval process and the generation conditioned on the retrieved content.

3.1 Taxonomy of RAG Reliability

At a high level, reliability requires the system to perform as expected under various conditions. Previous work [158] defines reliability from three aspects: the ability to express *uncertainty* in predictions, the capability of *robust generalization* under various conditions, and the extent to which the model can *adapt* to new tasks. Since RAG is inherently *adaptable* because of the retrieved context, we will only consider *uncertainty* and *robust generalization* in our following discussion.

3.2 Uncertainty

Uncertainty is a crucial factor for model reliability. Uncertainty quantification (UQ) helps quantify the confidence in the model's predictions, which is essential in high-stakes scenarios. Consider a medical question answering chatbot where a patient inquires about their condition. If the model can express uncertainty in its responses, it significantly reduces the risk associated with its predictions. The patient can then make more informed judgments based on the confidence level of the information provided. Thus, due to the imperativeness of accurately conveying uncertainty, we need to ensure that robust uncertainty quantification methods are integrated into the system.

Table 2. Taxonomy for Uncertainty Quantification in RAG

Module	Reference	White-Box	Task	Year
Generation	Ye et al. [185]	✗	Benchmarking	2024
	Su et al. [145]	✓	Open Domain Question Answering	2024
	Kumar et al. [78]	✓	Multiple Choice Question Answering	2023
	Quach et al. [123]	✗	Open Domain Question Answering	2023
Retrieval + Generation	Ni et al. [110]	✗	Multi-hop Question Answering	2024
	Li et al. [88]	✗	Open Domain Question Answering	2023

For a RAG system, uncertainty quantification presents two primary challenges: First, during the generation phase, uncertainty stems from the inherent limitations of LLMs. Standard techniques for quantifying uncertainty in LLMs, such as conformal prediction, can be applied here with few adaptations [185]. Second, uncertainty arises during the retrieval phase and its interaction with the LLM, introducing a more complex dynamics. The combination of retrieval and generation processes creates unique challenges for UQ, necessitating advanced methods to address the overall system complexity. The following section outlines ongoing efforts to tackle these challenges, with a summary of the relevant literature presented in Table 2.

3.2.1 Uncertainty Quantification in Generation. The generation phase in a RAG system is critically influenced by the UQ of LLMs. Recent studies have explored various approaches in this area, with a focus on techniques like conformal prediction (CP) – a model-agnostic, distribution-free method that uses a calibration set to estimate prediction confidence [136]. To apply conformal prediction, a *non-conformity score* is first defined to measure the confidence of a given prediction. Using a calibration set, the $1 - \alpha$ quantile of the non-conformity score is then calculated, where α represents the user-defined error rate. Finally, the prediction set is constructed by selecting valid predictions based on the quantile score, ensuring that the set satisfies the $1 - \alpha$ confidence level, assuming the calibration and test sets are exchangeable.

The cornerstone of CP lies in defining the *non-conformity score*. In traditional multi-class classification, a common approach is to use the softmax score of the from the class prediction. Extending the logit-based non-conformity score to LLMs, methods have been further developed. Typically, they assume white-box access to the model, making them unsuitable for commercial LLMs such as ChatGPT. For instance, Kumar et al. [78] applied standard CP to the Llama model [157] by leveraging softmax scores of token logits in multiple-choice tasks. Similarly, Ye et al. [185] extended logit-based approaches to multiple baselines and language models.

To compensate for the lack of application on black models, another promising direction is proposed for sampling-based techniques, where model confidence is estimated by repeatedly prompting the LLM. Quach et al. [123] adapted the learn-then-test risk-control framework [9] for LLMs, approximating the non-conformity score through sampling, which allows uncertainty quantification in black-box models without direct logit access. Su et al. [145] further advanced these methods by introducing non-conformity measures that integrate both coarse-grained and fine-grained notions of uncertainty, leading to smaller, more refined prediction sets.

3.2.2 Uncertainty Quantification in Retrieval and Generation. As shown in Figure 1, a traditional RAG system includes multiple components from retrieval to generation. Due to its complex, multi-component nature, directly applying LLM-based UQ methods will produce less accurate, sub-optimal results [88, 131]. This necessitates the development of specialized techniques tailored to the unique structure and requirements of RAG models.

Recently, researchers proposed a multi-step calibration framework to enhance the retrieval process of RAG [131]. Specifically, this framework uses conformal prediction to quantify retrieval

111:8

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

uncertainty, ensuring trustworthiness in RAG systems. The framework involves constructing a calibration set of questions answerable from the knowledge base and comparing their embeddings against document embeddings to identify the most relevant chunks containing the answers. By analyzing similarity scores and determining a cutoff threshold based on a user-specified error rate (α), the system retrieves all chunks exceeding this threshold during inference. This multi-step calibration ensures the true answer is captured in the context with a $(1 - \alpha)$ confidence level.

Moreover, TRAQ [88] expanded the conformal prediction framework to include a Bayesian optimization module that minimizes the prediction set during the multi-step calibration. Because of the complexity of RAG, the constructed prediction set will be very large after aggregating the error rates of multiple components. By leveraging Bayesian optimization, the framework efficiently searches for the optimal parameters that reduce the size of the prediction set while maintaining the desired confidence level. TRAQ ensures that the retrieval process remains both accurate and computationally feasible, enhancing the overall reliability and performance of RAG systems.

Besides uncertainty quantification in the process of document retrieval, UAG [110] attempted to address the gap in uncertainty quantification of knowledge graph reasoning. Unlike vector databases, knowledge graphs encode knowledge as triplets and include structural information. One representative task of knowledge graph reasoning is multi-hop question answering, where the system must infer answers by traversing multiple edges in the graph to connect the initial query node with the answer node. The UAG framework involves combining information from several related entities and relationships within the graph, further complicating the process of uncertainty quantification. UAG proposed to leverage a general risk control framework to find the optimal parameter for each stage of calibration, ensuring reliable uncertainty estimates while maintaining a reasonable prediction set size.

3.3 Uncertainty Evaluation

Metrics. Traditionally, uncertainty quantification is evaluated from two key perspectives: *coverage* and *efficiency* [58]. Recall that the goal of uncertainty quantification is to ensure that the returned answer set satisfies a user-defined error tolerance of $1 - \alpha$. Thus, the *coverage rate* measures how effectively the model meets this requirement.

Given a returned set of answers, \mathcal{A}_{ret} , and the correct answer set, $\mathcal{A}_{\text{true}}$, the coverage rate, C , is calculated as the proportion of instances where the correct answer is included in the returned set. Formally, $C = \frac{N_{\text{correct}}}{N_{\text{total}}}$, where N_{correct} represents the number of times the correct answer $\mathcal{A}_{\text{true}}$ is contained in the returned set \mathcal{A}_{ret} , $\mathcal{A}_{\text{true}} \subseteq \mathcal{A}_{\text{ret}}$, and N_{total} is the total number of instances.

A model is considered reliable when $C \geq 1 - \alpha$. However, simply exceeding the threshold does not necessarily indicate optimal performance. Overestimating the returned set size while still satisfying the desired error rate implies inefficiency, as a smaller set could suffice for the same error rate.

As a result, alongside coverage, *efficiency*, denoted as E , is another critical metric, often evaluated by the size of the returned answer set (i.e. the number of returned answers per question), i.e., $E = |\mathcal{A}_{\text{ret}}|$. Efficiency reflects the utility of the model's output, as larger sets may contain more irrelevant information, reducing their usefulness to the user. Thus, an efficient uncertainty quantification process minimizes E while maintaining the desired coverage rate, C .

3.4 Robust Generalization

Previous work [158] defines *robustness* as the ability to make accurate estimates or forecasts about unseen events caused by out-of-distribution data, covariate shift, domain change, concept change, or population shift, etc. In the context of RAG, the most significant challenge is the shift in the distribution of the database. Realistically, the database will always be evolving, introducing new

knowledge into the system. Without dedicated robustness measures, this can cause the model to underperform in various situations. Consequently, it is essential to develop approaches that allow the model to continually learn from new data and adjust its retrieval and generation processes accordingly such as in concept drifts. Specifically, we will consider two aspects of robustness for RAG: resilience against irrelevant context and resilience against corrupted contexts. It is worth mentioning that there is another type of context that we define as *adversarially constructed corrupted context*. Sometimes they are closely related to *corrupted context*, but due to their adversarial nature, we will consider them related to adversarial robustness and discussed in Section 5 for Safety. This section will focus on the corrupted context that occurs *organically* over time.

3.4.1 Irrelevant Context. Fang et al. [40] considers the noise robustness of RAG with adaptive adversarial training. The paper explores three types of retrieval noises: (i) contexts that appear to be related to the query but do not contain the correct answer, (ii) contexts that are entirely unrelated to the query, and (iii) contexts that are thematically related to the query but include incorrect information. With the conclusion that type (i) and (iii) noise are the most misleading to the language models, the authors developed Retrieval-augmented Adaptive Adversarial Training (RAAT) to regulate the retrieval of noisy text. To improve the robustness under the noisy data, RAAT generates adversarial samples (noises) by considering the model's sensitivity to various types of noises and shows significant robustness improvement. The study further demonstrates that RAAT can be integrated seamlessly with existing RAG systems, enhancing their performance without substantial computational overhead.

In addition, Yoran et al. [189] further explores the negative impact of the retrieval of irrelevant context on the model performance. They argue that the negative impact of the irrelevant context is a result of the lack of training data with the retrieved passages. As a result, the brittleness to noisy passages is expected during inference. To address this observation, the author propose to finetune the language models on noisy contexts, which allows the model to learn to differentiate between useful and distracting information and minimize the negative effect of irrelevant context. Their experiment result on five open domain datasets has shown significant improvements of robustness against irrelevant context for both single-hop and multi-hop retrieval based question answering.

3.4.2 Corrupted Context. Recently, Xu et al. [176] proposed a theoretical framework to explore the benefits and detriments of the RAG, in the situation where there's a discrepancy between the retrieved knowledge and the LLM knowledge. Specifically, they observed that the similarity between the RAG representation and the retrieved representation is bounded by the benefits and detriments, and the similarity is positively correlated with the value of benefits minus detriments. These results suggest that the similarities can be used as a proxy for the benefits and detriments of the RAG. Building upon the theoretical results, they further proposed an interactive inference framework X-RAG that leverages the benefit of both retrieved knowledge and LLM knowledge.

3.5 Robustness Evaluation

Metrics. The evaluation of the model's robustness focuses on assessing its performance when noise is present in the data. Thus, the setup of the noisy data, which will be detailed in the dataset section, plays a key role in this evaluation. Existing metrics outlined in 2.2 will be applied to assess the model's performance. It is important to note there are different reporting styles for these metrics in the context of model robustness. Some authors present standard tables comparing the proposed model's performance against baselines [40, 176], while others report only the performance delta between the proposed fine-tuned model and corresponding baseline for better visualization [189].

Table 3. Taxonomy for RAG Privacy

	REFERENCE	TRAINING	TASKS	LEAKAGE	YEAR
<i>Attack</i>	Zeng et al. [140]	✓	Document Extraction & Training Data	Internal & External	2024
	Liu et al. [92]	✓	Membership Inference Attack	External	2024
	Cohen et al. [26]	✗	MIA & Document Extraction	External	2024
	Jiang et al. [67]	✗	Document Extraction	Internal & External	2024
	Peng et al. [115]	✓	Document Extraction	External	2024
<i>Defense</i>	Zeng et al. [192]	✓	External Database	External	2024
	Zeng et al. [140]	✓	Document Extraction	External	2024

Datasets. Currently, there is no widely used benchmark for RAG robustness. To evaluate robustness, existing works create customized datasets that incorporate generated noise. Typically, a common QA benchmark (e.g., TriviaQA) is used, and during the retrieval process, noises are injected into the retrieved content. Depending on the problem setting (*irrelevant context* vs. *corrupted context*), the noise is either randomly selected or filtered using heuristic techniques [176, 189]. These datasets attempt to replicate the type of challenges encountered in realistic environments where the retrieved information may not perfectly align with the query.

Recently, Fang et al. [40] proposed a benchmark for noise-robust RAG. For each QA instance, the proposed dataset includes three types of augmented retrieval noise: relevant retrieval noise, irrelevant retrieval noise, and counterfactual retrieval noise where the answer entity is intentionally incorrect. A golden retrieval document is also provided for each query. We recognize this as one of the first publicly available datasets for RAG robustness evaluation, and future works could benefit from using this for benchmarking.

3.6 Future Directions of RAG Reliability

Improving reliability remains a central challenge in trustworthy RAG systems. While we currently treat uncertainty and robustness separately, future work should integrate these aspects to better capture their intersections. Uncertainty quantification can enhance trustworthiness under noisy or imprecise contexts, while robust generalization methods can, in turn, reduce uncertainty. A unified framework combining both could help RAG systems adapt to complex, real-world inputs where irrelevant and corrupted contexts often coexist. Such integration would enable models to both resist noise and adjust confidence dynamically based on context quality.

Evaluation benchmarks should be improved. While datasets like RAG-Bench [40] offer a starting point, their rule-based filtering limits practicality. Future benchmarks should reflect the diverse, noisy nature of real-world data, allowing models to be tested under scenarios where uncertainty and robustness are tightly coupled in the real-world.

Finally, drawing insights from areas like dynamic knowledge graphs and active learning can improve adaptability. As external knowledge changes, RAG systems must handle inconsistencies, generalize effectively, and quantify uncertainty in shifting contexts. These advances could enhance RAG reliability in more practical environments.

4 PRIVACY OF RETRIEVAL AUGMENTED GENERATION

Although privacy risks in LLMs are well-studied, integrating external data in RAG introduces additional challenges. This section introduces the threat model for privacy leaks in RAG, discusses current mitigation efforts, and explores future directions for enhancing RAG trustworthiness.

4.1 Taxonomy of RAG Privacy

Table 3 summarizes existing efforts addressing privacy in RAG systems. We briefly describe the relevant taxonomy below.

4.1.1 Training. Training indicates whether attacks or defenses require *prior* data training. Approaches requiring training usually assume white-box access, allowing fine-tuning of RAG components. In contrast, methods without training often rely on prompt-based or zero-shot techniques.

4.1.2 Tasks. Current RAG privacy research involves three tasks: *Document Extraction*, *Membership Inference*, and *Training Data Extraction*. *Document Extraction* aims to retrieve confidential data (e.g., Personally Identifiable Information) from external databases. *Membership Inference* determines if specific passages exist in databases, potentially exposing sensitive data. Lastly, *Training Data Extraction* explores leakage of sensitive information from the LLM’s internal training data.

4.1.3 Leakage. Privacy leakage can originate from two sources: external retrieval databases and internal training data. External leakage involves exposing sensitive information from retrieved knowledge sources, while internal leakage occurs when LLMs inadvertently reproduce confidential training data. The following discussion is structured around these two aspects.

4.2 Data Leakage From the External Retrieval Database

The goal of the attacker is to exploit privacy vulnerabilities within the retrieval dataset. Zeng et al.[140] introduced a composite structured prompt, formulated as $q = \text{information} + \text{command}$, which leverages the context retriever’s propensity for similarity-based matching. However, a significant limitation of this approach is its reliance on fixed queries, which cannot dynamically adapt to varying contexts. To address this limitation, Jiang et al.[67] proposed a learning-based method. Their framework begins with an initial adversarial query and iteratively refines it based on the model’s responses, progressively generating queries to extract as many documents as possible from the retrieval database.

When considering white-box access to the model, Peng et al. [115] focused on data extraction through backdoor attacks. Their method trains a model to associate specific triggers with desired outputs. Beyond directly extracting documents, their approach also explores generating stealthy outputs using a language model to paraphrase the retrieved content, thereby increasing the difficulty of detecting the attack. Furthermore, Cohen et al. [26] demonstrated that these attacks can escalate beyond isolated cases. By crafting an *adversarial self-replicating prompt*, attackers can initiate a chain reaction that propagates through the entire RAG system.

Although distinct from direct extraction methods and based on a different threat model, *membership inference attacks* have also proven effective for data extraction. These attacks allow malicious users to infer whether specific content is present in the retrieval database. Liu et al. [92] introduced a mask-based attack that obscures portions of documents, compelling the language model to predict the masked words.

4.3 Data Leakage From the LLM Training Data

The goal of the attacker is to extract data from the LLM’s training and fine-tuning data that are encoded in the model parameters. In their paper, Zeng et al. [140] compared the effect of RAG in preventing data leakage from the LLM training data. The result shows that incorporating retrieved passages greatly reduces LLM’s propensity to reproduce content memorized during its training/fine-tuning process. To isolate the effect of retrieval data integration, the author also attached 50 tokens of random noise injection as prefix. Although the random noise could also mitigate the data leakage, it is far less effective than integrating the retrieved content.

4.4 Defense on Privacy Attacks

Although still a relatively under-explored area, some works have proposed defenses to mitigate privacy vulnerabilities in RAG systems. In their foundational work, Zeng et al. [140] observed that using a separate model to summarize the retrieved documents effectively reduces privacy leakage by abstracting sensitive information into generalized content. Additionally, they proposed implementing a distance threshold in the retrieval database, ensuring that only documents with certain relevance requirements are returned. However, this approach introduces a trade-off between system performance and privacy protection, as stricter thresholds can limit retrieval accuracy.

Building on these mitigation strategies, the authors further suggested the use of purely synthetic data as a way to entirely avoid potential leakage of real data [192]. This method involves identifying importing attributes of the data through few-shot samples, extracting key information associated with these attributes, and generating synthetic data that mirrors the original data without exposing sensitive information. This approach has shown promise in effectively mitigating privacy leakage while maintaining the performance of the RAG system.

4.5 Privacy Evaluation

Metrics. Metrics for evaluating privacy attacks focus on quantifying the extent of information leakage and the effectiveness of extraction methods. Commonly used metrics include the total volume of retrieved context, the number of prompts that successfully yield substantial overlaps (e.g., at least 20 matching tokens) with the dataset, and the number of unique excerpts extracted. For targeted attacks, the evaluation centers on the precision of the extracted information, assessing how accurately specific targets are retrieved. In the case of untargeted attacks, metrics often rely on content similarity measures, such as ROUGE-L scores, to determine the degree of alignment between the retrieved content and the original dataset [140, 192].

Datasets. While there are no established datasets or baselines specifically for evaluating privacy in RAG systems, current work has leveraged existing datasets to provide initial insights. The Enron Email dataset [76] contains sensitive employee communications, including names, contact details, and internal messages. The HealthcareMagic-101 dataset [192], on the other hand, includes doctor-patient dialogues with personally identifiable and private health information. These datasets offer realistic, sensitive content that serves as a valuable starting point for privacy evaluations.

4.6 Future Directions of RAG Privacy

While Zeng et al. [140] highlight the dual nature of privacy risks in RAG systems, much remains unknown about how these risks play out across domains. Future work should focus on domain-specific privacy techniques for high-stakes areas like healthcare, finance, and law, where breaches can be particularly damaging. Additionally, incorporating advanced cryptographic tools like homomorphic encryption into RAG pipelines could enhance data protection. Likewise, adapting differential privacy to RAG can help strike a better balance between privacy and utility.

Finally, as the field of RAG continues to evolve, it will be crucial to establish comprehensive benchmarks and evaluation metrics for privacy in these systems. These benchmarks should account for the diverse range of privacy threats, including both direct data leakage and more subtle inferential attacks, to ensure that the proposed solutions are rigorously tested and validated across a wide spectrum of scenarios.

5 SAFETY OF RETRIEVAL AUGMENTED GENERATION

Recent studies have shown that LLMs are vulnerable to adversarial attacks [64, 96, 137, 184], such as prompt engineering and input perturbation. RAG systems, which combine LLMs with external

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56

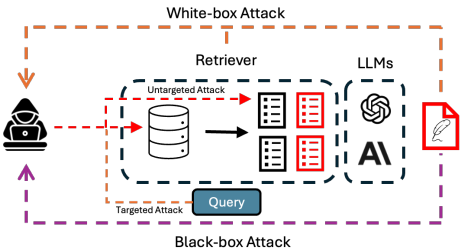


Fig. 2. An illustration of the RAG Safety Threat Model

databases, introduce additional safety risks. For example, attackers may inject adversarial content into retrieved data, bypassing safety alignment and producing malicious outputs [34]. As RAG adoption expands, especially in high-stakes contexts like education, ensuring safety becomes crucial to prevent harmful outputs. This section briefly summarizes adversarial attacks on RAG and highlights promising research directions to enhance RAG safety.

5.1 Taxonomy of RAG Safety

Table 4 summarizes existing RAG methods based on an adversarial taxonomy, briefly defined in this section. Although numerous attacks on LLMs (e.g., backdoor [99, 179], jailbreaking [169, 204], prompt injection [52, 95, 180]) primarily target the underlying LLM, the effect of retrieved contexts on attack surfaces remains unclear. Preliminary studies [140] indicate retrieval can mitigate simple attacks, but complex scenarios involving combined backdoor and retrieval attacks require further research. Due to space constraints, we omit detailed discussion of LLM-specific attacks (see [64, 94]) and instead focus on the retriever’s robustness and its interaction with the generator. Figure 2 provides an overview of the RAG safety taxonomy.

5.1.1 Threat Model. We define the threat model based on three main RAG components: the external database, the output generator (the underlying language model), and the retriever. A realistic assumption is that attackers have write-only access to the external database, reflecting scenarios where users can upload documents to a database but cannot access all of the content. Attackers typically have no detailed knowledge about the underlying language model, as commercial LLMs are usually proprietary black-box systems.

For the retriever, we differentiate between two threat scenarios. In the white-box setting, attackers have full access to the retriever’s parameters, allowing sophisticated adversarial examples crafted by exploiting known weaknesses (e.g., misleading yet highly-ranked documents [205]). Conversely, in the black-box setting, attackers lack direct access and must indirectly infer the retriever’s behavior by manipulating external data, relying solely on observing retrieval outputs [34].

5.1.2 Attacker’s Goal. In traditional machine learning, attacks are typically categorized as targeted or untargeted. For generative models like RAG, we identify two corresponding categories: targeted attacks and jailbreak attacks. Targeted attacks aim to subtly manipulate responses to specific queries or topics, evading detection and potentially skewing sensitive information (e.g., influencing perceptions on political or social issues). On the other hand, jailbreak attacks broadly seek to bypass built-in safety constraints, provoking models to generate unsafe or inappropriate content.

5.2 Methods of RAG Safety

Targeted Attacks. Zou et al. [205] introduced PoisonedRAG, an attack exploiting the retrieval component by injecting carefully designed passages into the external database. The goal is to manipulate RAG systems to produce attacker-specified answers to particular questions. PoisonedRAG

Table 4. Taxonomy for RAG Safety

	REFERENCE	WHITE-BOX	BLACK-BOX	YEAR
<i>Targeted</i>	Zou et al. [205]	✓	✓	2024
	Xue et al. [178]	✓	✗	2024
	Long et al. [97]	✓	✗	2024
	Zhong et al. [202]	✓	✗	2023
<i>Jailbreak</i>	Wang et al. [168]	✗	✓	2024
	Deng et al. [34]	✗	✓	2024

considers both black-box and white-box scenarios. In the black-box setting, where the attacker lacks access to model parameters, it uses a heuristic: passages closely matching the query are more likely retrieved. In the white-box setting, with model parameter access, the passages are optimized to maximize similarity between the encoded query and passage: $P = \operatorname{argmax}_P \operatorname{Sim}(f(Q), f(P'))$, where Q is the user query, P is the adversarial passage, f is the encoder, and Sim measures similarity.

However, PoisonedRAG targets specific queries, neglecting broader group-based attacks on semantically related query categories (e.g., political or social groups). To address this, Xue et al.[178] proposed the BadRAG framework, extending targeted attacks to semantic groups. For instance, given a topic such as *Republican*, BadRAG identifies related terms (*Governor*, *Red States*, *Pro-Life*) as triggers. It employs contrastive learning to craft adversarial passages, maximizing similarity with triggered queries and minimizing it with normal queries. BadRAG also enables other adversarial behaviors, like Denial-of-Service (DoS) attacks[178], further exploring the vulnerabilities.

Dense retrieval has been extensively studied within the Information Retrieval community [201]. Many attacks on dense retrievers, such as adversarial manipulation, are also applicable to passage retrieval tasks. Recently, Long et al.[97] introduced a backdoor attack framework that exploits grammatical errors as triggers to spread misinformation. By employing contrastive learning to fine-tune the retriever, the model can retrieve adversarial passages specified by an attacker when it detects these grammatical anomalies. Additionally, Zhong et al.[202] demonstrated that adversarial passages trained on one domain can effectively transfer to out-of-domain queries, broadening the scope and potential impact of such attacks.

Despite these advances, their impact on downstream generation remains unclear, as most were developed without considering generation tasks. For instance, certain RAG methods equipped with safety guardrails could potentially diminish the impact of adversarially retrieved passages. As we will discuss in the Future Directions section, bridging this gap presents a key research opportunity in Trustworthy RAG.

Jailbreak Attacks. When specific attack targets are absent, the threat model shifts toward jailbreak attacks. Wang et al.[168] examine jailbreaking in the context of LangChain, a popular RAG framework. They analyzed jailbreak vulnerabilities in major Chinese Large Language Models and introduced the Poisoned-LangChain (PLC) method. By embedding jailbreak prompts into the retrieval database, PLC achieved jailbreak success across three scenarios, maintaining a consistent success rate exceeding 80%. More recently, Deng et al.[34] introduce Pandora, which extends jailbreaking attacks to English-based LLMs and more generalized RAG frameworks. Pandora enhances the malicious prompts by categorizing them into distinct topics and storing them in PDF format. This approach ensures that only titles and abstracts are retrieved, by which circumventing potential defense mechanisms that might detect the malicious content.

5.3 Safety Evaluation

Metrics. For targeted attacks, researchers typically evaluate results from two perspectives. First, they measure the exclusivity of the trigger query’s effectiveness. To avoid detection, it’s essential that the same adversarial effects do not occur for non-triggered queries. To quantify this, retriever-based methods like BadRAG [178] assess the proportion of adversarial passage retrievals for clean queries compared to triggered queries. Specifically, they report the percentage of queries that retrieve at least one adversarial passage in the top- k results (where $k = 1, 10, 50$). Second, they measure the effectiveness using the Attack Success Rate (ASR). Notably, in generative tasks, ASR requires nuance due to variations in language expression. For example, responses like “*Sam Altman*” and “*The CEO of OpenAI is Sam Altman*” both correctly answer “*Who is the CEO of OpenAI?*” Thus, researchers often employ *substring matching* rather than *exact matching* for this evaluation [205].

Jailbreak attacks are also evaluated using ASR. However, lacking a targeted question, the criteria for successful jailbreaks need to be carefully defined. Deng et al.[34] manually label a generation as a successful attack based on the *relevance* and *quality* of the generated content. Similarly, Yang et al.[168] also manually count successful attacks to calculate ASR. We identify this as a methodological gap and will further discuss it in the future directions section.

Datasets. Currently, there are no widely accepted standardized datasets for evaluating robustness in RAG systems. For targeted attacks, researchers often follow a structured paradigm: first, identifying the downstream task. The current evaluations primarily focus on question answering, leveraging widely used benchmark datasets such as Natural Questions (NQ)[79], MS MARCO[12], and SQuAD [124]. Researchers then select questions based on the specific characteristics of the targeted attacks. For instance, in BadRAG, Xue et al. [178] chose *Republican* and *Democrats* as targets for sentiment steering. Evaluation involves comparing results between clean (untargeted) queries and targeted queries to measure the impact of the attack.

For jailbreak attacks, evaluation methods vary due to the challenges of assessing open-ended text generation. Current approaches rely on manually curated adversarial questions categorized by theme. For example, Pandora [34] groups questions into *Adult*, *Harmful*, *Privacy*, and *Illegal*, while Poisoned-Langchain [168] uses *Dangerous Behaviors*, *Misuse of Chemicals*, and *Illegal Discrimination*. Despite these efforts, RAG literature lacks standardized datasets and unified evaluation frameworks, underscoring the need for comprehensive benchmarks and methodologies.

5.4 Future Directions of RAG Safety

Defenses against adversarial attacks in RAG remain underdeveloped. Xue et al. [178] explore token masking to learn links between triggers and adversarial passages, but broader research is still lacking. As discussed in Sectio 3.2, integrating adversarial examples into training could improve generalization, yet targeted strategies tailored to RAG-specific challenges are still needed.

Moreover, most current RAG research focuses on text-based retrieval, but emerging applications now incorporate structured sources like knowledge graphs [110]. These bring new challenges for attack and defense due to their unique semantic and relational structures. Future work should develop modality-specific strategies to enhance RAG robustness across diverse data sources.

Finally, as highlighted in preceding sections, the field currently lacks standardized evaluation protocols. This absence hinders the ability to conduct fair comparisons and benchmark the effectiveness of different approaches. We advocate for the establishment of comprehensive evaluation frameworks and benchmarks that consider diverse trustworthy related metrics.

111:16

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

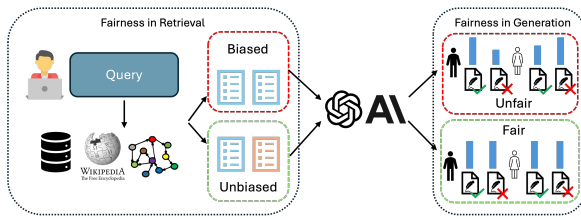


Fig. 3. An overview of the fairness challenge in RAG. The biased retriever includes content related to the blue sensitive attributes, and the biased language model generates unfair content based on protected attributes.

6 FAIRNESS OF RETRIEVAL AUGMENTED GENERATION

As generative models such as LLMs and image-generative models become increasingly integrated into real-world applications, it is critical to ensure fair outputs. RAG systems combine generative models with external knowledge retrieval, providing improvements in accuracy and relevance by incorporating up-to-date information. However, the external data retrieved by these models may contain societal biases due to biased pre-trained knowledge [15, 163], which leads to biased outputs. This also introduces the risk of amplifying disparities in gender, race, and other demographic attributes, particularly when the retrieved data is drawn from biased or unregulated sources. Figure 3 demonstrates how biased documents and retrieval can lead to biases in generation.

6.1 Taxonomy of RAG Fairness

In this section, we investigate various approaches that aimed at promoting fairness in RAG models. As shown in Table 5, Ensuring fairness in RAG systems requires addressing biases at two stages: the retrieval of external data and the generation of outputs. These stages introduce unique fairness challenges, from biased data sources to unfair generative behavior, which need to be mitigated to create equitable and unbiased systems.

6.2 Fairness in Retrieval

One major challenge for fairness in RAG systems lies in the retrieval phase, where external knowledge or data used to enhance generation is sourced. Fairness issues during this stage can arise from various factors, including the retrieval model, the retrieval process, and the re-ranking mechanism. To address these challenges, several frameworks have been proposed to ensure that the retrieval system itself is fair. Reskabsaz et al. [126] introduces a framework for measuring bias that quantifies gender-related biases in ranking lists and assesses the impact of both BM25 and neural retrieval models. Furthermore, Reskabsaz et al. [125] investigates how re-ranking methods can mitigate biases present in initial retrieval results. Then, Wang et al. [167] recognizes a gap between fairness and ranking performance when using LLMs for re-ranking and proposes a method with LoRA. To ensure demographic diversity, FairRAG [141] incorporates external data sources that cover a broad range of age, gender, and skin tone categories. This approach uses post-hoc sampling techniques to debias the retrieval process, preventing disproportionate representation of specific demographic groups in the retrieved data. Beyond addressing bias in the data itself, methods such as BadRAG [178] have shown how maliciously inserted or poisoned data in the retrieval corpus can lead to biased and unfair outputs. Kong et al. [77] proposes Post-hoc Bias Mitigation (PBM), which balances retrieved image sets to ensure more equitable representation across gender/race.

Table 5. Taxonomy for RAG Fairness

MODULE	REFERENCE	BIAS MITIGATION	FOCUS	YEAR
Retrieval	Wang et al. [167]	LoRA Fine-tuning	Ranking Fairness vs. Performance	2024
	Shrestha et al. [141]	Diverse Sampling	Demographic Diversity	2024
	Rekabsaz et al. [125]	Re-ranking Methods	Societal Bias Mitigation	2021
	Rekabsaz et al. [126]	Post-hoc Re-ranking	Gender Bias in Retrieval	2020
Generation	Wu et al. [172]	Empirical Evaluation	Cross-task Fairness	2024
	Wang et al. [164]	Output Conditioning	GPT-3.5/4 Bias Detection	2023
	Liang et al. [89]	Representation Adjustment	Fair Question Answering	2022
	Parrish et al. [114]	Benchmark Evaluation	Stereotype Analysis	2021
Retrieval + Generation	Kong et al. [77]	Post-hoc Bias Mitigation (PBM)	Gender and Race Fairness	2024
	Kim et al. [72]	Fair Retrieval + Generation	Fairness-Quality Trade-off	2024
	Shrestha et al. [141]	Cross-Modal Guidance	Demographic Balancing	2024

6.3 Fairness in Generation

Once the data is retrieved, the next challenge lies in ensuring that the generative process itself is fair. Even with fair retrieval, generative models may introduce biases based on how the retrieved data is integrated. To promote fairness in generation, Liang et al. [89] assesses the accuracy of question-answering systems while accounting for fairness through measures of toxicity and representation bias. Similarly, Wang et al. [164] identifies the demographic imbalances in models like GPT-3.5 and GPT-4 under both zero-shot and few-shot question-answering settings. FairRAG [141] employs conditioning techniques where generative models are guided by references that are demographically diverse. By incorporating external images or data from a wide range of demographic groups, these models produce more balanced and representative outputs. Parrish et al. [114] introduces the BBQ benchmark to evaluate biases in LLM-generated responses by examining the reliance on stereotypes and anti-stereotypes in both ambiguous and disambiguated contexts. Next, to fully explore fairness throughout all stages and components of RAG pipelines, Wu et al. [172] conducts an empirically evaluation of fairness across various RAG methods. Similarly, Kim et al. [72] evaluates RAG systems with a fairness-aware retriever across seven different tasks and identifies the overall trend of fairness-quality trade-off, considering both retrieval and generation performance.

6.4 Fairness Evaluation

Metrics. First, to assess the accuracy of generated answers, it is common to use Exact Match (EM) [124] and ROUGE-1 scores [90]. For fairness evaluation, the focus is on metrics such as Group Disparity (GD) [46] and Equalized Odds (EO) [56]. Group Disparity measures the performance difference between protected and non-protected groups by calculating the ratio of exact matches within each group to the total number of exact matches across all groups. In contrast, EO evaluates whether the likelihood of correct answers (true positives) and incorrect answers (false negatives) is similar across different demographic groups, ensuring that no group is disproportionately advantaged or disadvantaged. Works like BadRAG [178] identify vulnerabilities and attacks on retrieval components (RAG database) and their indirect effects on generative parts (LLMs). They evaluate metrics such as retrieval success rate and rejection rate to assess the system’s robustness and ensure that biases or attacks in retrieval do not impact generative outputs.

Datasets. A popular dataset is the TREC Fair Ranking Track [27, 36], which includes subsets such as gender and location. The track aims to provide a platform for participants to develop and evaluate novel retrieval algorithms that ensure fair exposure to a mix of demographics or attributes, such as ethnicity, represented by relevant documents in response to a search query. The BBQ dataset [114] includes samples with contexts that are either ambiguous or unambiguous.

111:18

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

Ambiguous contexts test model behavior with insufficient evidence by providing only a general setting, while disambiguated contexts offer enough details to identify the correct individual for negative/non-negative questions. The LaMP benchmarks [132] include various prediction tasks like classification, regression, and generation, and are ideal for scenarios where multiple items can be relevant, unlike typical QA tasks. With clear item providers and consumers, LaMP aligns with the goal of ensuring fairness for item providers and evaluates language models' personalization capability through retrieval-augmentation of user interaction histories.

6.5 Future Directions of RAG Fairness

While frameworks like FairRAG [141] and PBM [77] have reduced bias in RAG systems, they often neglect personalized fairness. Tailoring fairness constraints to specific domains—such as emphasizing equity in healthcare trials or reducing bias in hiring decisions—is essential for real-world impact. Future research should explore adaptive methods that adjust fairness criteria based on contextual and personal demands. Another key challenge is managing the trade-offs between fairness, accuracy, and relevance. Although recent work [89, 164] focuses on fairness during retrieval and generation, few works effectively balance all objectives. Multi-objective optimization could offer a more practical path forward, enabling more balanced and domain-aware fairness solutions.

Finally, multimodal RAG systems that handle both text and image data raise new fairness concerns. Biases may emerge differently across modalities, and current research [72, 172] only begins to address this. Future work should ensure consistent fairness across modalities to build more trustworthy multimodal systems.

7 EXPLAINABILITY OF RETRIEVAL AUGMENTED GENERATION

Explainability is a crucial aspect of trustworthiness, explaining model behavior and providing valuable insights for decision making [127, 129]. Its importance has grown with the rise of black-box LLMs [89, 146]. While prior surveys [203] address transparency in RAG, our focus is distinct: While transparency is more general and seeks to understand the algorithms and underlying rationales, our explainability (of the output) specifically aims to elucidate why a particular input (transparency) leads to a given output through a specific model (interpretability).

As shown in Figure 4, RAG systems pose unique explainability challenges due to the complexity of their multi-stage nature. Beyond explaining the generation step, it's critical to understand the retrieval decisions—e.g., which input terms trigger the selection of specific contexts. In systems with post-retrieval processing [51, 71, 181], further explanation is needed for how inputs influence reranking or filtering. This section reviews methods for explaining both the retrieval and generation components of RAG.

7.1 Taxonomy of RAG Explainability

7.1.1 Explainability in Retrieval. As illustrated in Table 6, we situate our discussion of explainability in RAG in the distinction between retrieval, generation, and dual enhancement. To the best of our knowledge, no dedicated research efforts have been made to explain retrieval within the context of RAG. However, several studies have indeed explored explainability in the general information retrieval, particularly in recommender systems and search [195, 197, 198]. Therefore, we provide a high-level summarization of representative explanation techniques in information retrieval, with the expectation of inspiring similar success in explaining the retriever of RAG.

Based on [8], the explanation methods in information retrieval can be categorized into post-hoc explanation, axiomatic strategies, probing strategies, and self-interpretable designs. The post-hoc explainers explain the models after they make decisions, the representative examples of which used in information retrieval are feature attribution and generative approach. The feature attribution

Table 6. Taxonomy for RAG Explainability

MODULE	REFERENCE	TASK	YEAR
Generation	Sudhi et al. [146]	English and German QA	2024
	Luo et al. [101]	Knowledge Graph QA	2024
	Rorseth et al. [130]	Open-book QA	2024
Retrieval + Generation	Kunze et al. [153]	Scene-Understanding	2024
	Hussien et al. [65]	Road User Intention Explanation	2024
	Ferraretto et al. [44]	Document Retrieval	2023

works by ascribing the retrieved outcomes to certain input (i.e., the attribution). Some of the methods find the explanation features by computing the feature importance, such as [122] uses interpretable textual features to explain rankings, [120] understands the BERT-based ranking models by the attention scores of tokens, and [160] estimates the point-wise explanations by analyzing the contribution of each token to the output of the ranking model. Other methods try to explain the model outputs by finding the most explanatory features, e.g. [142] uses a greedy search-based algorithm to obtain a subset of features that serve as the explanations. Axiomatic explainers provide explanations using axioms [162] and probing explainers provide valuable insights into the inner workings of neural models by revealing what types of information are encoded in their embeddings and model parameters [25], how sensitive they are to various textual properties [102], and what knowledge they possess [23, 45]. Although self-interpretable explainers inherently provide explanations by their designs, making the model fully transparent is extremely challenging, and usually, only specific components are interpretable and transparent [45, 199].

7.1.2 Explainability in Generation. Explaining the generator in RAG systems is as important as explaining the retriever since the final output is directly produced through generation. Errors in this stage can arise from shortcuts or reliance on misleading features [33, 60], and LLM-based generators are particularly prone to hallucinations. This raises concerns about whether the model’s output genuinely reflects the query-relevant content and retrieved evidence [134, 146].

Explanation techniques for generation can be grouped into two main categories: post-hoc and ante-hoc. Post-hoc methods aim to interpret the generator after inference. For example, RAG-Ex [146] offers a model-agnostic, perturbation-based framework that measures how changes to input tokens affect the output. It proposes six types of perturbations (e.g., token removal, noise injection, entity or word substitutions), then ranks token importance based on the impact on generated responses. Among them, the “leave-one-token-out” strategy proves most effective at highlighting critical inputs. Another method, RAGE [130], generates counterfactuals to determine the provenance and salience of external knowledge used by the generator. It incorporates pruning strategies to reduce the search space, making the explanation process more efficient.

In contrast, ante-hoc methods incorporate explainability directly into the generation process. For example, RoG [101], leverages knowledge graphs to generate reasoning paths for each query, which provide intermediate steps that the LLM follows to arrive at an answer. This both enhances transparency and improves generation quality by encouraging structured faithful reasoning.

7.2 Dual Enhancement of Explanations and RAG

In addition to exploring how to explain RAG systems, existing work also investigated the dual enhancement of explanations and RAG. On the one hand, explanations can be integrated to augment RAG systems. On the other hand, RAG systems can also be employed to provide explanations.

As retrieval plays a critical role in RAG [29], the enhancement of adding explanations during retrieval could benefit the whole RAG system. ExaRanker [44] utilizes the explanations as additional

111:20

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

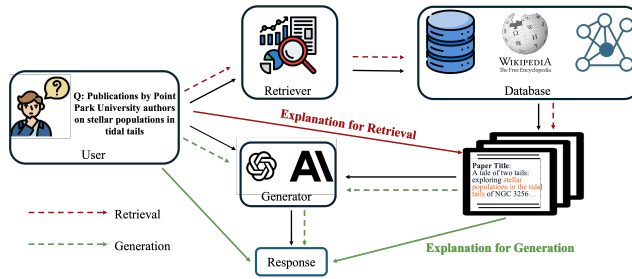


Fig. 4. An Overview of Explainability in RAG

labels to train the ranking models in the information retrieval task. Specifically, given the question-passage pair and the label indicating whether the passage can be used to answer the question, an LLM is first employed to generate the explanations of why the question can/cannot be answered by the given passage. With these ground-truth explanations, the ranking model is trained on question-passage pairs to predict not only whether the question can be answered but also the corresponding explanations. By integrating explanations as additional training labels, the ranking model can better understand the relationships between the questions and passages, which could benefit the ranking performance and ease the demand for a large number of training examples.

Apart from leveraging explanations to augment RAG systems, the reverse relationship—using RAG to improve explanations—is also worth exploring. For example, Tekkesinoglu and Kunze [153] use RAG in scene-understanding tasks to create explanations through a question-answering approach. For each input with a class label, the model predicts the probability of belonging to that class. To assess the impact of each semantic feature, it generates predictions without specific features, enabling the calculation of feature importance. These outputs, along with the contrastive cases, contribute to an external knowledge repository for LLMs. This RAG design can thus generate faithful explanations for the prediction model of the scene-understanding tasks. Another work [65] studying road user behavior also uses RAG to generate explanations. In particular, this work first creates a human-readable document that explains why the road user may/may not have a specific behavior. The document is then processed to form a database, serving as the external knowledge base of the RAG system. Given a tailored prompt and a query derived from the prediction frame, the RAG system will create a detailed explanation of the road user's intention.

7.3 Explainability Evaluation

Metrics. As very few works investigate explainability in the context of RAG, we first review the conventional explanation metrics used in explainable artificial intelligence (XAI). *Fidelity* is one commonly used evaluation metric. It measures to which extent the explanation can accurately reflect the decision-making process of the model [6]. Mathematically, fidelity is often defined as the proportion of data samples where the predictive model and the explanation produce the same decision, but there are some variations on computing fidelity, such as using Kullback-Leibler divergence between outputs, conditional entropy, and correlation [108]. Another metric often used is *stability* [50, 86, 119]. It measures the consistency of a method in producing similar explanations for similar or closely related inputs [161].

To evaluate the explanations of the generator in a RAG framework, Sudhi et al. [146] use two key metrics: significance and plausibility. In their framework, significance measures whether the explanations capture the core information present in the input. This is quantified using the F1-score and Mean Reciprocal Rank (MRR). On the other hand, plausibility is assessed through human evaluation. Annotators identify specific tokens from the input as ground-truth explanations, and

the generated explanations are then evaluated against these selected tokens using the F1-score. This dual-metric approach ensures both objective and subjective evaluation of explanation quality.

Datasets. Although the explainability of RAG enhances user trust and transparency in the generated outputs, evaluating the explanations remain challenging due to the lack of standardized datasets specifically designed for this purpose. Currently, researchers often adapt existing QA datasets to test explainability methods. For example, Sudhi et al. [146] utilized randomly sampled English and German QA pairs from the validation split of the XQUAD dataset to evaluate their proposed explainer. Similarly, Luo et al. [101] demonstrated the self-explanatory capabilities of their RoG method by automatically generating explanations while performing question-answering tasks on two benchmark knowledge graph question-answering (KGQA) datasets: WebQuestionsSP (WebQSP) [186] and Complex WebQuestions (CWQ) [152]. These efforts illustrate that, in the absence of standardized benchmarks, evaluation datasets are specific to different tasks and domains, and researchers rely on existing datasets. The development of dedicated datasets for explainability in RAG remains a critical area for future research, offering the potential to advance systematic evaluation and comparison of explainability methods.

7.4 Future Directions of RAG Explainability

Integrating knowledge graphs with LLMs offers a promising path to improve retrieval faithfulness in RAG systems [110]. The hybrid approach combines GNN-based retrieval with LLM prompting to support more robust and flexible solutions. Extending this integration to semi-structured knowledge bases could enrich context for generation and personalization. Future research can focus on optimizing the interaction between components and the structure of the underlying knowledge bases for more diverse applications.

Moreover, the multi-component nature of RAG systems, with tightly coupled retrievers and generators, introduces unique explainability challenges. Jointly trained models, such as those sharing embeddings [39, 84], make it difficult to isolate the contributions of each component. New retrievers using LLMs for graph traversal [68, 166] further complicate interpretation. Traditional explanation techniques often fall short here. Future work should develop new methods to clarify both the roles of individual components and their interactions.

Similar to fairness, the trade-off between performance and explainability remains a contested issue. While some argue that improving one harms the other [28], others disagree [13]. In RAG systems, replacing black-box components with explainable ones could provide insight into this balance. More research is needed to assess whether transparency can coexist with performance.

Evaluating explainability in RAG is still an open challenge. While metrics like fidelity and stability are helpful, they often miss context-specific concerns. Emerging approaches like RAG-Ex introduce metrics such as significance (alignment with known facts) and plausibility (alignment with human judgment). Expanding domain-aware metrics will better assess explanation quality and user trust.

Last but not least, misinformation is a major risk in RAG, particularly in multi-step reasoning. Explainability methods that trace how information flows through retrieval and generation components could help detect and mitigate errors. Understanding how false information impacts generation can guide the development of more reliable systems. Future work should explore how explainability can support misinformation detection and prevention.

8 ACCOUNTABILITY OF RETRIEVAL AUGMENTED GENERATION

Accountability in AI refers to determining whether outputs comply with established procedural and substantive standards and identifying responsible entities when violations occur [35]. In generative AI, a key technical challenge here is attributing *ownership* as the origin of the outputs is often

Table 7. Taxonomy for RAG Accountability

MODULE	REFERENCE	TECHNIQUE	TYPE	YEAR
<i>Retrieval</i>	Liu et al. [91]	Format-based	Text Watermarking	2024
	Xu et al. [175]	Embedding-based	Data Watermarking	2024
	Sun et al. [150]	Trigger-based	Data Watermarking	2022
<i>Generation</i>	Christ et al. [24]	Semantic-based	Sentence-level Watermarking	2024
	Hou et al. [61]	Sampling-based	Token-level Watermarking	2023
	Kirchenbauer et al. [73]	Logit-based	Global Watermarking	2023
	Yang et al. [183]	Post-generation	Lexical/Syntactic	2022
<i>Retrieval + Generation</i>	Jovanovic et al. [70]	Integrated Pipeline	Red-Green Token Scheme	2024

uncertain. Robust content-tracing techniques, such as *watermarking*, are therefore crucial for linking outputs back to source models or datasets. By clearly associating content with its origins, watermarking bridges technical and policy-oriented accountability. Thus, we focus this section on watermarking in RAG-based LLMs.

8.1 Taxonomy of RAG Accountability

Retrieval. Accountability in retrieval in this survey refers to embedding watermarks within the sources of retrieval, encompassing both Text Watermarking and Data Watermarking techniques. These approaches aim to safeguard content and data ownership, ensuring accountability and traceability in RAG systems.

Generation. Watermarking is key to ensuring accountability in LLM outputs, with strategies applied at different stages. Pre-generation embeds markers during training, in-generation integrates them during inference, and post-generation adds them after text is produced. Each method offers distinct advantages, collectively enhancing traceability and authenticity. We summarize the RAG accountability taxonomy in Table 7.

8.2 Accountability in Retrieval

8.2.1 Text Watermarking. Text Watermarking involves embedding identifiable markers into textual content to protect copyright and authenticate ownership of the author [91]. The common techniques for watermarking are shown in Figure 5. Format-based watermarking algorithms are commonly employed as they only embed watermarks in the text format without altering the author’s contents [91]. This includes line or word shifting methods [18] and unicode-based approaches [121, 128, 133]. Specifically, the former involves adjusting text lines or words vertically and horizontally, effectively used in image-format texts, while the latter usually involves inserting or replacing Unicode codepoints such as whitespace for watermarking. Beyond the above methods, other innovative techniques have emerged, including variation in text color or font [104] and feature embedding, e.g., bookmarks or variables [66]. While these approaches signify the promise of format-based watermarking in preserving text copyright, we should also be wary of their potential vulnerability to adversary attacks such as removal using canonicalization [16] and watermark forgery due to the detectable pattern in watermarked text formats [91].

8.2.2 Data Watermarking. Data Watermarking addresses the increasing need to protect datasets used during the training of machine learning models, ensuring proper attribution and preventing unauthorized usage. A key technique in this area is backdoor watermarking, which embeds ownership information directly into the trained model by introducing trigger-specific input modifications that prompt unique, identifiable behaviors in the model.

Trigger-based watermarking is widely employed due to its flexibility and effectiveness as a type of backdoor watermarking. These triggers can take various forms, including word- or sentence-level modifications [150], semantically invariant transformations in code [149], or distinctive input formats designed to be recognizable [175]. Embedding such triggers ensures that ownership can be verified through the model's behavior, even if the dataset itself is no longer accessible.

While trigger-based watermarking is robust, its effectiveness depends on careful design to ensure triggers remain inconspicuous yet detectable. Additionally, as models become more complex and versatile, challenges such as ensuring trigger persistence and avoiding unintended activations must be addressed. As the field evolves, continued innovation in embedding mechanisms and detection strategies will be essential for robust and scalable dataset protection.

8.3 Accountability in Generation

8.3.1 Pre-generation Watermarking. Pre-generation watermarking involves embedding watermarks during the training phase of LLMs, creating inherent markers within the model's outputs. This approach can be categorized into trigger-based watermarks and global watermarks.

Trigger-based watermarking has been described in detail in Section 8.2.2. As a localized method, trigger-based watermarking relies on specific inputs to reveal ownership, which minimizes its impact on regular outputs but may limit its detection capabilities in broad use cases.

On the other hand, global watermarking ensures pervasive traceability across outputs but requires careful design to balance robustness with imperceptibility, ensuring the watermark does not degrade quality or usability of generated content. *Global watermarking* embeds markers in all generated outputs, enabling consistent content tracking without the need for specific triggers. This integrates watermarking directly into the model's parameters, with methods including sampling-based or logit-based watermark distillation [53] and reinforcement learning with feedback from watermark detectors [177]. Global watermarking's broader applicability makes it an attractive option for large-scale deployment, especially in scenarios where consistent tracking of all outputs is critical.

Together, pre-generation watermarking provides a proactive means of embedding accountability into LLMs. Future research should focus on hybrid approaches that combine the specificity of trigger-based watermarking with the universality of global watermarking, enabling robust, scalable solutions that balance protection and practicality.

8.3.2 In-generation Watermarking. In-generation watermarking directly embeds watermarks as the LLM produces text at inference time. Unlike pre-generation watermarking, which embeds markers into the model during training, in-generation watermarking dynamically modifies generated outputs, providing flexibility without altering model parameters. This can be classified into two main methods: watermarking in logit generation and watermarking in token sampling.

Watermarking in logit generation involves modifying the logits during inference. This approach is both versatile and cost-effective, as it avoids the need for model retraining. Typically, KGW [73] partitions the vocabulary into distinct categories, such as red and green token lists, using a hash function that depends on the preceding token. The watermark is embedded by biasing the selection of tokens from one category (e.g., green tokens) during text generation. The detection of KGW involves analyzing the proportion of green tokens in the output and computing a z-score to determine whether the text is watermarked, reported with a high detection performance. However, its performance can degrade in low-entropy contexts, such as code generation, where token probabilities are unevenly distributed. Optimization techniques, such as entropy-based weighting [100] and sliding window methods [74], have been proposed to enhance detectability and robustness in these challenging scenarios.

111:24

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

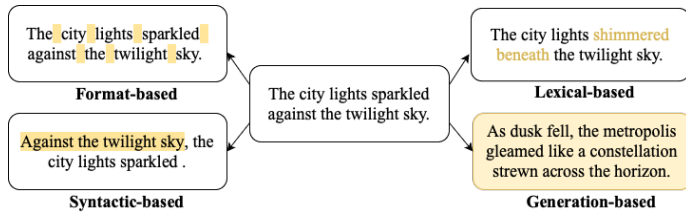


Fig. 5. An illustration of the common watermarking techniques.

In addition to modifying logits, in-generation watermarking can be applied during token sampling at both the token and sentence levels. Token-level methods embed watermarks by biasing the random seed or pseudo-random number generator used in token selection. For example, Christ et al. [24] use fixed random sequences to align outputs with pre-defined patterns, though such methods are sensitive to text edits. To improve robustness, sentence-level approaches like SemStamp [61] operate on the semantic space, partitioning it into watermarked and non-watermarked regions. This ensures that entire sentences preserve watermark properties even after paraphrasing, enabling more resilient and semantically meaningful detection.

In-generation watermarking offers significant advantages, such as flexibility and adaptability. However, challenges remain, including mitigating the impact on text quality, enhancing robustness against removal attacks, and achieving public verifiability. Techniques such as fine-grained vocabulary partitioning [43] and semantic-aware watermarking [47] show promise in addressing these issues. Future research should explore hybrid methods that combine the strengths of logits-based and sampling-based approaches, providing both robustness and adaptability. Additionally, developing standards for evaluating the effectiveness and detectability of in-generation watermarks across diverse applications will be critical for their broader adoption in real-world scenarios.

8.3.3 Post-generation Watermarking. Post-generation watermarking involves embedding watermarks into already-generated text, which can be categorized into four primary methods: Format-based watermarking, lexical-based, syntactic-based, and generation-based watermarking[91].

Format-based watermarking has already been discussed in Section 8.2.1 which displays vulnerability to adversary attacks due to detectable patterns. To address this, Lexical-based watermarking advances by word substitution without altering the original textual semantics. Early methods, such as those using WordNet or Word2Vec [42, 105, 156], were limited by their lack of context awareness. Recent advances, like BERT-based infill models [183, 187], have improved context sensitivity, enabling more robust watermarking.

Syntactic-based watermarking modifies the grammatical structure of sentences to embed watermarks. This involves transformations such as adjunct movement and clefting. Research later expanded with activization and topicalization [11, 155]. While effective, these methods are often language-dependent and may require customization to adhere to grammatical rules. Excessive syntactic changes can also disrupt the original style and fluency of the text [91].

Generation-based watermarking leverages advanced neural network models to directly generate watermarked text from the original content and a watermark message. These approaches, such as AWT [1] and REMARK-LLM [194], utilize transformer-based architectures to embed high-capacity watermarks while maintaining the quality and naturalness of the text. Techniques like WATERFALL [81] further enhance fluency by using LLMs for paraphrasing, ensuring seamless integration of watermarks. This method offers high detectability, robustness, and scalability, making it a promising solution for embedding watermarks in LLM-generated content.

Table 8. Watermarking Datasets and Metrics

Dataset	Metric		
	Detectability	Quality Impact	Robustness
WaterBench [159]	✓		
WaterJudge [106]	✓	✓	
Mark My Words [117]	✓		✓
MarkLLM [112]	✓	✓	✓

8.4 Accountability in RAG Systems

WARD (Watermarking for RAG Dataset Inference) [70] introduces a unified approach to watermarking in RAG systems by embedding imperceptible signals into datasets to enable traceability through both retrieval and generation, even after paraphrasing. Its red-green token scheme ensures statistically robust detection, resists content blending, and offers low false detection rates with scalable query aggregation. We recognize it as one of the first methods to provide tracking across both retrieval and generation stages in RAG. This marks a step toward accountable and transparent data usage in multi-stage RAG pipelines.

8.5 Accountability Evaluation

8.5.1 Metrics. Evaluating watermarking techniques for large language models involves a comprehensive set of metrics to assess detectability, quality, output performance, diversity, and robustness [91]. These ensure watermark effectiveness while minimizing quality loss and resisting attacks.

Detectability measures how reliably a watermark can be identified. Zero-bit watermarking detects presence without decoding content, using statistical tools like z-scores or p-values. Multi-bit watermarking recovers embedded data and is evaluated using Bit Error Rate (BER) [187], bit accuracy [188], and required watermark size [116], with longer texts aiding detection but limiting short-text applicability. Watermarked text must maintain high quality. Comparative metrics (e.g., BLEU [113], Meteor [7], Semantic Score, Entailment Score) compare with unmarked outputs, while single-text metrics like Perplexity (PPL) assess coherence directly. Human evaluation remains the gold standard. Output performance ensures that watermarking does not hinder LLM capabilities. Text tasks use PPL [187], GPT-4 scoring [38], and semantic similarity. Code generation uses CodeBLEU [54] and Edit Sim [159]. Other tasks (e.g., summarization, translation, QA) are evaluated with BLEU, ROUGE [90], and task-specific metrics. Diversity metrics assess how watermarking affects variability. Seq-Rep-N [53] and Log Diversity [75] measure lexical variation. Entropy-based metrics like Ent-3 [61] and Sem-Ent [55] capture lexical and semantic diversity, with higher scores indicating better variability. Robustness assesses a watermark’s resilience to both untargeted and targeted attacks (see Section 5). A strong watermark must remain detectable even under adversarial transformations.

8.5.2 Datasets. According to various metrics discussed above, several benchmarks and toolkits have been developed to standardize the evaluation of text watermarking techniques in LLMs, including WaterBench [159], WaterJudge [106], Mark My Words [117], and MarkLLM [112]. The details about each dataset can be viewed in Table 8.

8.6 Future Direction of RAG Accountability

Ensuring accountability in RAG systems calls for a unified approach that integrates watermarking across both retrieval and generation stages. Current techniques operate independently, leaving gaps in traceability and weakening intellectual property protection. Future research should develop holistic frameworks that embed imperceptible signals throughout the entire RAG pipeline, enabling consistent attribution and reinforcing data and model integrity as retrieval and generation

111:26

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

increasingly intertwine. As AI systems become more adaptive to real-time inputs and evolving user contexts, static watermarking schemes may fall short. **Dynamic watermarking** methods—capable of adjusting to model updates, resisting adversarial attacks, and preserving traceability across system changes—will be essential for robust accountability. Beyond technical solutions, effective **governance and ethical integration** are also critical. Collaborations among researchers, policy-makers, and ethicists are needed to embed watermarking into legal and regulatory frameworks, ensuring RAG systems meet standards for transparency, intellectual property protection, and responsible AI use.

9 APPLICATIONS

In previous sections, we outlined the six dimensions of Trustworthy RAG systems. We now explore how these principles apply in high-stakes domains—Healthcare, Legal, and Education—highlighting recent advancements, current use cases, and domain-specific challenges.

9.1 Healthcare

9.1.1 RAG Use Cases in Healthcare.

- **Clinical Decision Support:** LLMs are widely used for Clinical Decision Support (CDS), providing access to diagnostic guidelines, treatment plans, and patient history [83]. While many models rely on fine-tuning, RAG enhances decision-making by retrieving relevant medical documents and personalized patient data [174].
- **Patient Communication:** RAG systems support patient-facing applications like question answering and dialogue [57]. With 58% of U.S. adults searching online for medical information [20], there is a strong need for accurate and personalized health advice. RAG-based chatbots can provide symptom assessments and guide next steps with improved reliability.
- **Knowledge Discovery:** In drug discovery, RAG can accelerate research by retrieving up-to-date literature and clinical trial data [182]. LLMs complement traditional graph-based approaches by aiding hypothesis generation and scientific exploration [103, 111].
- **Precharting:** Precharting involves reviewing patient records before visits [17]. While still underexplored, RAG systems can improve efficiency by enhancing personalization and contextual relevance [143, 182].

9.1.2 Trustworthiness Challenges in Healthcare. **Reliability:** Due to the high-stakes nature of healthcare, RAG systems must be highly reliable. Although uncertainty quantification is common in traditional settings [135], its use in LLMs is limited. Integrating it into CDS and patient communication can help users assess confidence in outputs [101, 110, 148]. Domain-specific robustness and adaptation to individual differences are key. **Privacy:** Protecting sensitive patient data is essential. RAG systems must prevent personal information leakage, especially in applications like precharting [17, 140]. Future work should explore privacy-preserving mechanisms suitable for real-world healthcare contexts. **Others:** Beyond reliability and privacy, RAG systems must be safe, fair, explainable, and accountable. They should resist adversarial manipulation, avoid bias, offer transparent reasoning, and support traceability of errors.

9.2 Law

9.2.1 RAG Use Cases in Law.

- **Legal Question Answering:** RAG systems are increasingly used in Legal Question Answering (LQA), which involves responding to queries about laws, legal procedures, and case precedents [22, 80]. Because laws differ across jurisdictions, many approaches fine-tune LLMs on domain-specific legal corpora [3]. RAG systems complement these methods by

dynamically retrieving relevant statutes and precedents, enabling more contextualized and accurate responses [170].

- **Legal Document Summarization:** Legal documents are often long, technical, and challenging to interpret. Legal Document Summarization helps condense content into digestible and accurate summaries [10, 22]. While current approaches leverage LLMs, often in combination with rule-based templates, the integration of RAG remains underexplored. By retrieving case law, statutes, and relevant legal commentary, RAG-enhanced summarization could significantly improve factual grounding, contextual relevance, and overall summary utility [93].
- **Legal Judgment Prediction:** Legal Judgment Prediction (LJP) seeks to predict court rulings based on factual case descriptions [22]. Early models treated LJP as a classification problem [30], but this oversimplifies legal reasoning. Recent approaches integrate RAG to retrieve similar precedent cases and applicable laws [173], thereby enhancing transparency and improving the model's ability to reflect legal argumentation and judicial logic.

9.2.2 Trustworthiness Challenges in Legal Applications. **Fairness:** Ensuring fairness is essential in legal RAG applications, as biased outputs could contribute to unjust legal outcomes or perpetuate discrimination [141]. Biases in training data or retrieval mechanisms can lead to the selection of prejudiced precedents or the omission of underrepresented legal viewpoints. Addressing these issues requires targeted de-biasing techniques and retrieval strategies designed specifically for legal domains [48]. **Explainability:** Legal reasoning demands transparency. Legal professionals must be able to trace how conclusions were reached, especially in tasks like legal opinion drafting or judgment prediction [22, 30]. RAG systems must clearly indicate the sources of retrieved content and how it contributed to the final output. Recent work has begun to explore interpretable methods for long-form legal QA [98], but more domain-specific explainability frameworks are needed to meet legal standards. **Others:** Like in healthcare, legal RAG systems must also be reliable, private, and accountable. Outputs must be factually accurate and robust to noisy or ambiguous input. Privacy is critical when dealing with confidential client records or sealed case documents. Finally, accountability mechanisms should ensure that errors or harmful outputs can be traced back and addressed appropriately to maintain trust in legal applications.

9.3 Education

9.3.1 RAG Use Cases in Education.

- **Personalized Learning:** Personalized learning is a core educational application of LLMs [165], with recent research exploring learning path planning [109]. However, many existing approaches lack adaptability to individual needs. Integrating RAG allows systems to dynamically retrieve relevant educational content based on student progress and knowledge gaps, leading to more adaptive, engaging, and effective learning experiences.
- **Student Support:** RAG systems can serve as intelligent academic assistants, answering students' questions and clarifying concepts in real time [32]. Unlike static systems, RAG can pull updated and contextually relevant information to provide more accurate explanations. This reduces the support burden on educators and helps students overcome obstacles independently and efficiently.
- **Teacher Support:** Teachers can benefit from RAG systems that assist in lesson planning, content generation, and responding to student queries using real-time, domain-specific resources. Preliminary studies using LLMs with sources like Reddit show promise for enhancing classroom support [107]. However, further research is needed to develop high-quality educational datasets and refine RAG models for diverse classroom settings.

111:28

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

9.3.2 Trustworthiness Challenges in Educational Applications. **Fairness:** Fairness is essential in educational contexts to ensure equal learning opportunities for all students. LLMs and RAG systems may unintentionally reinforce biases present in training or retrieved data [138, 139], leading to unequal treatment or content that disadvantages underrepresented groups. Educational RAG systems must be carefully designed to detect and mitigate such biases. **Safety:** RAG systems are susceptible to adversarial attacks and jailbreaks [34, 169, 178], which can be particularly harmful in educational environments. Students might be exposed to misleading or inappropriate content if safety mechanisms are bypassed. Tailored defenses are crucial to ensure RAG systems remain safe, especially when deployed for minors or in public learning platforms. **Others:** Additional trustworthiness concerns include reliability, privacy, and robustness. Educational RAG systems must consistently deliver accurate and relevant content. They must also protect sensitive data like student performance and demographics. Robustness is needed to handle diverse learners, ambiguous inputs, and evolving educational goals while maintaining trustworthy performance.

10 CONCLUSION

In this survey, we provide a comprehensive review of RAG systems through the lens of trustworthiness, focusing on six critical aspects: reliability, privacy, safety, fairness, explainability, and accountability. For each trustworthiness aspect, we introduce definitions and key concepts to establish an understanding of the topics. We also offer a structured taxonomy to help researchers navigate the diverse approaches in the specific trustworthy aspect. Beyond methodological summary, we highlight the evaluation protocols commonly used for trustworthy RAG systems, including the specific datasets and metrics. By including these discussions, we aim to facilitate the development of benchmarks tailored to trustworthiness challenges.

Finally, we provide an in-depth discussion of future research directions for each aspect of trustworthiness, including promising directions within individual aspects and potential synergies across multiple areas. By addressing these directions, we hope to inspire innovative approaches that enhance the overall trustworthiness of RAG systems and drive their broader adoption in critical applications. This survey not only serves as a comprehensive roadmap for researchers aiming to advance trustworthy RAG systems but also underscores the importance of addressing these challenges to ensure the safe deployment of RAG applications.

REFERENCES

- [1] Sahar Abdelnabi and Mario Fritz. 2021. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 121–140.
- [2] Daniel Adiwardana, Minh-Thang Luong, et al. 2020. Towards a Human-like Open-Domain Chatbot. In *ICLR 2020*.
- [3] Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. In *EMNLP 2020*. Association for Computational Linguistics, Online, 743–749.
- [4] Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4. arXiv:2311.07361 [cs.CL]
- [5] Rama Akkiraju, Anbang Xu, Deepak Bora, Tan Yu, Lu An, Vishal Seth, et al. 2024. FACTS About Building Retrieval Augmented Generation-based Chatbots. arXiv:2407.07858 [cs.LG]
- [6] Nourah Alangari, Mohamed El Bachir Menai, Hassan Mathkour, and Ibrahim Almosallam. 2023. Exploring evaluation methods for interpretable machine learning: A survey. *Information* 14, 8 (2023), 469.
- [7] Mohammed Hazim Alkawaz, Ghazali Sulong, Tanzila Saba, et al. 2016. Concise analysis of current text automation and watermarking approaches. *Security and Communication Networks* 9, 18 (2016), 6365–6378.
- [8] Avishkek Anand, Procheta Sen, Sourav Saha, Manisha Verma, and Mandar Mitra. 2023. Explainable information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference*, 3448–3451.
- [9] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. 2022. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. arXiv:2110.01052 [cs.LG]
- [10] Anthropic. 2024. Legal Summarization - Claude Use Case Guide. <https://docs.anthropic.com/en/docs/about-claude/use-case-guides/legal-summarization> Accessed: 2025-02-03.

- [11] Mikhail J Atallah, Victor Raskin, et al. [n. d.]. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In Information Hiding: 4th International Workshop.
- [12] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, et al. 2018. MS MARCO: A Human Generated MACHine Reading COMprehension Dataset. [arXiv:1611.09268 \[cs.CL\]](#)
- [13] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In ACM FAccT. 248–266.
- [14] Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the Gap Between Consumers' Medication Questions and Trusted Answers. Studies in Health Technology and Informatics 264 (August 21 2019), 25–29.
- [15] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In WACV.
- [16] Nicholas Boucher, Iliia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad characters: Imperceptible nlp attacks. In 2022 IEEE Symposium on Security and Privacy (SP). IEEE, 1987–2004.
- [17] CA Bowman and H Holzer. 2021. EMR Precharting Efficiency in Internal Medicine: A Scoping Review. J Med Educ Curric Dev 8 (Jul 2021), 23821205211032414.
- [18] Jack T Brassil, Steven Low, Nicholas F Maxemchuk, and Lawrence O'Gorman. 1995. Electronic marking and identification techniques to discourage document copying. Journal on Selected Areas in Communications (1995).
- [19] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2019. Assessing the Usability of a Chatbot for Mental Health Care. In Internet Science.
- [20] Centers for Disease Control and Prevention (CDC). 2023. Internet Use for Health Information and Communications Among Adults: United States, 2022. <https://www.cdc.gov/nchs/products/databriefs/db482.htm> Accessed: 2025-01-29.
- [21] Mark Chen, Jerry Tworek, Heewoo Jun, and Others. 2021. Evaluating Large Language Models Trained on Code. (2021). [arXiv:2107.03374 \[cs.LG\]](#)
- [22] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. [arXiv preprint arXiv:2405.01769](#) (2024).
- [23] Jaekel Choi, Euna Jung, Sungjun Lim, and Wonjong Rhee. 2022. Finding Inverse Document Frequency Information in BERT. [arXiv preprint arXiv:2202.12191](#) (2022).
- [24] Miranda Christ, Sam Gunn, and Or Zamir. 2024. Undetectable watermarks for language models. In The Thirty Seventh Annual Conference on Learning Theory. PMLR, 1125–1139.
- [25] Daniel Cohen, Brendan O'Connor, and W Bruce Croft. 2018. Understanding the representational power of neural retrieval models using NLP tasks. In Proceedings of the SIGIR 2018. 67–74.
- [26] Stav Cohen, Ron Bitton, and Ben Nassi. 2024. Unleashing Worms and Extracting Data: Escalating the Outcome of Attacks against RAG-based Inference in Scale and Severity Using Jailbreaking. [arXiv preprint arXiv:2409.08045](#).
- [27] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 deep learning track. [arXiv preprint arXiv:2003.07820](#) (2020).
- [28] Barnaby Crook et al. 2023. Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). In 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW).
- [29] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonello, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In ACM SIGIR.
- [30] Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges. [arXiv:2204.04859 \[cs.CL\]](#)
- [31] Eliot Dai, Tianhao Zhao, Hongfu Zhu, et al. 2024. A Comprehensive Survey on Trustworthy Graph Neural Networks: Privacy, Robustness, Fairness, and Explainability. Machine Intelligence Research 21 (2024), 1011–1061.
- [32] Sagnik Dakshit. 2024. Faculty Perspectives on the Potential of RAG in Computer Science Higher Education. [arXiv:2408.01462 \[cs.CY\]](#)
- [33] Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2024. CrAM: Credibility-Aware Attention Modification in LLMs for Combating Misinformation in RAG. [arXiv preprint arXiv:2406.11497](#) (2024).
- [34] Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, and Yang Liu. 2024. Pandora: Jailbreak GPTs by Retrieval Augmented Generation Poisoning. [arXiv:2402.08416 \[cs.CR\]](#)
- [35] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, et al. 2017. Accountability of AI under the law: The role of explanation. [arXiv preprint arXiv:1711.01134](#) (2017).
- [36] Michael D Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2023. Overview of the TREC 2022 fair ranking track. [arXiv preprint arXiv:2302.05558](#) (2023).
- [37] OpenAI et al. 2024. GPT-4 Technical Report. [arXiv:2303.08774 \[cs.CL\]](#)
- [38] Jaiden Fairoze, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Mingyuan Wang. 2023. Publicly-Detectable Watermarking for Language Models. Cryptology ePrint Archive, Paper 2023/1661.

111:30

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

- [39] Wenqi Fan, Yajuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey On Rag Meeting LLMs: Towards retrieval-augmented large language models. In *ACM KDD*. 6491–6501.
- [40] Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing Noise Robustness of Retrieval-Augmented Language Models with Adaptive Adversarial Training. *arXiv preprint arXiv:2024.05* (2024).
- [41] Philip Feldman, James R. Foulds, and Shimei Pan. 2024. RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots. *arXiv:2403.01193* [cs.CL]
- [42] Christiane Fellbaum. 1998. WordNet: An electronic lexical database. MIT Press *google schola* 2 (1998), 678–686.
- [43] Pierre Fernandez, Antoine Chaffin, et al. 2023. Three bricks to consolidate watermarks for large language models. In *IEEE International Workshop on Information Forensics and Security (WIFS)*. 1–6.
- [44] Fernando Ferraretto, Thiago Laitz, Roberto Lotufo, and Rodrigo Nogueira. 2023. Exaranker: Explanation-augmented neural ranker. *arXiv preprint arXiv:2301.10521* (2023).
- [45] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. A white box analysis of ColBERT. In *ECIR*. Springer, 257–263.
- [46] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*.
- [47] Yu Fu, Deyi Xiong, and Yue Dong. 2024. Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy. In *Proceedings of the AAAI Conference*, Vol. 38. 18003–18011.
- [48] Isabel O. Gallegos, Ryan A. Rossi, et al. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* (09 2024).
- [49] Yunfan Gao, Yun Xiong, Xinyu Gao, et al. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2023).
- [50] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *AAAI*.
- [51] Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. *arXiv preprint arXiv:2207.06300* (2022).
- [52] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. 79–90.
- [53] Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. 2024. On the Learnability of Watermarks for Language Models. In *ICLR*.
- [54] Batu Guan et al. 2024. CodeIP: A Grammar-Guided Multi-Bit Watermark for Large Language Models of Code. In *EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 9243–9258.
- [55] Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and Improving Semantic Diversity of Dialogue Generation. In *EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.
- [56] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* (2016).
- [57] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion* 118 (2025), 102963.
- [58] Wenchong He, Zhe Jiang, Tingsong Xiao, Zelin Xu, and Yukun Li. 2024. A Survey on Uncertainty Quantification Methods for Deep Learning. *arXiv:2302.13425* [cs.LG]
- [59] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv:2009.03300* [cs.CY]
- [60] Giwon Hong, Jeonghwan Kim, Junmo Kang, et al. 2023. Why So Gullible? Enhancing the Robustness of Retrieval-Augmented Models against Counterfactual Noise. *arXiv preprint arXiv:2305.01579* (2023).
- [61] Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, et al. 2023. Semstamp: A semantic watermark with paraphrastic robustness for text generation. *arXiv preprint arXiv:2310.03991* (2023).
- [62] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, et al. 2024. Large Language Models for Software Engineering: A Systematic Literature Review. *ACM Trans. Softw. Eng. Methodol.* 33, 8 (2024).
- [63] Mengxuan Hu, Hongyi Wu, Zihan Guan, Ronghang Zhu, Dongliang Guo, Daiqing Qi, and Sheng Li. 2024. No Free Lunch: Retrieval-Augmented Generation Undermines Fairness in LLMs, Even for Vigilant Users. *arXiv:2410.07589* [cs.IR]
- [64] Xiaowei Huang, Wenjie Ruan, Wei Huang, et al. 2023. A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation. *arXiv:2305.11391* [cs.AI]
- [65] Mohamed Manzour Hussien, Angie Nataly Melo, Augusto Luis Ballardini, Carlota Salinas Maldonado, Rubén Izquierdo, and Miguel Ángel Sotelo. 2024. RAG-based Explainable Prediction of Road Users Behaviors for Automated Driving using Knowledge Graphs and Large Language Models. *arXiv preprint arXiv:2405.00449* (2024).

- [66] Muhammad Munwar Iqbal, Umair Khadam, Ki Jun Han, Jihun Han, and Sohail Jabbar. 2019. A robust digital watermarking algorithm for text document copyright protection based on feature coding. In *IWCMC*. IEEE.
- [67] Changyue Jiang, Xudong Pan, Geng Hong, Chenfu Bao, and Min Yang. 2024. RAG-Thief: Scalable Extraction of Private Data from Retrieval-Augmented Generation Applications with Agent-based Attacks. *arXiv:2411.14110 [cs.CR]*
- [68] Bowen Jin, Chulin Xie, Jiawei Zhang, et al. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. *arXiv preprint arXiv:2404.07103 (2024)*.
- [69] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th ACL*.
- [70] Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. 2024. Ward: Provable RAG Dataset Inference via LLM Watermarks. *arXiv preprint arXiv:2410.03537 (2024)*.
- [71] Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. SuRe: Summarizing Retrievals using Answer Candidates for Open-domain QA of LLMs. *arXiv preprint arXiv:2404.13081 (2024)*.
- [72] To Eun Kim and Fernando Diaz. 2024. Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation. *arXiv preprint arXiv:2409.11598 (2024)*.
- [73] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*. PMLR, 17061–17084.
- [74] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, et al. 2023. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634 (2023)*.
- [75] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, et al. 2024. On the Reliability of Watermarks for Large Language Models. In *ICLR*.
- [76] Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *Machine Learning: ECML 2004 (Lecture Notes in Computer Science, Vol. 3201)*. Springer, 217–226.
- [77] Fanjie Kong, Shuai Yuan, Weituo Hao, and Ricardo Henao. 2023. Mitigating test-time bias for fair image retrieval. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
- [78] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal Prediction with Large Language Models for Multi-Choice Question Answering. *arXiv:2305.18404 [cs.CL]*
- [79] Tom Kwiatkowski, Jennimaria Palomaki, et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466.
- [80] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2023. Large Language Models in Law: A Survey. *arXiv:2312.03718 [cs.CL]*
- [81] Gregory Kang Ruey Lau, Xinyuan Niu, et al. 2024. Waterfall: Framework for robust and scalable text watermarking. In *ICML 2024 Workshop on Foundation Models in the Wild*.
- [82] Peter Lee, Sebastian Bubeck, and Joseph Petro. 2023. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine* 388, 13 (2023), 1233–1239.
- [83] Alessandro Gaj Levra, Mauro Gatti, Roberto Mene, Dana Shiffer, et al. 2025. A large language model-based clinical decision support system for syncope recognition in the emergency department: A framework for clinical workflow integration. *European Journal of Internal Medicine* (2025).
- [84] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*. 9459–9474.
- [85] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2022. Trustworthy AI: From Principles to Practices. *arXiv:2110.01167 [cs.AI]*
- [86] Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi. 2020. Evaluating explanation methods for neural machine translation. *arXiv preprint arXiv:2005.01672 (2020)*.
- [87] Jiming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni. 2023. GPT4Rec: A Generative Framework for Personalized Recommendation and User Interests Interpretation. *arXiv:2304.03879 [cs.IR]*
- [88] Shuo Li et al. 2023. TRAQ: Trustworthy Retrieval Augmented Question Answering via Conformal Prediction. *arXiv preprint arXiv:2307.04642 (2023)*.
- [89] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110 (2022)*.
- [90] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.
- [91] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024. A survey of text watermarking in the era of large language models. *Comput. Surveys* 57, 2 (2024), 1–36.
- [92] Mingrui Liu, Sixiao Zhang, and Cheng Long. 2024. Mask-based Membership Inference Attacks for Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.20142 (2024)*.
- [93] Shengjie Liu, Jing Wu, Jingyuan Bao, Wenyi Wang, Naira Hovakimyan, and Christopher G Healey. 2024. Towards a Robust Retrieval-Based Summarization System. *arXiv:2403.19889 [cs.CL]*

111:32

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

- [94] Yang Liu et al. 2024. Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv preprint arXiv:2308.05374* (2024).
- [95] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt injection attack against LLM-integrated applications. *arXiv preprint arXiv:2306.05499* (2023).
- [96] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058* (2024).
- [97] 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* [cs.CL]
- [98] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. *arXiv:2309.17050* [cs.CL]
- [99] Dong Lu, Tianyu Pang, Chao Du, Qian Liu, Xianjun Yang, and Min Lin. 2024. Test-time backdoor attacks on multimodal large language models. *arXiv preprint arXiv:2402.08577* (2024).
- [100] Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. 2024. An Entropy-based Text Watermarking Detection Method. *arXiv preprint arXiv:2403.13485* (2024).
- [101] Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. In *International Conference on Learning Representations*.
- [102] 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.
- [103] J. Miao, C. Thongprayoon, S. Suppadungsuk, et al. 2024. Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications. *Medicina (Kaunas)* 60, 3 (Mar 2024), 445.
- [104] Nighat Mir. 2014. Copyright for web content using invisible text watermarking. *Comput. Hum. Behav.* 30 (2014).
- [105] Eric Mitchell, Yoonho Lee, et al. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *ICML*. PMLR.
- [106] Piotr Molenda, Adian Liusie, and Mark JF Gales. 2024. WaterJudge: Quality-Detection Trade-off when Watermarking Large Language Models. *arXiv preprint arXiv:2403.19548* (2024).
- [107] Elizabeth A Mullins, Adrian Portillo, Kristalys Ruiz-Rohena, and Aritran Piplai. 2024. Enhancing classroom teaching with LLMs and RAG. *arXiv:2411.04341* [cs.LG]
- [108] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, et al. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM CSUR* 55, 13s (2023), 1–42.
- [109] Chee Ng and Yuen Fung. 2024. Educational Personalized Learning Path Planning with Large Language Models. *arXiv:2407.11773* [cs.CL]
- [110] Bo Ni, Yu Wang, Lu Cheng, Erik Blasch, and Tyler Derr. 2025. Towards Trustworthy Knowledge Graph Reasoning: An Uncertainty Aware Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.
- [111] S. Pal, M. Bhattacharya, M. A. Islam, and C. Chakraborty. 2023. ChatGPT or LLM in next-generation drug discovery and development: pharmaceutical and biotechnology companies can make use of the artificial intelligence-based device for a faster way of drug discovery and development. *International Journal of Surgery* (2023).
- [112] Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, et al. 2024. Markllm: An open-source toolkit for llm watermarking. *arXiv preprint arXiv:2405.10051* (2024).
- [113] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.
- [114] Alicia Parrish, Angelica Chen, Nikita Nangia, et al. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193* (2021).
- [115] Yuefeng Peng, Junda Wang, Hong Yu, and Amir Houmansadr. 2024. Data Extraction Attacks in Retrieval-Augmented Generation via Backdoors. *arXiv preprint arXiv:2411.01705* (2024).
- [116] Mike Perkins. 2023. Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice* 20, 2 (2023).
- [117] Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. 2023. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273* (2023).
- [118] Nicholas Pipitone and Ghita Houir Alami. 2024. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. *arXiv:2408.10343* [cs.AI]
- [119] Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, et al. 2020. Regularizing black-box models for improved interpretability. *Advances in Neural Information Processing Systems* (2020).
- [120] Sayantan Polley. 2022. Towards Explainable Search in Legal Text. In *ECIR*. Springer, 528–536.
- [121] Lip Yee Por, KokSheik Wong, and Kok Onn Chee. 2012. UniSpaCh: A text-based data hiding method using Unicode space characters. *Journal of Systems and Software* 85, 5 (2012), 1075–1082.
- [122] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. *arXiv preprint arXiv:1904.07531* (2019).

- [123] Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal Language Modeling. [arXiv:2306.10193 \[cs.CL\]](#)
- [124] P Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. [arXiv preprint arXiv:1606.05250](#).
- [125] Navid Rekasaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In [SIGIR](#).
- [126] Navid Rekasaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias?. In [SIGIR](#).
- [127] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In [ACM SIGKDD](#).
- [128] Stefano Giovanni Rizzo, Flavio Bertini, and Danilo Montesi. 2016. Content-preserving text watermarking through unicode homoglyph substitution. In [IDEAS](#). 97–104.
- [129] Marko Robnik-Šikonja and Marko Bohanec. 2018. Perturbation-based explanations of prediction models. [Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent](#) (2018), 159–175.
- [130] Joel Rorseth, Parke Godfrey, Lukasz Golab, Divesh Srivastava, and Jaroslav Szlichta. 2024. RAGE Against the Machine: Retrieval-Augmented LLM Explanations. [arXiv preprint arXiv:2405.13000](#) (2024).
- [131] Pouria Rouzrokhi, Shahriar Faghani, Cooper U. Gamble, Moein Shariatnia, and Bradley J. Erickson. 2024. CONFLARE: CONFormal Large language model REtrieval. [arXiv preprint arXiv:2404.04287](#) (2024).
- [132] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. [arXiv preprint arXiv:2304.11406](#) (2023).
- [133] Ryoma Sato, Yuki Takezawa, Han Bao, Kenta Niwa, and Makoto Yamada. 2023. Embarrassingly simple text watermarks. [arXiv preprint arXiv:2310.08920](#) (2023).
- [134] Johannes Schneider. 2024. Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda. [Artificial Intelligence Review](#) 57, 11 (2024), 289.
- [135] Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U. Rajendra Acharya. 2023. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). [Computers in Biology and Medicine](#) 165 (2023), 107441.
- [136] Glenn Shafer and Vladimir Vovk. 2007. A tutorial on conformal prediction. [arXiv:0706.3188 \[cs.LG\]](#)
- [137] Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks. [arXiv:2310.10844 \[cs.CL\]](#)
- [138] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "Nice try, kiddo": Investigating Ad Hominems in Dialogue Responses. In [Proceedings of the NAACL 2021](#). Association for Computational Linguistics, Online.
- [139] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal Biases in Language Generation: Progress and Challenges. In [Proceedings of the 59th ACL 2021](#). Association for Computational Linguistics, Online.
- [140] Pengfei He Shenglai Zeng, Jiankun Zhang et al. 2024. The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG). [ACL](#) (2024).
- [141] Robik Shrestha, Yang Zou, Qiuyu Chen, Zhiheng Li, Yusheng Xie, and Siqi Deng. 2024. FairRAG: Fair human generation via fair retrieval augmentation. In [Computer Vision and Pattern Recognition](#).
- [142] Jaspreet Singh, Megha Khosla, Wang Zhenye, and Avishek Anand. 2021. Extracting per query valid explanations for blackbox learning-to-rank models. In [SIGIR](#). 203–210.
- [143] K. Singhal, S. Azizi, T. Tu, et al. 2023. Large language models encode clinical knowledge. [Nature](#) 620 (2023).
- [144] Karan Singhal, Tao Tu, Juraj Gottweis, et al. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. [arXiv:2305.09617 \[cs.CL\]](#)
- [145] Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. API Is Enough: Conformal Prediction for Large Language Models Without Logit-Access. [arXiv:2403.01216 \[cs.CL\]](#)
- [146] Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation. In [SIGIR](#). 2776–2780.
- [147] Ryuichi Sumida, Koji Inoue, and Tatsuya Kawahara. 2024. Should RAG Chatbots Forget Unimportant Conversations? Exploring Importance and Forgetting with Psychological Insights. [arXiv:2409.12524 \[cs.CL\]](#)
- [148] Jiashuo Sun, Chengjin Xu, Luminyuan Tang, Saizhuo Wang, Chen Lin, et al. 2023. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model with Knowledge Graph. In [ICLR](#).
- [149] Zhensu Sun, Xiaoning Du, Fu Song, and Li Li. 2023. Codemark: Imperceptible watermarking for code datasets against neural code completion models. In [ESEC/FSE](#). 1561–1572.
- [150] Zhensu Sun, Xiaoning Du, Fu Song, Mingze Ni, and Li Li. 2022. Coprotector: Protect open-source code against unauthorized training usage with data poisoning. In [Proceedings of the ACM Web Conference 2022](#). 652–660.
- [151] Elham Tabassi. 2022. Trustworthy AI: Managing the Risks of Artificial Intelligence. <https://www.nist.gov/speech-testimony/trustworthy-ai-managing-risks-artificial-intelligence> Accessed: 2024-07-01.
- [152] Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. [arXiv preprint arXiv:1803.06643](#) (2018).

111:34

B. Ni, Z. Liu, Y. Lei, L. Wang et al.

- [153] Sule Tekkesinoglu and Lars Kunze. 2024. From Feature Importance to Natural Language Explanations Using LLMs with RAG. *arXiv preprint arXiv:2407.20990* (2024).
- [154] Maung Thway, Jose Recatala-Gomez, Fun Siong Lim, Kedar Hippalgaonkar, and Leonard W. T. Ng. 2023. Battling Botpoop using GenAI for Higher Education: A Study of a Retrieval Augmented Generation Chatbots Impact on Learning. *arXiv preprint arXiv:2312.10997* (2023).
- [155] Mercan Topkara, Umut Topkara, and Mikhail J Atallah. 2006. Words are not enough: sentence level natural language watermarking. In *Proceedings of the 4th ACM international workshop on Contents protection and security*. 37–46.
- [156] Umut Topkara, Mercan Topkara, and Mikhail J Atallah. 2006. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *MM&Sec*. 164–174.
- [157] Hugo Touvron et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971* [cs.CL]
- [158] Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, et al. 2022. Plex: Towards Reliability using Pretrained Large Model Extensions. *arXiv preprint arXiv:2207.07411* (2022).
- [159] Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. 2024. WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models. In *ACL*. 1517–1542.
- [160] Manisha Verma and Debasis Ganguly. 2019. LIRME: locally interpretable ranking model explanation. In *SIGIR*.
- [161] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (2021), 89–106.
- [162] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards axiomatic explanations for neural ranking models. In *ICTIR*. 13–22.
- [163] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. 2022. REVISE: A tool for measuring and mitigating bias in visual datasets. *IJCV* (2022).
- [164] Boxin Wang, Weixin Chen, Hengzhi Pei, et al. 2023. DecodingTrust: a comprehensive assessment of trustworthiness in GPT models. In *NeurIPS*. Article 1361, 108 pages.
- [165] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large Language Models for Education: A Survey and Outlook. *arXiv:2403.18105* [cs.CL]
- [166] Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference*.
- [167] Yuan Wang, Xuyang Wu, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2024. Do Large Language Models Rank Fairly? An Empirical Study on the Fairness of LLMs as Rankers. *arXiv preprint arXiv:2404.03192* (2024).
- [168] Ziqiu Wang, Jun Liu, Shengkai Zhang, and Yang Yang. 2024. Poisoned LangChain: Jailbreak LLMs by LangChain. *arXiv:2406.18122* [cs.CL]
- [169] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: how does LLM safety training fail?. In *NeurIPS*.
- [170] Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, et al. 2023. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. *arXiv preprint arXiv:2312.10997* (2023).
- [171] Caesar Wu, Yuan-Fang Lib, and Pascal Bouvry. 2023. Survey of Trustworthy AI: A Meta Decision of AI. *arXiv:2306.00380* [cs.AI]
- [172] Xuyang Wu, Shuowei Li, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2024. Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems. *arXiv preprint arXiv:2409.19804* (2024).
- [173] Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, et al. [n. d.]. Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration. In *EMNLP*.
- [174] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. *arXiv:2402.13178* [cs.CL]
- [175] Hengyuan Xu, Liyao Xiang, Xingjun Ma, Borui Yang, and Baochun Li. 2024. Hufu: A Modality-Agnostic Watermarking System for Pre-Trained Transformers via Permutation Equivariance. *arXiv preprint arXiv:2403.05842* (2024).
- [176] Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2024. A Theory for Token-Level Harmonization in Retrieval-Augmented Generation. *arXiv:2406.00944* [cs.CL]
- [177] Xiaojun Xu, Yuanshun Yao, and Yang Liu. 2024. Learning to Watermark LLM-generated Text via Reinforcement Learning. *arXiv:2403.10553* [cs.LG]
- [178] Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. 2024. BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models. *arXiv preprint arXiv:2406.00083* (2024).
- [179] Jiaqi Xue, Mengxin Zheng, Ting Hua, Yilin Shen, Yepeng Liu, Ladislau Bölöni, and Qian Lou. 2024. TrojLLM: A black-box Trojan prompt attack on large language models. *NeurIPS* 36 (2024).
- [180] Jun Yan, Vikas Yadav, Shiyang Li, et al. 2023. Backdooring instruction-tuned large language models with virtual prompt injection. *arXiv preprint arXiv:2307.16888* (2023).

- [181] Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, et al. 2023. Prca: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. *arXiv preprint arXiv:2310.18347*.
- [182] R. Yang, Y. Ning, E. Keppo, et al. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems* 2 (2025), 2.
- [183] Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. 2022. Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI Conference*, Vol. 36. 11613–11621.
- [184] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing* 4, 2 (June 2024), 100211.
- [185] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking LLMs via Uncertainty Quantification. *arXiv:2401.12794 [cs.CL]*
- [186] Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In *ACL*. 201–206.
- [187] KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust Multi-bit Natural Language Watermarking through Invariant Features. In *Proceedings of the 61st ACL (Volume 1: Long Papers)*. Toronto, Canada.
- [188] KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2024. Advancing Beyond Identification: Multi-bit Watermark for Large Language Models. In *NAACL*.
- [189] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. *arXiv:2310.01558 [cs.CL]*
- [190] Cyril Zakka, Akash Chaurasia, Rohan Shad, et al. 2023. Almanac: Retrieval-Augmented Language Models for Clinical Medicine. *arXiv preprint arXiv:2312.10997* (2023).
- [191] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? *ACL 2019* (2019). *arXiv:1905.07830 [cs.CL]*
- [192] Shenglai Zeng, Jiankun Zhang, Pengfei He, Jie Ren, Tianqi Zheng, et al. 2024. Mitigating the Privacy Issues in Retrieval-Augmented Generation (RAG) via Pure Synthetic Data. *arXiv preprint arXiv:2406.14773* (2024).
- [193] Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A Survey for Efficient Open Domain Question Answering. In *ACL*.
- [194] Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. 2024. REMARK-LLM: A robust and efficient watermarking framework for generative large language models. In *USENIX Security*. 1813–1830.
- [195] Yongfeng Zhang, Xu Chen, et al. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.
- [196] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, et al. 2023. Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2305.14551* (2023).
- [197] Yongfeng Zhang, Jiaxin Mao, and Qingyao Ai. 2019. Tutorial on explainable recommendation and search. In *WWW*.
- [198] Yongfeng Zhang, Yi Zhang, and Min Zhang. 2018. SIGIR 2018 workshop on explainable recommendation and search (EARS 2018). In *The 41st International ACM SIGIR Conference*. 1411–1413.
- [199] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and predict, and then predict again. In *Proceedings of the 14th ACM international conference on web search and data mining*. 418–426.
- [200] Zongmeng Zhang, Yufeng Shi, Jinhua Zhu, Wengang Zhou, Xiang Qi, Peng Zhang, and Houqiang Li. 2024. Trustworthy alignment of retrieval-augmented large language models via reinforcement learning. In *ICML*.
- [201] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Trans. Inf. Syst.* 42, 4, Article 89 (Feb. 2024), 60 pages.
- [202] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning Retrieval Corpora by Injecting Adversarial Passages. *arXiv:2310.19156 [cs.CL]*
- [203] Yujia Zhou, Yan Liu, et al. 2024. Trustworthiness in Retrieval-Augmented Generation Systems: A Survey. *arXiv preprint arXiv:2409.10102* (2024).
- [204] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).
- [205] Wei Zou, Runkeng Geng, Binghui Wang, and Jinyuan Jia. 2024. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. *arXiv:2402.07867 [cs.CR]*

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009