arXiv:2502.06872v1 [cs.CL] 8 Feb 2025

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

Bo Ni[1], **Zheyuan Liu**[†2], **Leyao Wang**[†1] **Yongjia Lei**[†3], **Yuying Zhao**[1], **Xueqi Cheng**[1], **Qingkai Zeng**[2], **Luna Dong**[4], **Yinglong Xia**[4], **Krishnaram Kenthapadi**[5], **Ryan Rossi**[6], **Franck Dernoncourt**[6], **Md Mehrab Tanjim**[6], **Nesreen Ahmed**[7], **Xiaorui Liu**[8], **Wenqi Fan**[9], **Erik Blasch**[10], **Yu Wang**[*3], **Meng Jiang**[*2], **Tyler Derr**[*1]

[1]Vanderbilt University, [2]University of Notre Dame, [3]University of Oregon, [4]Meta, [5]Oracle Health AI, [6]Adobe Research, [7]Cisco AI Research, [8]North Carolina State University, [9]The Hong Kong Polytechnic University, [10]Air Force Research Lab

{bo.ni, leyao.wang, yuying.zhao, xueqi.cheng, tyler.derr}@vanderbilt.edu, {zliu29, qzeng, mjiang2}@nd.edu, {yongjia, yuwang}@uoregon.edu, {lunadong, yxia}@meta.com, krishnaram.kenthapadi@oracle.com, {ryrossi,dernonco,tanjim}@adobe.com, nesahmed@cisco.com, xliu96@ncsu.edu, wenqi.fan@polyu.edu.hk, erik.blasch.1@us.af.mil

Abstract

Retrieval-Augmented Generation (RAG) is an advanced technique designed to address the challenges of Artificial Intelligence-Generated Content (AIGC). By integrating context retrieval into content generation, RAG provides reliable and up-to-date external knowledge, reduces hallucinations, and ensures relevant context across a wide range of tasks. However, despite RAG's success and potential, recent studies have shown that the RAG paradigm also introduces new risks, including robustness issues, privacy concerns, adversarial attacks, and accountability issues. Addressing these risks is critical for future applications of RAG systems, as they directly impact their trustworthiness. Although various methods have been developed to improve the trustworthiness of RAG methods, there is a lack of a unified perspective and framework for research in this topic. Thus, in this paper, we aim to address this gap by providing a comprehensive roadmap for developing trustworthy RAG systems. We place our discussion around five key perspectives: reliability, privacy, safety, fairness, explainability, and accountability. For each perspective, we present a general framework and taxonomy, offering a structured approach to understanding the current challenges, evaluating existing solutions, and identifying promising future research directions. To encourage broader adoption and innovation, we also highlight the downstream applications where trustworthy RAG systems have a significant impact. For more information about the survey, please check our GitHub repository[*].

[†]Significant Contribution.
[*]Corresponding Authors.
[*]https://github.com/Arstanley/Awesome-Trustworthy-Retrieval-Augmented-Generation

Preprint. Under review.

Contents

1 Introduction **3**

2 Preliminaries **4**
2.1 Retrieval Augmented Generation 4
2.2 Tasks and Evaluations 5
2.3 Motivation 6
2.4 Paper Collection 8
2.5 Notes on the Organization of the Survey 8

3 Reliability of Retrieval Augmented Generation **8**
3.1 Taxonomy of RAG Reliability 8
3.2 Uncertainty 8
3.3 Uncertainty Evaluation 10
3.4 Robust Generalization 10
3.5 Robustness Evaluation 11
3.6 Future Directions of RAG Reliability 11

4 Privacy of Retrieval Augmented Generation **12**
4.1 Taxonomy of RAG Privacy 12
4.2 Data Leakage From the External Retrieval Database 13
4.3 Data Leakage From the LLM Training Data 13
4.4 Defense on Privacy Attacks 14
4.5 Privacy Evaluation 14
4.6 Future Directions of RAG Privacy 14

5 Safety of Retrieval Augmented Generation **15**
5.1 Taxonomy of RAG Safety 15
5.2 Methods of RAG Safety 16
5.3 Safety Evaluation 17
5.4 Future Directions of RAG Safety 18

6 Fairness of Retrieval Augmented Generation **18**
6.1 Taxonomy of RAG Fairness 18
6.2 Fairness in Retrieval 18
6.3 Fairness in Generation 19
6.4 Fairness Evaluation 19
6.5 Future Directions of RAG Fairness 20

7 Explainabilty of Retrieval Augmented Generation **20**
7.1 Taxonomy of RAG Explainability 21
7.2 Dual Enhancement of Explanations and RAG 22
7.3 Explainability Evaluation 23
7.4 Future Directions of RAG Explainability 23

8 Accountability of Retrieval Augmented Generation **24**
8.1 Taxonomy of RAG Accountability 24
8.2 Accountability in Retrieval 25
8.3 Accountability in Generation 25
8.4 Accountability in RAG Systems 27
8.5 Accountability Evaluation 28
8.6 Future Direction of RAG Accountability 29

9 Applications **29**
9.1 Healthcare 29
9.2 Law 31
9.3 Education 32

10 Conclusion **33**

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address the challenges faced by Large Language Models (LLMs), such as hallucinations, reliance on outdated knowledge, and the lack of explainability [55, 223]. By incorporating external information into the generation context, RAG improves the accuracy and reliability of the generated content. The recency of information also enables the model to stay current with minimal training costs by reducing the need for extensive retraining of the entire system to update its parameters. These benefits have profound implications for real-world applications. For example, RAG has been effectively applied in medical question answering [195, 216, 162], legal document drafting [190, 132], educational chatbots [172], and financial report summarization [208] due to their adaptability in various domains.

The definition of trustworthiness often depends on the context of discussion [102, 191, 95, 58, 36, 107, 219, 105]. In the context of machine learning and artificial intelligence, trustworthy AI must exhibit characteristics that make the system *worthy of trust*. In 2022, the National Institute of Standards and Technology (NIST) published guidelines for trustworthy AI, defining trustworthiness from several perspectives [169]: Reliability, Privacy, Explainability, Fairness, Accountability, and Safety.

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 robust generalization, which are critical for enhancing system dependability. For instance, in a legal case analysis system, reliability involves balancing uncertainty quantification (e.g., confidence scores for retrieved legal citations and the number of retrieved legal citations) and robust generalization (e.g., applying precedents to new cases) to ensure lawyers are not misled during case preparation.

Privacy focuses on safeguarding user data, ensuring control over personal information. Since RAG has been applied to sensitive domains like the medical field, protecting patient 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.

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.

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, such as those related to gender, race, or socioeconomic status.

Accountability pertains to the governance of AI, including policymaking and law enactment, but also extends to technical aspects such as tracing the origins and processes behind AI-generated content. For example, ensuring that a news-generating system can trace its retrieved sources to improve content accountability and reduce misinformation is critical. Techniques like content watermarking help identify the provenance of retrieved information and the generation process, providing a clear audit trail for future verification.

Safety addresses the system's capacity to prevent and mitigate harm, with a particular focus on defending against adversarial attacks and reducing risks from malicious actors. Current chatbot systems often interact with high-risk users, such as teenagers, who may unknowingly 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, such as adversarial training and ethical guardrails, is crucial for ensuring safety and preventing harm in such interactions.

Despite their recent success, concerns about the trustworthiness of RAG-based systems have become an increasing subject of debate. First, RAG systems are susceptible to reliability issues since developers must ensure the output is accurately grounded on the retrieved content [98, 55]. Second, the reliance on an external database introduces a new attack surface, exposing the systems to a range of adversarial threats [198, 45, 194, 214, 215, 227]. As a result, robustness improvements are needed

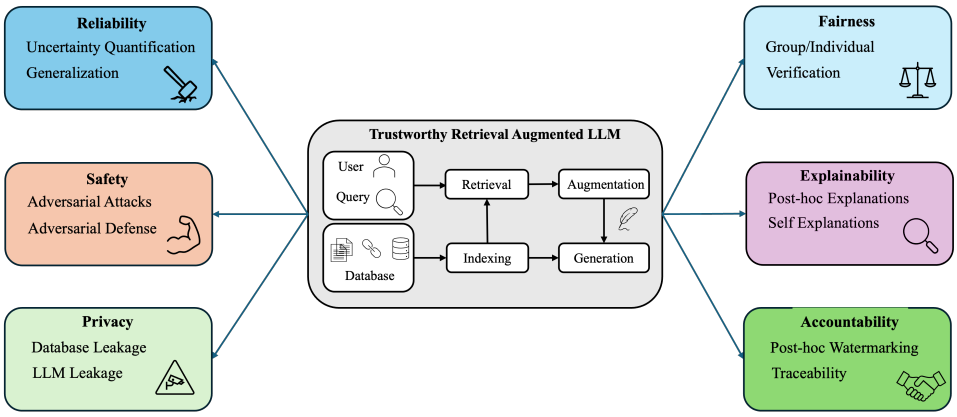

Figure 1: An overview of the key components and dimensions of Trustworthy Retrieval Augmented Generation (RAG) for Large Language Models (LLMs) that are covered in this survey.

to safeguard the systems. Third, RAG systems pose new challenges regarding data privacy [158]. 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. Additionally, RAG could be susceptible to fairness issues [159] from both the retrieving 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 [73]. 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 [195, 216, 190, 172]. Thus, it is essential to incorporate the trustworthy perspective while advancing the RAG systems.

Due to the importance of trustworthy AI, a plethora of research has been developed to advance the application of RAG in Large Language Models with heterogeneous definitions, tremendous implementation, and inconsistent evaluation metrics. 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) [44]. As a simplification of the terminology, in the rest of this survey, we use RAG, RA-LLM, and RAG-based systems interchangeably.

2.1 Retrieval Augmented Generation

As illustrated in Figure **??**, a typical RAG framework consists of three stages: information retrieval, knowledge augmentation, and content generation. Given a query, the retrieval process aims to provide relevant information and context to facilitate the reasoning of the query. Following the classification of previous work [55], the retrieval process contains two stages - *indexing* and *retrieving*. The *indexing* stage takes inputs from a diverse range of formats (PDFs, HTML, words, Markdowns, etc.) and converts them into chunks of data. Subsequently, the chunks of data are converted into vector representations and stored in a vector database for access during inference. However, it is worth noting that for some retrieval-augmented tasks, such as those involving knowledge graphs

[†]https://github.com/Arstanley/Awesome-Trustworthy-Retrieval-Augmented-Generation

(e.g., GraphRAG frameworks), the external databases may not rely on vector databases. Instead, they often use symbolic knowledge structures or relational data to store and retrieve information, bypassing the need for vectorization. The *retrieval* stage takes the incoming query, converts it into vector representation (or aligns it with the symbolic knowledge structures in non-vector cases), and then fetches the matching chunks or nodes from the database. *Knowledge augmentation* involves integrating the retrieved knowledge into the underlying generation models. Often it is achieved through injecting the knowledge into the prompt or finetuning the language model on the retrieved knowledge. Finally, the *generation* stage utilizes the augmented knowledge to produce coherent and contextually accurate responses based on the query.

Compared to the direct generation by LLMs without any contextual information retrieved externally, the additional consideration of external knowledge in RAG paradigm leads to new challenges in trustworthiness. Each stage - retrieval, augmentation, and generation - introduces unique trustworthy challenges, such as retrieval bias, hallucination, and the injection of irrelevant information, all of which require careful mitigation to ensure a trusted response. These concerns will be further explored in subsequent sections of this survey.

2.2 Tasks and Evaluations

2.2.1 Tasks

The RAG paradigm has enabled multiple applications in the natural language processing (NLP) domains. The following briefly introduces some of the common tasks including question answering and chatbots, while also introducing some of the associated and commonly used datasets.

Question Answering. One of the primary downstream tasks for RAG-based language models is question answering (QA), which includes various sub-tasks such as long-form question answering, multi-hop question answering, domain-specific question answering, and open-domain question answering. In these tasks, the system is given a user query and aims to generate the most relevant and accurate answer. The RAG paradigm assists these tasks by integrating the relevant external context into the generation.

Different QA sub-tasks might require different evaluation paradigms. For example, for multi-hop and multiple choice question answering, since the answer is relatively structured or limited to a pre-defined set of answers, the performance is often evaluated through metrics such as hits@n, where n represents the rank of the correct answer in the list of retrieved candidates, and F1, where precision and recall are balanced to assess both the correctness of the generated answers and the system's ability to retrieve all relevant answers [114, 166]. On the other hand, for open domain question answering where answers are more unstructured, Exact Match (EM) and Lexical Match are commonly used [220]. These metrics both provide a good measure of quality of the generated answers. It is also worth mentioning that for some works that emphasize the retrieval process also report results that measures the quality of the retrieval. Commonly used QA datasets include the following:

- **MMLU** [68]: 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 [207, 88].
- **TriviaQA** [79]: 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, making it suitable for testing reading comprehension and retrieval-based models. It is used by most open-domain question answering models [163, 137, 98].
- **WebQSP** [209]: 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 [124, 114, 166].

Chatbots. Another common application for RAG-based language models is Chatbots. Chatbots are designed to handle an array of dialogue types, ranging from task-oriented interactions to open-domain conversations. The RAG paradigm can enhance Chatbots performance by integrating external knowledge into the conversation, allowing the system to access up-to-date information that would otherwise be out of its parameters trained from outdated text corpus [6, 165, 46]. Contemporary knowledge will be especially important for those Chatbots operating in dynamic or domain-specific environments, such as medical and financial dialogue systems.

When evaluating the performance of a chatbot, two primary goals are considered. The first goal is to assess the quality and coherence of the dialogue. To achieve this, various metrics have been proposed, including those that measure utility [24], response understanding [211], and overall aesthetics [188]. Additionally, some metrics evaluate the similarity between the generated responses and human responses, as seen in works such as [3, 197]. While these metrics are generally effective for open-domain dialogues, they often fail to account for specific use cases. Thus, the second goal focuses on evaluating the chatbot's effectiveness in meeting users' needs, particularly in real-world, business-oriented use cases. In these scenarios, beyond the aforementioned dialogue quality metrics, practical effectiveness is a key consideration. General natural language generation (NLG) metrics, such as BLEU [127] and ROUGE [100], are frequently employed to measure this aspect.

There is currently no standardized holistic evaluation of chatbots. For large commercial models, the performance of the chatbots is evaluated on the subtasks such as code generation, problem solving, and complex reasoning. On the other hand, for the business-oriented, task specific models, they are evaluated on the corresponding task-specific or industry-specific datasets. We introduce some of the commonly used datasets below.

- **HellaSwag** [217]: A commonly used dataset for commonsense reasoning and story completion tasks. HellaSwag presents models with scenarios requiring contextually appropriate completions, testing their ability to reason beyond surface-level semantics. It has been widely adopted for benchmarking commonsense reasoning capabilities in large language models [42, 23].
- **HumanEval** [26]: A widely used dataset for evaluating code generation capabilities of language models. HumanEval includes programming problems of varying difficulty levels, along with unit tests to validate the correctness of generated code. It is a standard benchmark for assessing the coding performance of generative models [42].
- **MedicationQA** [17]: 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 [42, 92].

Others. Beyond language-based tasks, RAG-based language models can be applied to a diverse range of downstream tasks, including recommendation systems [97], software engineering [72], and AI for scientific discovery [5]. 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. To comprehensively assess trustworthiness, additional metrics are necessary, including those that evaluate bias, fairness, and reliability in the generated answers. As AI trust evaluation has well been debated across cognitive, communication, information, and social dimensions, the focus was on the use of the result. To maintain trust, a trust worthy evaluation is also needed to determine the efficacy under changing conditions. These metrics will be designed to ensure that the system aligns well with trustworthy aspects. Due to their heterogeneous nature, we will delve into these specific metrics in detail in the corresponding sections throughout the rest of the survey.

2.3 Motivation

Although trustworthiness in deep learning and LLMs has been well-explored in the general AI community, it is rather critical to consider the trustworthiness for RAG-based LLMs because (a) their growing usage in real-world applications often involves high-stakes decision-making, where errors or biases can lead to significant consequences; (b) RAG models are vulnerable to trustworthiness

Table 1: Comparison with Existing Surveys on RAG and Trustworthy LLMs. (We mark some of the items ✓ for LLM related survey as they do not focus on RAG)

	Surveys	Pillars of Trustworthiness											
		Reliability		Privacy		Safety		Fairness		Explainability		Accountability	
		Uncertainty	Generalizability	External	Training Data	Jailbreaking	Defense	Retrieval	Generation	Retrieval	Generation	Retrieval	Generation
LLM	[105]	✓	✓	✗	✓	✓	✓	✗	✓	✗	✓	✗	✓
	[74]	✗	✗	✗	✓	✓	✓	✗	✓	✗	✓	✗	✓
RAG	[44]	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
	[55]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗
	[230]	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗
	Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

issues as the multiple stages in RAG—such as indexing, retrieval, and generation—can introduce compounding errors, biases, or hallucinations, which can be difficult to trace and mitigate; and (c) RAG-based LLMs are inherently different from traditional standalone LLMs as they rely on external data sources during the retrieval process, leading to a dynamic interaction between the model and potentially unreliable or biased external information. Thus, directly extending the trustworthiness framework for standalone LLMs to RAG-based LLMs is challenging due to evolving complexities. Therefore, developing a comprehensive and robust trustworthiness framework specifically tailored for RAG-based LLMs is imperative to ensure reliable and safe deployment in various domains.

While prior work [158, 200, 194, 40, 230] has touched on these issues individually, there remains a gap in the literature for a unified survey that reviews the current advancements and challenges in ensuring trustworthiness for RAG-based LLMs. This survey aims to be a comprehensive review by summarizing existing efforts, categorizing approaches, and identifying opportunities for future research. By providing a systematic review, we hope to facilitate further research and development of trustworthy RAG-based LLMs across diverse domains.

2.3.1 Related Surveys and Differences

As shown in Table **??**, recently, several surveys have been conducted on RAG [55, 44] and trustworthy LLMs [105, 74]. On one hand, although the RAG surveys provide a comprehensive overview of the state-of-the-art models and architectures, detailing some of the methods' efforts to address challenges in the realm of trustworthiness (e.g., robustness), few of them provide a comprehensive and focused discussion on the specific challenges and solutions related to trustworthiness across the entire RAG pipeline, particularly in areas such as reliability, fairness, privacy, and safety [55, 44]. On the other hand, the surveys on LLM trustworthiness focus mostly on the generation aspects of LLMs, such as mitigating hallucinations or enhancing explainability, and cannot be directly applied to RAG models due to the added complexity introduced by the retrieval and augmentation stages [105].

Most recently, there is a survey dedicated to trustworthiness in RAG systems [230]. Although the survey provides reviews on existing works, most of its content focuses on experiments related to trustworthy generation across different language models within a basic retrieval-augmented generation setup. We acknowledge the importance of such empirical studies in advancing the field but would like to emphasize the different focus of our survey. Rather than empirical studies, our survey centers on providing a more detailed and comprehensive literature review. We systematically categorize trustworthiness challenges and solutions across the RAG applications. By focusing on these broader aspects, our work aims to establish a unifying framework to guide future research and development in trustworthy RAG systems.

Nonetheless, all of the existing surveys have acknowledged the importance of trustworthiness and included it as a critical area for future research directions [55, 44], demonstrating the aligned interests in the community.

2.4 Paper Collection

To construct a comprehensive survey of trustworthy RAG systems in LLMs, we follow a systematic literature review methodology. We began by identifying relevant papers through keyword searches across major academic databases, including Google Scholar, ACM Digital Library, and arXiv, using general terms such as "Retrieval-Augmented Generation," "trustworthiness," as well as specific terms for the different aspects. Papers were included if they directly discussed the trustworthiness aspects of RAG systems, such as reliability, privacy, safety, fairness, and accountability. Because of the modular nature of RAG-based systems where external databases are coupled with a language model, we also take works that discuss each into consideration. After an initial screening, we categorize the papers into each aspect of trustworthiness, situating them into corresponding taxonomies. Our final collection of papers, as discussed in this survey, reflects a diverse range of perspectives on RAG and trustworthiness. The cut off date for the papers to be included in this survey is October 2024.

2.5 Notes on the Organization of the Survey

As outlined in the previous sections, the remainder of this survey is structured around the six key aspects of trustworthiness: Reliability, Privacy, Safety, Fairness, Explainability, and Accountability. Given the distinct nature of each aspect, every section will follow its own taxonomy, introducing relevant works accordingly. To further emphasize these differences, we include specific future directions and evaluation protocols within each section. A general discussion of the future directions for trustworthy Retrieval-Augmented Generation (RAG) systems will follow at the conclusion, providing a broader discussion that encompasses and integrates the section-specific insights.

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 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. Just as other deep learning models that train on the big data, RAG models are susceptible to common pitfalls of reliability. Previous work [176] defines reliability from three granular 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. However, 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.

For Retrieval-Augmented Generation (RAG) systems, uncertainty quantification presents two primary challenges: First, during the generation phase, uncertainty stems from the inherent limitations of large language models (LLMs). Standard techniques for quantifying uncertainty in LLMs, such as conformal prediction, can be applied here with few adaptations [207]. 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.

Table 2: Taxonomy for Uncertainty Quantification in RAG

MODULE	REFERENCE	WHITE-BOX	TASK	YEAR
Generation	Ye et al. [207]	✗	Benchmarking	2024
	Su et al. [163]	✓	Open Domain Question Answering	2024
	Kumar et al. [88]	✓	Multiple Choice Question Answering	2023
	Quach et al. [137]	✗	Open Domain Question Answering	2023
Retrieval + Generation	Ni et al. [124]	✗	Multi-hop Question Answering	2024
	Li et al. [98]	✗	Open Domain Question Answering	2023

3.2.1 Uncertainty Quantification in Generation

The generation phase in Retrieval-Augmented Generation (RAG) systems 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 [154]. 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. [88] applied standard CP to the Llama model [175] by leveraging softmax scores of token logits in multiple-choice tasks. Similarly, Ye et al. [207] extended logit-based approaches to multiple baselines and language models.

To compensate 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. [137] adapted the learn-then-test risk-control framework [11] for LLMs, approximating the non-conformity score through sampling, which allows uncertainty quantification in black-box models without direct logit access. Su et al. [163] 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 **??**, 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 [98, 146]. 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 [146]. Specifically, this framework uses conformal prediction to quantify retrieval 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 - α) confidence level.

Moreover, TRAQ [98] 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 [124] 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* [67]. 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, $\mathcal{A}_{\text{ret}}$, 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, it is defined as:

$$C = \frac{N_{\text{correct}}}{N_{\text{total}}},$$

where N_{correct} represents the number of times the correct answer $\mathcal{A}_{\text{true}}$ is contained in the returned set $\mathcal{A}_{\text{ret}}$, $\mathcal{A}_{\text{true}} \subseteq \mathcal{A}_{\text{ret}}$, and N_{total} is the total number of instances.

For the model to be considered reliable, C should be at least $1 - \alpha$. However, simply exceeding this 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.

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):

$$E = |\mathcal{A}_{\text{ret}}|.$$

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 [176] 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 or misinformation 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 in Section 5 for Adversarial Robustness. This section will focus on the context that occurs *organically* over time.

3.4.1 Irrelevant Context

Fang et al. [45] 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. [214] further explores the negative impact of the retrieval of irrelevant context on the model performance. They argue that the negative impact of the irrlevant 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. Finetuning on this additional context allows the model to learn to differentiate between useful and distracting information and minimizing the negative effect of the irrelevant context. The experiment result on five different 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. [198] 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, the author further proposed an interactive inference framework X-RAG that leverages the benefit of both worlds of 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. Exsting metrics outlined in 2.2.1 will be applied to assess the model's performance. It is important to note that 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 [45, 198], while others report only the performance delta between the proposed fine-tuned model and the corresponding baseline for better visualization [214].

Datasets. Currently, there is no widely-used benchmark for RAG robustness. To simulate real-world conditions and 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 [198, 214]. 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. [45] 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, as well as the golden retrieval data. *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

Reliability remains an important challenge in the development of trustworthy RAG systems. While our current sections are structured independently for uncertainty and robustness, future research should aim for a more integrated approach that captures the intricate interactions between these aspects. These concepts are not necessarily exclusive; uncertainty quantification can help the model produce more trustworthy results when faced with less accurate or noisy contexts, while better robustness can reduce the model's overall uncertainty.

Table 3: Taxonomy for RAG Privacy

	Reference	Training	Tasks	Leakage	Year
Attack	Zeng et al. [158]	✓	Document Extraction & Training Data	Internal & External	2024
	Liu et al. [103]	✓	Membership Inference Attack	External	2024
	Cohen et al. [31]	✗	MIA & Document Extraction	External	2024
	Jiang et al. [77]	✗	Document Extraction	Internal & External	2024
	Peng et al. [129]	✓	Document Extraction	External	2024
Defense	Zeng et al. [218]	✓	External Database	External	2024
	Zeng et al. [158]	✓	Document Extraction	External	2024

An integrated framework addressing both uncertainty and robustness could enhance the adaptability and robustness of RAG systems, particularly in complex real-world applications where the boundaries between irrelevant and corrupted contexts are blurred. By combining strategies from uncertainty quantification with robust robustness techniques, future research can explore models that are not only resilient to noise but also capable of dynamically adjusting their confidence levels based on the context quality.

Moreover, future efforts should extend current evaluation benchmarks to include more comprehensive benchmarks that reflect real-world conditions. Current benchmarks such as RAG-Bench [45] mark a good effort, but the data are manually filtered by rules, which limits the scope of the challenges they present. Real-world scenarios are more complex, with a greater variety of noise. By incorporating these complexities, new benchmarks could better simulate the unpredictable nature of real-world applications. With better benchmarks, researchers will be able to test models under various reliability scenarios where uncertainty and robustness challenges are intertwined. Expanding evaluation methodologies in this direction would push the frontier of research in reliable RAG systems and offer more actionable insights for practical deployments.

Finally, drawing from interdisciplinary fields like dynamic knowledge graphs and active learning could lead to the development of more adaptive RAG systems. In dynamically evolving environments, new knowledge introduced in external databases may create inconsistencies or conflicts with previously retrieved information, posing significant reliability challenges. These interactions would require RAG systems to not only generalize well but also quantify uncertainty effectively when knowledge shifts or drifts occur. Such advances would make RAG systems more robust and practical, improving their performance in dynamic, real-world scenarios.

4 Privacy of Retrieval Augmented Generation

Although privacy risks in LLMs have been extensively studied, RAG systems introduce additional complexities by leveraging external data. This integration poses new challenges in maintaining privacy, ensuring data integrity, and managing the overall trustworthiness of the RAG system. In this section, we will introduce the threat model concerning privacy leaks in RAG systems. We will then discuss the current efforts to address these challenges. Towards the end, we will explore potential future directions for enhancing the trustworthiness and privacy of RAG systems.

4.1 Taxonomy of RAG Privacy

In Table 3, we outline the existing efforts in addressing the privacy issues present in the RAG systems. We will briefly introduce the relevant taxonomy in the following.

4.1.1 Training

Training refers to whether the attack or defense requires *prior* training on the data. Models that require training typically assume a distinct threat model compared to those that do not. When a model requires training, it often presumes white-box access to either the retriever or the language model, allowing attackers or defenders to fine-tune or adjust components of the RAG system to exploit or mitigate vulnerabilities. On the other hand, models that do not require training typically rely on prompt-based methods or zero-shot techniques. This distinction has significant implications for the feasibility of privacy attacks and defenses in RAG systems.

4.1.2 Tasks

Admittedly, research on RAG privacy is still in its infancy. Current literature focuses on three main tasks: *Document Extraction*, *Training Data Extraction*, and *Membership Inference Attack*. *Document Extraction* seeks to extract confidential information from the retrieval database, such as Personally Identifiable Information (PII). *Membership Inference Attack* aims to determine whether specific passages are present in the retrieval database. While not a direct privacy attack, it introduces risks by exposing sensitive associations between queries and database contents, potentially enabling adversaries to infer private information. Lastly, *Training Data Extraction* examines the leakage of LLM training data in the context of retrieval-augmented generation, highlighting vulnerabilities that could lead to the unauthorized exposure of proprietary or sensitive datasets.

4.1.3 Leakage

We consider two sources of leakages. First, the leakage of the external retrieval database involves leaking targeted/untargeted information from external knowledge sources, such as sensitive data in proprietary databases or publicly available but privacy-relevant information inadvertently retrieved during query processing. Second, the leakage of the internal training data focuses on the exposure of the LLM training data in the context of retrieval-augmented generation. This occurs when the language model unintentionally reproduces sensitive or proprietary information from its training dataset during response generation, raising concerns about policy violations and privacy breaches. We organize the rest of the section from the above 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, targeting two main objectives: (1) eliciting specific information from the retrieval system with high accuracy, and (2) outputting the retrieved private data. Zeng et al.[158] 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.[77] 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. [129] 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 by employing a language model to paraphrase the retrieved content, thereby increasing the difficulty of detecting the attack. Furthermore, Cohen et al. [31] 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 Retrieval-Augmented Generation (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. [103] introduced a mask-based attack that obscures portions of documents, compelling the language model to predict the masked words. This technique not only reveals sensitive information but also highlights vulnerabilities in the retrieval system's training data.

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. [158] 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. [158] 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 [218]. 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 [158, 218].

Datasets. Similar to the generalization evaluation, there haven't been any established datasets or baselines specifically designed for evaluating privacy in RAG systems. However, current work has leveraged existing datasets to provide initial insights. The Enron Email dataset [86] contains employee emails, which often include sensitive personal information, such as names, contact details, and internal company communications. The HealthcareMagic-101 dataset [218], on the other hand, consists of doctor-patient dialogues, encompassing a variety of medical discussions that include personally identifiable information (PII) and private health information. These datasets serve as a valuable starting point for privacy evaluations due to their realistic, sensitive content that mimics the types of data RAG systems may encounter.

4.6 Future Directions of RAG Privacy

Domain Specific Applications. While the preliminary findings from Zeng et al. [158] shed light on the dual nature of privacy risks in RAG systems, there remains a significant gap in understanding the broader implications of these vulnerabilities across different applications and contexts. Future research should prioritize developing robust, application-specific privacy-preserving techniques tailored to the unique demands of various domains, such as healthcare, finance, and legal services, where the consequences of privacy breaches can be particularly severe.

Privacy Defense. Additionally, exploring the integration of advanced cryptographic methods, such as secure homomorphic encryption, within RAG frameworks could provide new avenues for safeguarding sensitive data during retrieval and generation processes. Another promising direction is the development of differential privacy techniques specifically adapted for RAG systems, aiming to balance privacy preservation with the utility of the generated outputs.

Evaluation. 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. By addressing these challenges, future research can contribute to the creation of more secure and trustworthy RAG systems, ultimately fostering greater confidence in the deployment of these technologies in sensitive and high-stakes environments.

Table 4: Taxonomy for RAG Safety

	Reference	White-Box	Black-Box	Year
Targeted	Zou et al. [232]	✓	✓	2024
	Xue et al. [200]	✓	✗	2024
	Long et al. [110]	✓	✗	2024
	Zhong et al. [229]	✓	✗	2023
Jailbreak	Wang et al. [187]	✗	✓	2024
	Deng et al. [39]	✗	✓	2024

5 Safety of Retrieval Augmented Generation

Recent research has shown that LLMs are vulnerable to a range of adversarial attacks [74, 155, 108, 206]. Through techniques such as prompt engineering, hint manipulation, and input perturbation, attackers can bypass safety mechanisms and exploit model weaknesses, posing significant threats to society. RAG systems, built upon the capabilities of LLMs and integrated with external databases, present distinct challenges related to safety concerns. For example, by letting the RAG system retrieve adversarial information, attackers can circumvent the alignment of LLMs with human integrity and produce malicious content [39]. As more adoption of RAG system emerges, the safety of RAG systems have arisen significant concerns on utilizing RAG in more high-stake applications. For example, in education applications, RAG is often leveraged to retrieve relevant domain-specific educational context (textbooks, problem set, etc.). Thus, these vulnerabilities will pose significant threats to underage minors if the system is adversarially compromised. As a result, developing robust RAG is essential for ensuring trustworthiness in RAG systems. Since this is a relatively new field of research, we will give a brief overview of the adversarial attacks on RAG and then point out potential future directions that are worth investigating.

5.1 Taxonomy of RAG Safety

In Table 4, we summarize existing Retrieval-Augmented Generation (RAG) methods based on an adversarial taxonomy. This section introduces the taxonomy and provides definitions for each category. It is worth noting that the wide range of adversarial attacks on LLMs, such as backdoor attacks [201, 112], jailbreaking attacks [189, 231], and prompt injection attacks [60, 106, 202], technically target the underlying LLM component of RAG. However, how the retrieved context influences the attack surface and defense strategies in RAG systems remains an open question requiring further investigation. Preliminary results [158] suggest that RAG can alleviate the effects of simple prefix attacks, as the retrieval step introduces an additional layer of complexity. Yet, the interaction between the retriever and generator under more complex adversarial conditions, e.g., combined backdoor and retrieval-based attacks, is still poorly understood warranting further study.

Due to the scope of this survey, we do not delve into existing LLM-specific attacks, which are covered extensively in works like [105, 74]. Instead, our focus in this section is on the robustness of the retriever model in RAG systems and how its interaction with the generator model can impact overall system security and resilience to adversarial attacks.

5.1.1 Threat Model

To systematically categorize the robustness of RAG systems, we consider three primary components: the external database, the output generator (the underlying language model), and the context retriever. A realistic threat model assumes that the attacker has no read or delete access to the external database but possesses write access. This mirrors real-world scenarios where users can upload documents to a database but cannot access all of the content. The attacker is also assumed to have no detailed knowledge of the underlying language model. In practice, most commercial language models are proprietary, making them black-box systems. This assumption is crucial for adversarial robustness studies, as the attacker cannot exploit specific weaknesses in the model architecture or training data.

For the retriever, we consider two distinct threat models. For white-box setting, the attacker has complete access to the retriever model. This includes the ability to inspect the model architecture, parameters, and sometimes retriever-specific training data. The attacker can thus craft sophisticated

adversarial examples by exploiting the retriever's known weaknesses. For example, by understanding the tokenization or ranking algorithm, an attacker might introduce documents that are highly ranked by the retriever but irrelevant or misleading for the generation task [232].

In contrast, the black-box setting assumes the attacker has no direct access to the retriever. The adversary can only query the retriever and observe the output (i.e., the ranked retrieved documents) without knowledge of the internal mechanisms. The attacker must infer patterns from these outputs and attempt to manipulate the retriever's behavior through indirect methods, such as poisoning the external database or introducing misleading or noisy entries [39].

5.1.2 Attacker's Goal

The attacker's target, under traditional machine learning contexts, is often categorized into two main types: targeted attacks and untargeted attacks [219, 105]. For generative models such as Retrieval-Augmented Generation (RAG), we identify two corresponding categories: targeted attacks and jailbreak attacks. Targeted attacks are aimed at manipulating the model's output in response to specific inputs, usually focusing on certain questions or topics. The goal is to subtly distort the generated response while keeping the attack as stealthy as possible, making it harder to detect through typical monitoring systems. These methods are often carefully crafted to evade detection mechanisms by introducing minimal disruptions. Such targeted attacks can have significant social implications. For example, by injecting retrieval bias, a targeted attack might skew the model's output to emphasize certain policies or downplay others, inadvertently affecting sensitive areas such as elections.

In the context of generative models, untargeted attacks manifest as jailbreaking attacks, where attackers attempt to bypass content restrictions or safety measures embedded within the model. The goal is to provoke unrestricted or harmful outputs without focusing on a specific topic. This type of attack poses a broad threat as it can force the model to generate inappropriate or unsafe content across a range of inputs, compromising the system's reliability and trustworthiness.

5.2 Methods of RAG Safety

Targeted Attacks. Zou et al. [232] first introduce PoisonedRAG as a novel method to attach the new attack surface brought by the retrieval component. By designing specific passages to be injected into the retrieval database based on specific questions, PoisedRAG is able to mislead the RAG system to generate specific answers desired by the attackers for specific questions. It considers both the white-box and black-box settings. When the attacker has no access to the model parameters (black-box), PoisonedRAG crafts the injected passage with a simple heuristic: passages that is more similar to the question would be more likely to be retrieved. On the other hand, when the attacker has access to the model parameters, the crafted message is constructed by further optimizing the following equation

$$P = \underset{P'}{\text{argmax}} Sim(f(Q), f(P')) \tag{1}$$

where P is the generated passage, Q is the user query, f is the encoder, and Sim is the function that measures the similarity between the encoded passage and question.

One limitation of PoisonedRAG is its focus on specific queries, neglecting broader group-based attacks that target semantically related query categories, such as those involving political affiliations, race, or religion. To address this gap, Xue et al. [200] propose the BadRAG framework, which extends the attack methodology to include group-query targeting. BadRAG allows the trigger to be semantic groups such as political parties or candidates by collecting target triggers from the given topic. For example, for *Republican*, BadRAG collects terms including *Governor*, *Red States*, and *Pro-Life*. To optimize the adversarial passages, BadRAG employs a contrastive learning paradigm, where the triggered queries are positive samples and the normal queries are negative samples. The adversarial passage will then maximize its similarity with the triggered queries and minimize its similarity with the normal ones. It's also worth mentioning that this approach can also support additional attacks, such as Denial of Service (DoS) [200], by aligning adversarial passages with targeted model behaviors.

Dense retrieval has been extensively studied within the Information Retrieval community [228]. Many attacks on dense retrievers, such as adversarial manipulation, are also applicable to passage retrieval tasks. Recently, Long et al.[110] 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.[229] 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, most attack strategies for dense retrieval have been developed without considering downstream generation tasks, leaving their effects on generated outputs unclear. For instance, certain Retrieval-Augmented Generation (RAG) methods equipped with safety guardrails could potentially diminish the impact of adversarially retrieved passages. As we will discuss in the Future Directions section, understanding and addressing the interaction between retrieval attacks and downstream generation presents a significant research opportunity in Trustworthy RAG.

Jailbreak Attacks. When specific attack targets are absent, the threat model shifts toward jailbreak attacks. Wang et al.[187] 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.[39] 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 [200] 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 [232].

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.[39] manually label a generation as a successful attack based on the *relevance* and *quality* of the generated content. Similarly, Yang et al.[187] 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 retrieval-augmented generation (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)[89], MS MARCO[15], and SQuAD [138]. Researchers then select questions based on the specific characteristics of the targeted attacks. For instance, in BadRAG, Xue et al. [200] 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 significantly due to the inherent challenges of assessing open-ended text generation. Current approaches typically rely on manually curated adversarial questions categorized into specific themes. For example, Pandora [39] organizes questions into categories such as *Adult*, *Harmful*, *Privacy*, and *Illegal*, whereas Poisoned-Langchain [187] uses categories like *Dangerous Behaviors*, *Misuse of Chemicals*, and *Illegal Discrimination*. Despite these efforts, the RAG literature lacks standardized datasets and unified evaluation frameworks, highlighting the need for future research to establish comprehensive benchmarks and methodologies.

5.4 Future Directions of RAG Safety

Adversarial Defense. Current defense mechanisms against adversarial attacks in retrieval-augmented generation (RAG) systems are rudimentary. For instance, Xue et al. [200] propose learning the connection between trigger words and adversarial passages through *token masking*. However, there is a notable lack of dedicated research specifically addressing adversarial defense in this domain. Existing approaches are limited in their scope and sophistication, leaving significant room for improvement. Arguably, advancements in generalization research, as discussed in Section 3.2, can be leveraged for adversarial defense. Training RAG models on distributions that include retrieved adversarial data could enable these models to better generalize and potentially mitigate the impact of adversarial passages. However, further exploration is needed to design and implement targeted approaches that address the unique challenges posed by adversarial attacks in RAG systems. The development of dedicated methodologies for adversarial defense remains an essential and largely uncharted area of research.

Modality. Current research predominantly focuses on RAG systems operating over text databases represented as vectors. However, recent advancements have introduced emerging applications that utilize alternative knowledge representations, such as knowledge graphs and databases [124]. The unique structural properties of knowledge graphs present novel opportunities and challenges for both adversarial attacks and defenses. Attack and defense mechanisms tailored to these modalities are necessary to account for their inherent characteristics, such as the interconnectedness of entities and the semantic richness of relationships. Future research should focus on developing methods that address these unique requirements to expand the applicability and robustness of RAG systems across diverse knowledge representations.

Evaluation. As highlighted in preceding sections, the field currently suffers from a lack of 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 metrics, such as robustness, generalization, and performance under adversarial conditions. Such benchmarks would provide a unified basis for assessing advancements in the field and drive progress through consistent and meaningful comparisons. Addressing this gap is critical for fostering innovation and ensuring the reliability of RAG systems in practical applications.

6 Fairness of Retrieval Augmented Generation

As generative models such as LLMs and image-generative models becomes increasingly integrated into real-world applications, it is critical to ensure fairness in their outputs. Retrieval-Augmented Generation (RAG) systems combine generative models with external knowledge retrieval, providing substantial 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 [182, 18, 19], which leads to biased outputs. This further introduces the risk of amplifying disparities in age, gender, race, and other demographic attributes, particularly when the retrieved data is drawn from biased or unregulated sources.

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 Retrieval-Augmented Generation (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

Table 5: Taxonomy for RAG Fairness

Module	Reference	Bias Mitigation	Focus	Year
Retrieval	Wang et al. [186]	LoRA Fine-tuning	Ranking Fairness vs. Performance	2024
	Shrestha et al. [159]	Diverse Sampling	Demographic Diversity	2024
	Rekabsaz et al. [139]	Re-ranking Methods	Societal Bias Mitigation	2021
	Rekabsaz et al. [140]	Post-hoc Re-ranking	Gender Bias in Retrieval	2020
Generation	Wu et al. [192]	Empirical Evaluation	Cross-task Fairness	2024
	Wang et al. [183]	Output Conditioning	GPT-3.5/4 Bias Detection	2023
	Liang et al. [99]	Representation Adjustment	Fair Question Answering	2022
	Parrish et al. [128]	Benchmark Evaluation	Stereotype Analysis	2021
Retrieval + Generation	Kong et al. [87]	Post-hoc Bias Mitigation (PBM)	Gender and Race Fairness	2024
	Kim et al. [82]	Fair Retrieval + Generation	Fairness-Quality Trade-off	2024
	Shrestha et al. [159]	Cross-Modal Guidance	Demographic Balancing	2024

these challenges, several frameworks have been proposed to ensure that the retrieval system itself is fair. Reskabsaz et al. [140] 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. [139] investigates how re-ranking methods can mitigate biases present in initial retrieval results. Then, Wang et al. [186] 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 [159] 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, frameworks such as BadRAG [200] have shown how maliciously inserted or poisoned data in the retrieval corpus can lead to biased and unfair outputs. Kong et al. [87] proposes the Post-hoc Bias Mitigation (PBM) technique, which balances retrieved image sets to ensure more equitable representation across gender and race.

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. [99] assesses the accuracy of question-answering systems while accounting for fairness through measures of toxicity and representation bias. Similarly, Wang et al. [183] identifies the demographic imbalances in models like GPT-3.5 and GPT-4 under both zero-shot and few-shot question-answering settings. FairRAG [159] 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. [128] 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. [192] conducts an empirically evaluation of fairness across various RAG methods. Similarly, Kim et al. [82] 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) [138] and ROUGE-1 scores [100]. For fairness evaluation, the focus is on metrics such as Group Disparity (GD) [52] and Equalized Odds (EO) [65]. 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 [200] 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 [41, 32], 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 [128] includes samples with contexts that are either ambiguous or unambiguous. 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 [149] 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

Personalized Fairness. While existing frameworks like FairRAG [159] and PBM [87] have made notable strides in mitigating biases in Retrieval-Augmented Generation (RAG) systems, they largely overlook the need for personalized fairness mechanisms. Personalized fairness involves tailoring fairness constraints to specific application contexts, such as healthcare or recruitment, where fairness definitions and requirements can vary significantly. For example, in healthcare, equity might require prioritizing underrepresented groups in clinical trials, while in recruitment, it might focus on reducing gender or racial biases in candidate evaluation. Future research should explore adaptive fairness constraints that dynamically align with these domain-specific requirements, ensuring equitable outcomes in high-stakes domains.

Fairness-Accuracy-Relevance Trade-offs. A persistent challenge in advancing fairness in RAG systems is the inherent trade-offs between fairness, relevance, and accuracy. Current approaches, such as those explored by Liang et al. [99] and Wang et al. [183], primarily emphasize fairness in the generation and retrieval phases. However, they lack robust mechanisms to balance these competing objectives. Multi-objective optimization techniques could be pivotal in navigating these trade-offs. By formalizing fairness, relevance, and accuracy as interconnected yet distinct objectives, researchers can develop frameworks that prioritize trade-off management, enabling more holistic and practical solutions for fairness in RAG systems.

Cross-Modal Fairness. The emergence of multimodal RAG systems that process both text and image data introduces unique challenges in fairness. Biases in these systems can manifest differently across modalities, yet current research, such as Wu et al. [192] and Kim et al. [82], has only scratched the surface of cross-modal fairness. Future work should aim to identify and mitigate modality-specific biases while ensuring consistent fairness across both text and image outputs. For example, ensuring that a system's fairness constraints for text descriptions align with fairness principles for corresponding visual elements could significantly enhance trustworthiness. Addressing cross-modal fairness is essential for building robust and equitable multimodal RAG systems.

7 Explainabilty of Retrieval Augmented Generation

Explainability, as a critical perspective of trustworthiness, has attracted significant attention due to its capability to elucidate the behavior of machine learning models and uncover novel data-driven insights that can even inspire domain experts [152, 143, 141, 109, 48, 210]. This demand for explainability has been further intensified with the rise of black-box LLMs [99, 164]. We acknowledge the existing survey [230] on trustworthiness, which includes a section aimed at uncovering the decision-making process of RAG systems called transparency. However, the concept of transparency in that context differs from the explainability we focus on in our work. 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).

In the context of RAG, explainability becomes even more crucial due to the inherent complexity of its multi-component architecture. Beyond explaining the generation process of LLMs, it is also important to understand why the retrieval process prefers specific contexts over others. For example,

Table 6: Taxonomy for RAG Explainability

MODULE	REFERENCE	TASK	YEAR
Generation	Sudhi et al. [164]	English and German QA	2024
	Luo et al. [114]	Knowledge Graph QA	2024
	Rorseth et al. [145]	Open-book QA	2024
Retrieval + Generation	Kunze et al. [171]	Scene-Understanding	2024
	Hussien et al. [75]	Road User Intention Explanation	2024
	Ferraretto et al. [50]	Document Retrieval	2023

what words in the input questions lead to the retrieval of a particular sentence? As some of the RAG systems include post-retrieval [59, 81, 203], it is also important to understand why the retrieved content needs to be post-processed in a specific way to augment the downstream generation. For example, what features of the input result in such ranking/importance scores when performing reranking after retrieval? Given the inherently multi-stage nature of RAG, this section reviews the literature on explaining the two crucial stages of RAG, retrieval and generation process, respectively.

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 [222, 224, 225]. 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 [10], 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 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 [136] uses interpretable textual features to explain rankings, [134] understands the BERT-based ranking models by the attention scores of tokens, and [178] 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. [160] uses a greedy search-based algorithm to obtain a subset of features that serve as the explanations. Axiomatic explainers provide explanations using axioms [180] 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 [30], how sensitive they are to various textual properties [115], and what knowledge they possess [28, 51]. 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 [51, 226].

7.1.2 Explainability in Generation

In addition to explaining the retriever, explaining the generator is equivalently or even more important. First, the output directly stems from this generation process, and any shortcuts or reliance on misleading features can lead to significant generation errors [69, 38]. Second, modern RAG systems typically rely on LLM-based generators, which are prone to hallucinations. This raises concerns about whether the generated output truly focuses on the query-relevant words and the retrieved contents, which can be naturally addressed via explanation [151, 164]. There are two main categories of explanation methods in the generation process of RAG systems, ante-hoc explanation methods and post-hoc explanation methods, respectively.

RAG-Ex [164], as a post-hoc explanation method, introduces a model and language-agnostic framework inspired by the philosophy of perturbation-based explanation. The core idea is to identify critical tokens or features such that removing them would significantly alter the output of the generator. Specifically, RAG-Ex proposes six perturbation methods: leave-one-token-out, random noise, entity manipulation, antonym injection, synonym injection, and order manipulation. After prompting LLMs with perturbed inputs (using the same parameters), the similarity between the generated responses and the original response is measured, which can be further used to calculate the importance scores for the perturbed tokens. After obtaining the importance score for each token, top-K important tokens are selected as the explanation, and they are referred to with the sentence containing the ground-truth answer or question to further assess the explanation quality. Through experiments, RAG-Ex finds that the "leave-one-token-out" perturbation performs the best in terms of accurately identifying the most critical tokens for the explanation. Aiming to trace the origin of LLM answers within the context of RAG, Rorseth et al. [145] proposed a post-hoc explainer called RAGE, which deduces the provenance and salience of external knowledge used during RAG. This framework is designed to generate counterfactual explanations for LLM answers by employing different combinations and perturbations of external knowledge sources. To enhance efficiency, it incorporates pruning strategies that reduce the search space for counterfactual explanations. Through three challenging use cases, RAGE demonstrates its effectiveness in explaining why LLMs produce specific answers, thereby improving their transparency and interpretability for users.

Another line of work provides ante-hoc explanations to either understand the decision-making process of the generator or enhance the generation performance after incorporating them into the model forward process. In the research of [114], a novel method called reasoning on graphs (RoG) is proposed to perform such an ante-hoc explanation. The presented framework utilizes knowledge graph relations to generate reasoning paths for the given question, making the reasoning paths interpretable and faithful. Then, an LLM is employed to generate the answer by conducting reasoning on the paths, which is intuitive to understand the decision-making process of the LLM. Through this framework, the generation process can be interpretable and traceable, not only enhancing the explainability in generation but also allowing the RAG system to generate more accurate answers.

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 information retrieval plays a critical role in RAG [34], the enhancement of adding explanations during retrieval could also benefit the whole RAG system. The ExaRanker method [50] utilizes the explanations as additional 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 [171] 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 features and contrastive cases, contribute to an external knowledge repository for LLMs. This RAG design can thus generate human-friendly and faithful explanations for the prediction model of the scene-understanding tasks. Another work [75] 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 [8]. 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 [122]. Another metric often used is *stability* [57, 96, 133]. It measures the consistency of a method in producing similar explanations for similar or closely related inputs [179].

To evaluate the explanations of the generator in a RAG framework, Sudhi et al. [164] 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.

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. [164] 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. [114] 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) [209] and Complex WebQuestions (CWQ) [170]. 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

Integration of Knowledge Graphs and LLMs. Our work pioneers the integration of knowledge graphs (KGs) and large language models (LLMs) to enhance retrieval faithfulness in Retrieval-Augmented Generation (RAG) systems [124]. Unlike existing approaches that rely solely on GNN-based retrieval methods or LLM-based prompting, our hybrid approach leverages the strengths of both. This integration opens opportunities for more robust and versatile solutions. Beyond question-answering tasks, we propose extending retrieval capabilities to semi-structured knowledge bases, providing richer contexts that enhance text generation and personalization. Future research should focus on optimizing the interaction between these components to improve retrieval accuracy and context integration. Additionally, optimizing the structure and accessibility of the knowledge base is essential for achieving better performance in diverse applications.

Explaining Multi-Component RAG Systems. RAG systems are inherently complex, with multiple interacting components, e.g., retrievers and generators. Explaining the behaviors of these components introduces unique challenges, especially when they are jointly trained, as seen in approaches by [44] and [94]. In such cases, the retriever and generator often share embeddings or feature representations, making it difficult to disentangle their individual contributions. The advent of novel retrievers, such as those using LLMs as agents for sequential graph traversal [78, 185], further complicates explainability. Traditional differentiable-based explanation methods are often impractical for such systems. Future work should prioritize developing explainability methods that provide insights into both individual components and their interactions within the system. This could include frameworks that combine the outputs of retrievers and generators to analyze their joint contributions to system behavior.

Performance vs. Explainability Trade-Off. An ongoing debate in explainable AI concerns the trade-off between model accuracy and explainability. Some studies [33, 13] argue that optimizing for one often compromises the other, while others contest this notion [16, 148], citing a lack of

Table 7: Taxonomy for RAG Accountability

Module	Reference	Technique	Type	Year
Retrieval	Liu et al. [101]	Format-based	Text Watermarking	2024
	Xu et al. [196]	Embedding-based	Data Watermarking	2024
	Sun et al. [168]	Trigger-based	Data Watermarking	2022
Generation	Christ et al. [29]	Semantic-based	Sentence-level Watermarking	2024
	Hou et al. [71]	Sampling-based	Token-level Watermarking	2023
	Kirchenbauer et al. [83]	Logit-based	Global Watermarking	2023
	Yang et al. [205]	Post-generation	Lexical/Syntactic	2022
Retrieval + Generation	Jovanovic et al. [80]	Integrated Pipeline	Red-Green Token Scheme	2024

conclusive evidence. In the context of RAG systems, replacing components like retrievers or generators with fully explainable counterparts offers a promising area for exploration. Researchers could investigate the performance implications of such replacements, analyzing whether transparency and interpretability can coexist with high performance. This line of inquiry could yield valuable insights into the relationship between explainability and system effectiveness, particularly when compared to black-box LLM-based approaches.

Evaluation Metrics for Explainability. Evaluating the quality of explanations in RAG systems is a challenging yet essential area of research. Existing metrics like fidelity and stability are useful but often fail to capture the subjective, context-dependent nature of explainability. Future work should focus on developing domain-specific metrics tailored to user needs and application contexts. For instance, RAG-Ex introduces metrics such as significance and plausibility. Significance evaluates whether explanations align with predefined information, while plausibility assesses whether they reflect annotators' choices. Expanding such metrics to encompass diverse domains and user preferences will provide a more nuanced understanding of explanation quality and its impact on system usability.

Propagation of Misinformation. The propagation of misinformation remains a critical concern in RAG systems, particularly in scenarios involving multi-step reasoning or complex retrieval processes. Explainability methods that trace the flow of information through RAG components could help identify and mitigate sources of misinformation. For example, analyzing how erroneous retrievals influence generated outputs could inform the design of more robust systems. Future research should investigate how explainability frameworks can incorporate misinformation detection and prevention mechanisms, ensuring that generated responses are not only accurate but also trustworthy.

8 Accountability of Retrieval Augmented Generation

Accountability, in the context of AI, refers to the ability to determine whether decisions or outputs align with established procedural and substantive standards and to identify who is responsible when those standards are violated [40]. While accountability focuses on policy and standards, we identify one key technical challenge towards accountability: the ability to identify *ownership*. Generative AI poses unique challenges regarding the attribution of 'speech'—that is, determining who should bear responsibility for its outputs. It is crucial to identify and hold the appropriate entities responsible when outputs deviate from procedural or substantive standards. This necessitates robust mechanisms for content traceability, such as watermarking techniques, which can link outputs back to their source models, datasets, or operators. *Watermarking* thus serves as a critical tool for implementing ownership, enabling the identification of stakeholders responsible for ensuring procedural compliance and addressing any deviations. By providing a transparent means of associating outputs with their origins, watermarking helps bridge the gap between technical and policy-oriented accountability, making it an essential component of the accountability framework for generative AI systems. For the above reasons, we dedicate this section of discussion to the watermarking in 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 plays a crucial role in ensuring accountability in the outputs of LLMs, with strate- gies applicable at different stages of the generation process. Pre-generation watermarking involves embedding watermarks during the training phase, enabling the model to generate content that inher- ently contains identifiable markers. In-generation watermarking integrates watermarking algorithms directly into the text generation process, embedding watermarks as the LLM produces text during inference. This approach ensures that the watermarks are seamlessly woven into the generated output in real time. Finally, post-generation watermarking applies text watermarking techniques to LLM- generated content after the text is produced, embedding markers without requiring modifications to the generation process itself. Each of these methods provides unique advantages, collectively strengthening the traceability and authenticity of LLM-generated text. 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 [101]. Format-based watermarking algorithms are commonly employed as they only embed watermarks in the text format without altering the author's contents [101]. This includes line or word shifting methods [22] and unicode-based approaches [135, 142, 150]. 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 [118] and feature embedding, e.g., bookmarks or variables [76]. 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 [20] and watermark forgery due to the detectable pattern in watermarked text formats [101].

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 [168], semantically invariant transformations in code [167], or distinctive input formats designed to be recognizable [196]. 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 the generated content. *Global watermarking* embeds markers in all generated outputs, enabling consistent content tracking without the need for specific triggers. This approach integrates watermarking directly into the model's parameters, with methods including sampling-based or logit-based watermark distillation [61] and reinforcement learning with feedback from watermark detectors [199]. Global watermarking's broader applicability makes it an attractive option for large-scale deployment, particularly 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 in the inference time. Unlike pre-generation watermarking, which embeds markers into the model during training, in-generation watermarking dynamically modifies the generated outputs, providing flexibility and adaptability without altering the 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 [83] 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 [113] and sliding window methods [84], have been proposed to enhance detectability and robustness in these challenging scenarios.

In addition to modifying logits, in-generation watermarking can be achieved during token sampling on both the token level and sentence level. Token-level sampling introduces watermarks by biasing the random seed or pseudo-random number generator guiding token selection. For instance, Christ et al. [29] proposed using a fixed random sequence to verify watermarked outputs by aligning tokens with pre-determined patterns. While effective, this method faces challenges related to robustness against text edits. To address these limitations, sentence-level sampling watermarking focuses on semantic-level modifications. Algorithms like SemStamp [71] partition the semantic space into watermarked and non-watermarked regions, ensuring that entire sentences maintain watermark integrity even after semantic-preserving modifications. This approach enhances robustness against text editing while enabling more meaningful watermark detection at a higher granularity.

In-generation watermarking offers significant advantages, such as its flexibility and adaptability during inference. 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 [49] and semantic-aware watermarking [53] 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[101].

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. Techniques often involve synonyms or contextually appropriate replacements. Early methods, such as those using WordNet or Word2Vec [174, 47, 119], were limited by their lack of context awareness, which could compromise text quality. Recent advances, like BERT-based infill models[205, 212], have improved context sensitivity, enabling more robust and semantically coherent watermarking. These methods enhance resilience against watermarking removal like reformatting, significantly enhancing robustness.

Syntactic-based watermarking modifies the grammatical structure of sentences to embed watermarks. This involves transformations such as adjunct movement, clefting, and passivization, [14], later expanded with activization and topicalization [173]. While effective in embedding watermarks, 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 [101].

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 [2] and REMARK-LLM [221], utilize transformer-based architectures to embed high-capacity watermarks while maintaining the quality and naturalness of the text. Techniques like WATERFALL [91] 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.

8.4 Accountability in RAG Systems

While watermarking techniques for retrieval and generation have been extensively studied, most existing research addresses these stages independently. Few approaches tackle the unique challenges of watermarking in RAG systems, which require seamless integration across both retrieval and generation processes. WARD (Watermarking for RAG Dataset Inference) [80] stands out as a pioneering method that integrates watermarking across both stages, distinguishing it from traditional approaches that focus solely on one.

WARD bridges the gap between retrieval and generation watermarking by embedding imperceptible signals into datasets during their creation or integration. This ensures traceability throughout both the retrieval and generation stages. Traditional watermarking techniques either track dataset provenance during retrieval or trace outputs during generation but fail to provide a unified solution for RAG systems. WARD's approach ensures that these embedded signals remain detectable even after transformations common in RAG workflows, such as paraphrasing or reformatting. By modifying token probabilities to embed watermarks, WARD offers a robust and unified solution for dataset ownership attribution that is resistant to obfuscation.

A key strength of WARD is its ability to handle complex scenarios that challenge traditional methods, such as fact redundancy where multiple documents contain overlapping or similar information. Watermarking approaches limited to the retrieval stage may fail to detect dataset usage after generative transformations, while those focusing only on the generation stage may overlook the contributions of source datasets. WARD's innovative red-green token-based scheme generates statistically significant signals that persist throughout the entire RAG pipeline. This enables data owners to confidently identify their datasets even when content is blended with other sources or altered during generation.

Furthermore, WARD provides statistical guarantees for watermark detection, significantly reducing false positives and negatives. Techniques such as aggregated queries enhance reliability without compromising computational efficiency, making WARD both scalable and practical for real-world applications. Its ability to handle large-scale datasets and the inherent complexities of RAG systems further sets it apart from traditional watermarking approaches that consider retrieval and generation stages independently.

Table 8: Watermarking Datasets and Metrics

Dataset	Metric		
	Detectability	Quality Impact	Robustness
WaterBench [177]	✓		
WaterJudge [120]	✓	✓	
Mark My Words [131]	✓		✓
MarkLLM [126]	✓	✓	✓

Overall, WARD represents a significant advancement in watermarking for RAG systems by effectively integrating techniques across both retrieval and generation stages. Its comprehensive approach not only ensures robust dataset ownership attribution but also addresses the limitations of traditional methods that treat these stages separately. As RAG systems become increasingly prevalent, solutions like WARD are essential for protecting intellectual property and ensuring ethical data usage. Future developments may build upon WARD's framework to enhance watermark resilience and adapt to evolving data transformation techniques.

8.5 Accountability Evaluation

8.5.1 Metrics

Evaluating watermarking techniques for large language models involves a comprehensive framework of metrics, focusing on detectability, quality impact, output performance, output diversity, and robustness [101]. These metrics ensure the effectiveness of watermarking systems while minimizing quality degradation and maximizing resilience against attacks.

Detectability is a fundamental metric that evaluates the ability to identify the presence of a watermark in text. For zero-bit watermarking, the focus is on determining whether a watermark exists, without recovering specific information. This is typically achieved using statistical methods such as z-scores or p-values, with careful calibration to minimize false positives, as they can misclassify human-generated text. On the other hand, multi-bit watermarking involves extracting encoded information, which is assessed using metrics like Bit Error Rate (BER) [212] and bit accuracy[213]. Additionally, watermark size [130], which refers to the text length required for reliable detection, is critical, with longer texts generally improving detectability at the cost of applicability in shorter content.

Watermarking techniques must preserve the quality of the generated text, ensuring that the output remains coherent and natural. Quality metrics are divided into comparative and single-text methods. *Comparative metrics* evaluate differences between watermarked and non-watermarked text, using surface-level measures like BLEU [127] and Meteor [9] or semantic-level measures like Semantic Score and Entailment Score, which leverage embeddings to capture deeper relationships. *Single-text* metrics, like Perplexity (PPL), focus on the coherence of the watermarked text independently. Lower PPL indicates higher fluency and coherence, while human evaluation remains the gold standard for quality assessment.

Output performance metrics assess whether the capabilities of watermarked LLMs remain intact across downstream tasks. For text completion, metrics like PPL [212], GPT-4-based scoring [43], and semantic similarity measures are used to ensure the generated text aligns with prompts. Code generation tasks require precise evaluation using metrics such as CodeBLEU [63], which captures lexical and semantic accuracy, and Edit Sim [177], which measures the similarity between reference and generated code. Other downstream tasks, including machine translation, summarization, and question answering, are evaluated using standard metrics like BLEU, ROUGE [100], and task-specific accuracy measures to ensure watermarking does not degrade model utility.

Output *diversity metrics* are essential to evaluate the potential restriction in creativity or variability introduced by watermarking. Metrics like Seq-Rep-N [62] measure lexical diversity by calculating the ratio of unique n-grams to total n-grams in text. Log Diversity [85] builds on this by quantifying the diversity logarithmically across different n-grams. Additionally, entropy-based metrics such as Ent-3 [70] and Sem-Ent [64] provide insights into lexical and semantic diversity, respectively. Higher entropy scores indicate greater diversity, ensuring that the watermarking process does not overly constrain model outputs.

Robustness evaluates the ability of watermarked systems to resist adversarial attacks, which can be classified into untargeted attacks and targeted attacks. For details, see Section 5 Robustness.

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 large language models (LLMs), including WaterBench [177], WaterJudge [120], Mark My Words [131], and MarkLLM [126]. The details about each dataset can be viewed in Table 8.

8.6 Future Direction of RAG Accountability

Unifying Retrieval and Generation Watermarking. Currently, retrieval and generation watermarking techniques operate independently within RAG systems, leaving gaps in overall accountability. Future research should develop unified frameworks that seamlessly integrate both methods. By embedding traceability throughout the entire RAG pipeline, these frameworks would enhance intellectual property protection, ensure responsible attribution, and maintain data and model integrity in complex AI systems where retrieval and generation increasingly overlap.

Dynamic Watermarking for Adaptive AI. As generative AI systems evolve to adapt to real-time inputs and changing user contexts, static watermarking becomes insufficient. Developing dynamic watermarking techniques that adjust to system updates, counter adversarial attacks, and respond to shifts in model behavior is crucial. These adaptive methods would enhance robustness and maintain traceability in RAG architectures, supporting accountability in ever-changing AI environments.

Governance and Ethical Integration. Besides technical innovation, alignment with legal and ethical frameworks is essential. Future efforts should foster collaborations among technologists, policymakers, and ethicists to establish governance models that incorporate watermarking as a foundational tool for AI accountability and intellectual property protection. Such interdisciplinary work would ensure that RAG systems adhere to global ethical standards and legal requirements.

9 Applications

In the preceding sections, we delved into the six dimensions of Trustworthy RAG systems, outlining their foundational principles and challenges. Building on this foundation, we now extend the discussion to their practical applications across high-stakes domains. Specifically, we focus on Healthcare, Legal, and Education. For each domain, we provide a summary of the current landscape and highlight recent advancements. Additionally, we identify domain-specific open problems that present opportunities for further research and development.

9.1 Healthcare

9.1.1 RAG Use Cases in Healthcare

Clinical Decision Support. One of the primary healthcare applications of LLMs is in Clinical Decision Support (CDS) for medical professionals [93]. These systems enhance the decision-making process by providing access to diagnostic guidelines, treatment plans, and patient history for medical professionals. Although the majority of clinical models leverage fine-tuning for knowledge injection, recent results have shown that by integrating the retrieval of relevant medical documents and personalized patient data, RAG systems can improve the performance of the model to make more informed and accurate decisions [195].

Patient Communication. While CDS systems are geared toward assisting medical professionals, RAG systems can also play a vital role in patient-facing communication. There are mainly two categories of patient-facing communication: question answering (QA) and dialogue support [66]. The Centers for Disease Control and Prevention (CDC) has shown that 58% of US adults have used the internet to search for medical information [25]. This widespread reliance on online health information underscores the need for personalized and reliable medical guidance. RAG systems, by integrating structured retrieval with generative capabilities, can enhance the accuracy and relevance

of patient interactions. For instance, a RAG-based chatbot might help patients assess symptoms and recommend next steps, such as scheduling a medical consultation or seeking immediate care.

Knowledge Discovery. AI has been extensively used in various medical knowledge discovery tasks such as drug discovery. Traditional works prioritize graph-based approaches for analyzing molecular structures and drug-drug interactions, leveraging methods such as GNN [144, 7]. Recently, researchers started to look at LLM for science and explore the application of LLMs in medical-related fields, particularly for accelerating drug discovery and development through improved information processing and hypothesis generation [125]. RAG systems can significantly accelerate the research process by retrieving the latest research findings, publications, and clinical trial data [204]. For example, research has shown that RAG can greatly enhance the prediction on Nephrology [117]. By synthesizing large volumes of medical literature into a retrieval database, these systems provide valuable insights that can inform future studies, improve hypotheses, and guide the direction of medical innovation [147].

Precharting. Precharting is a crucial healthcare application that involves a physician reviewing a patient's medical records before a visit [21]. Studies have demonstrated that LLMs encode extensive medical knowledge [161], while retrieval-augmented generation (RAG) can enhance personalization and contextual relevance in clinical settings [204]. Although the integration of RAG into precharting remains largely unexplored, these observations suggest its potential to transform the process by improving efficiency and reducing load of healthcare professionals. Future research could further investigate how RAG can optimize precharting workflows and enhance patient care.

9.1.2 Trustworthiness Challenges in Healthcare

Reliability. Previous research have identified reliability as one of the main challenge to healthcare applications [66]. Healthcare applications demand an exceptionally high level of reliability due to the high-stakes impact of decisions made using RAG outputs. Uncertainty quantification has been extensively used in various healthcare applications [153], but its use in LLM-based applications remains largely unexplored. Applications such as CDS and Patient Communication should include uncertainty quantification measures to enable end-users to make informed decisions by understanding the confidence levels of the outputs. Furthermore, these systems require robustness to ensure consistent performance under challenging or unexpected input. Current research often focuses on general question answering [124, 114, 166]; however, we advocate for increased focus on domain-specific and individual difference challenges in medical knowledge. Tailoring RAG systems to the requirements of healthcare can enhance their reliability and effectiveness in real-world applications.

Privacy. Another critical challenge in healthcare RAG applications is maintaining the privacy of sensitive medical data. The personal nature of medical information makes it imperative to safeguard against data breaches and misuse. Zeng et al. [158] highlight the risks of personal identification information leakage in medical data, representing an initial effort to address this issue. Since applications such as precharting heavily rely on the retrieval of personal medical records, ensuring the privacy of the data is imperative given the high stakes [21]. Future research should extend these efforts to more complex, real-world scenarios where multiple stakeholders, data sources, and regulatory frameworks intersect. Developing privacy-preserving mechanisms tailored to healthcare RAG systems will be crucial to fostering trust and ensuring compliance with legal and ethical standards.

Others. While Reliability and Privacy are particularly critical in healthcare applications, the other dimensions of trustworthiness also play significant roles. For instance, healthcare RAG systems must be safe and secure, i.e., resilient to adversarial attacks, as malicious actors could manipulate critical outputs, potentially leading to harmful consequences for patients [56]. Fairness is equally essential, as healthcare is a fundamental right, and biases in RAG outputs could disproportionately disadvantage certain populations or exacerbate health disparities [66]. Additionally, Explainability and Accountability are crucial to building trust with both medical professionals and patients. Explainable systems allow users to understand the rationale behind recommendations, while accountability mechanisms ensure that errors or unintended consequences can be traced and addressed effectively. Together, these dimensions create a comprehensive framework for trustworthy healthcare RAG systems.

9.2 Law

9.2.1 RAG Use Cases in Law

Legal Question Answering. One of the key applications of RAG systems in the legal field is a legal question answering (LQA) [90, 27]. LQA systems aim to provide answers to queries related to law, cases, and theoretical analysis. Due to the idiosyncratic nature of the laws in different jurisdictions, current methods focus on fine-tuning existing language models on specific laws or jurisdictions [4, 116, 1]. These systems streamline the process of finding precedents, statutes, and relevant case law, enabling legal professionals to build stronger arguments and make informed decisions. Recent research has shown that RAG can also help provide useful context for LQA [190] besides fine-tuning. Given the highly specialized nature of law, future research should focus on effectively integrating context to LQA through various methods (Fine-tuning, RAG, etc.).

Legal Document Summarization. Legal Document Summarization (LDS) has become an increasingly popular use case of large language models (LLMs) in the legal community [12, 27]. Legal documents are often long and complex, making manual summarization time-consuming and error-prone. LDS assists legal professionals and researchers by enhancing efficiency in legal analysis. Commercial models such as Claude have incorporated rule-based summarization techniques alongside state-of-the-art LLMs to improve the quality and relevance of summaries. However, the integration of RAG into LDS remains underexplored. By incorporating RAG into LDS systems, legal summaries can better capture relevant case law, statutes, and contextual references, leading to improved factual consistency and enhanced robustness [104].

Legal Judgment Prediction. Legal Judgment Prediction (LJP) aims to forecast court rulings based on case fact descriptions [27]. It plays a crucial role in legal decision support, assisting judges, lawyers and clients in analyzing case outcomes. Early research formulated LJP as a classification task using traditional machine learning models [35]. However, to better reflect real-world judicial reasoning, recent studies have integrated external legal databases and RAG techniques to incorporate precedent cases and legal statutes [193], enhancing the interpretability and reliability of predictions.

9.2.2 Trustworthiness Challenges in Legal Applications

Fairness. Fairness is a critical concern in legal RAG applications, as biases in data or algorithms can lead to unjust or discriminatory outcomes. For instance, biases in training data may disproportionately affect marginalized groups by retrieving prejudiced legal precedents or overlooking relevant cases. Existing research has shown that LLMs and RAG systems are susceptible to both explicit and implicit biases [159, 181]. While initial methods have been proposed to mitigate these biases [54], dedicated research in legal contexts remains limited. Addressing fairness in legal RAG systems requires not only debiasing training data but also designing algorithms that promote equity in document retrieval and output generation. Future research should explore domain-specific challenges to ensure just and transparent legal RAG systems.

Explainability. Explainability is essential in legal RAG applications due to the complexity and high-stakes nature of legal decision-making [35, 27]. Legal professionals require transparent insights into how these systems retrieve and synthesize information. For instance, when generating judicial opinions, it is crucial to trace the origins of retrieved precedents and legal arguments to assess their relevance, credibility, and potential biases. While existing research has explored explainability in general LLMs, recent studies have specifically examined interpretable methods for long-form legal question answering [111]. Future work should focus on developing domain-specific explainability frameworks that ensure legal RAG systems provide justifications aligned with legal reasoning.

Others. In addition to Fairness and Explainability, other dimensions of trustworthiness play a vital role in legal RAG applications. Reliability needs to be ensured as the output of the legal applications needs to be factually grounded with reliable uncertainty and robustness estimation. Privacy is paramount when handling sensitive client information or confidential case details, necessitating the development of privacy-preserving mechanisms compliant with legal and ethical standards. Furthermore, similar to the healthcare applications, accountability are essential to fostering trust among legal professionals and clients. Accountability mechanisms ensure that errors, biases, or

unintended consequences can be identified and addressed. Together, these dimensions establish a comprehensive foundation for trustworthy RAG systems in legal applications.

9.3 Education

9.3.1 RAG Use Cases in Education

Personalized Learning. LLMs have demonstrated significant potential in education [184], with personalized learning being one of their key applications. Recent advancements in educational research highlight the effectiveness of LLMs in learning path planning [123]. However, existing approaches often lack dynamic retrieval mechanisms to adapt to different student needs and contextual knowledge gaps. Integrating RAG can enhance personalized learning by incorporating up-to-date information, ensuring more adaptive and tailored learning experiences. This approach has the potential to improve student engagement, optimize learning trajectories, and foster deeper comprehension.

Student Support. RAG systems can assist students by providing real-time answers to academic queries and offering guidance on educational content [184, 37]. For example, a RAG-powered chatbot could help students understand complex concepts that might not be necessarily present in the training material. By addressing common questions and providing actionable insights, these systems can reduce the burden on teachers and counselors while providing valuable assistance to the students.

Teacher Support. Another valuable application of RAG systems is assisting teachers in classroom instruction by integrating external databases to provide up-to-date knowledge and resources. This can help educators generate lesson materials and answer student queries based on real-time information. Current research has shown promising preliminary results in leveraging LLMs for classroom support using reddit as a data source [121]. However, there remains a gap in developing high-quality, domain-specific datasets and more sophisticated models tailored for real-world classroom settings. Future research should focus on refining dataset curation and improving model adaptability to better integrate RAG into educational practices.

9.3.2 Trustworthiness Challenges in Educational Applications

Fairness. Recent research has highlighted that LLMs often exhibit explicit and implicit biases [181, 156, 157]. In the context of educational applications, ensuring fairness is crucial to provide all students with equal access to learning opportunities and unbiased content. For example, a RAG-based learning system might retrieve or synthesize information influenced by biases present in its training data, potentially reinforcing stereotypes or disadvantaging students from underrepresented groups. Furthermore, if the system inadvertently retrieves or generates content based on protected attributes such as race, gender, or socioeconomic status, it could lead to unequal treatment or negative educational outcomes.

Safety. As shown in our previous discussion, studies have demonstrated that RAG systems are vulnerable to various adversarial attacks [39, 200, 189]. These vulnerabilities present significant safety concerns in educational settings. For instance, a student using a RAG-based learning platform could be exposed to harmful or misleading content if a malicious actor successfully jailbreaks the system to bypass safety filters. Such breaches could lead to the dissemination of inappropriate or even dangerous materials, undermining the utility of the system and harming the minors. To mitigate these risks, robust adversarial defense mechanisms tailored to the application are essential to ensure that educational RAG applications remain secure and reliable for learners.

Others. Beyond Fairness and Safety, other dimensions of trustworthiness are equally critical for educational applications of RAG systems. First, Reliability must be ensured to emphasize the accuracy and consistency of retrieved or synthesized content. Next, Privacy is a vital consideration, as educational platforms often handle sensitive personal information such as student performance, learning behaviors, and demographic data. A failure to protect privacy not only undermines trust but could also expose students to risks such as data breaches or identity theft. Lastly, Robustness is crucial for maintaining system trustworthiness in dynamic educational environments where students and learning objectives might constantly change. Educational RAG systems must adapt to diverse user queries, varying levels of prior knowledge, and potential ambiguities in input while ensuring stable 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. This work addresses the pressing need for a unified perspective in the field, bridging the gap in understanding and systematically categorizing the challenges and solutions in developing trustworthy RAG systems. 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 and ethical deployment of AI technologies.

References

[1] Abdelrahman Abdallah, Bhaskar Piryani, and Adam Jatowt. 2023. Exploring the state of the art in legal QA systems. Journal of Big Data 10, 127 (2023). https://doi.org/10.1186/s40537-023-00802-8

[2] 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.

[3] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. In Proceedings of the International Conference on Learning Representations (ICLR). arXiv:2001.09977 [cs.CL] https://arxiv.org/abs/2001.09977

[4] Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. In Findings of the Association for Computational Linguistics: EMNLP 2020, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 743–749. https://doi.org/10.18653/v1/2020.findings-emnlp.66

[5] 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] https://arxiv.org/abs/2311.07361

[6] Rama Akkiraju, Anbang Xu, Deepak Bora, Tan Yu, Lu An, Vishal Seth, Aaditya Shukla, Pritam Gundecha, Hridhay Mehta, Ashwin Jha, Prithvi Raj, Abhinav Balasubramanian, Murali Maram, Guru Muthusamy, Shivakesh Reddy Annepally, Sidney Knowles, Min Du, Nick Burnett, Sean Javiya, Ashok Marannan, Mamta Kumari, Surbhi Jha, Ethan Dereszenski, Anupam Chakraborty, Subhash Ranjan, Amina Terfai, Anoop Surya, Tracey Mercer, Vinodh Kumar Thanigachalam, Tamar Bar, Sanjana Krishnan, Samy Kilaru, Jasmine Jaksic, Nave Algarici, Jacob Liberman, Joey Conway, Sonu Nayyar, and Justin Boitano. 2024. FACTS About Building Retrieval Augmented Generation-based Chatbots. arXiv:2407.07858 [cs.LG] https://arxiv.org/abs/2407.07858

[7] M. H. Al-Rabeah and A. Lakizadeh. 2022. Prediction of drug-drug interaction events using graph neural networks based feature extraction. Scientific Reports 12 (2022), 15590. https://doi.org/10.1038/s41598-022-19999-4

[8] 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.

[9] Mohammed Hazim Alkawaz, Ghazali Sulong, Tanzila Saba, Abdulaziz S Almazyad, and Amjad Rehman. 2016. Concise analysis of current text automation and watermarking approaches. Security and Communication Networks 9, 18 (2016), 6365–6378.

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

[11] 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] https://arxiv.org/abs/2110.01052

[12] 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.

[13] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion 58 (2020), 82–115.

[14] Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. 2001. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In Information Hiding: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, April 25–27, 2001 Proceedings 4. Springer, 185–200.

[15] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL] https://arxiv.org/abs/1611.09268

[16] 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 Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. 248–266.

[17] 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. https://doi.org/10.3233/SHTI190176

[18] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In WACV.

[19] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963 (2021).

[20] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad characters: Imperceptible nlp attacks. In 2022 IEEE Symposium on Security and Privacy (SP). IEEE, 1987–2004.

[21] CA Bowman and H Holzer. 2021. EMR Precharting Efficiency in Internal Medicine: A Scoping Review. J Med Educ Curric Dev 8 (Jul 2021), 23821205211032414. https://doi.org/10.1177/23821205211032414

[22] Jack T Brassil, Steven Low, Nicholas F Maxemchuk, and Lawrence O'Gorman. 1995. Electronic marking and identification techniques to discourage document copying. IEEE Journal on Selected Areas in Communications 13, 8 (1995), 1495–1504.

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[24] 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 (Lecture Notes in Computer Science, Vol. 11551), Tatiana Antipova and Alvaro Rocha (Eds.). Springer International Publishing, Cham, 121–132. https://doi.org/10.1007/978-3-030-17705-8_10

[25] 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.

[26] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. (2021). arXiv:2107.03374 [cs.LG]

[27] 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). https://arxiv.org/abs/2405.01769

[28] Jaekeol Choi, Euna Jung, Sungjun Lim, and Wonjong Rhee. 2022. Finding Inverse Document Frequency Information in BERT. arXiv preprint arXiv:2202.12191 (2022).

[29] 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.

[30] 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 2018 ACM SIGIR International Conference on Theory of Information Retrieval. 67–74.

[31] 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 (2024). https://doi.org/10.48550/arXiv.2409.08045

[32] 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).

[33] Barnaby Crook, Maximilian Schlüter, and Timo Speith. 2023. Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). In 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW). IEEE, 316–324.

[34] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 719–729.

[35] 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] https://arxiv.org/abs/2204.04859

[36] 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. https://doi.org/10.1007/s11633-024-1510-8

[37] Sagnik Dakshit. 2024. Faculty Perspectives on the Potential of RAG in Computer Science Higher Education. arXiv:2408.01462 [cs.CY] https://arxiv.org/abs/2408.01462

[38] 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).

[39] 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] https://arxiv.org/abs/2402.08416

[40] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. 2017. Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134 (2017).

[41] 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).

[42] OpenAI et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[43] 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. https://eprint.iacr.org/2023/1661

[44] Wenqi Fan, Yujuan 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 Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 6491–6501.

[45] 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). https://arxiv.org/abs/2024.05

[46] Philip Feldman, James R. Foulds, and Shimei Pan. 2024. RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots. arXiv:2403.01193 [cs.CL] https://arxiv.org/abs/2403.01193

[47] Christiane Fellbaum. 1998. WordNet: An electronic lexical database. MIT Press google schola 2 (1998), 678–686.

[48] Qizhang Feng, Ninghao Liu, Fan Yang, Ruixiang Tang, Mengnan Du, and Xia Hu. 2023. Degree: Decomposition based explanation for graph neural networks. arXiv preprint arXiv:2305.12895 (2023).

[49] Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. 2023. Three bricks to consolidate watermarks for large language models. In 2023 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 1–6.

[50] Fernando Ferraretto, Thiago Laitz, Roberto Lotufo, and Rodrigo Nogueira. 2023. Exaranker: Explanation-augmented neural ranker. arXiv preprint arXiv:2301.10521 (2023).

[51] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. A white box analysis of ColBERT. In Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43. Springer, 257–263.

[52] 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.

[53] 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 on Artificial Intelligence, Vol. 38. 18003–18011.

[54] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. Computational Linguistics 50, 3 (09 2024), 1097–1179. https://doi.org/10.1162/coli_a_00524

[55] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997 (2023).

[56] N. Ghaffari Laleh, D. Truhn, G. P. Veldhuizen, et al. 2022. Adversarial attacks and adversarial robustness in computational pathology. Nature Communications 13 (Sep 2022), 5711. https://doi.org/10.1038/s41467-022-33266-0

[57] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In Proceedings of the AAAI conference on artificial intelligence, Vol. 33. 3681–3688.

[58] Nicole Gillespie, Simon Lockey, Claire Curtis, Julian Pool, and Arian Akbari. 2023. Trust in Artificial Intelligence: A Global Study. Technical Report. The University of Queensland and KPMG Australia. https://doi.org/10.14264/00d3c94

[59] 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).

[60] 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.

[61] Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. 2023. On the learnability of watermarks for language models. arXiv preprint arXiv:2312.04469 (2023).

[62] Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. 2024. On the Learnability of Watermarks for Language Models. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=9k0krNzvlV

[63] Batu Guan, Yao Wan, Zhangqian Bi, Zheng Wang, Hongyu Zhang, Pan Zhou, and Lichao Sun. 2024. CodeIP: A Grammar-Guided Multi-Bit Watermark for Large Language Models of Code. In Findings of the Association for Computational Linguistics: EMNLP 2024, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9243–9258. https://doi.org/10.18653/v1/2024.findings-emnlp.541

[64] Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and Improving Semantic Diversity of Dialogue Generation. In Findings of the Association for Computational Linguistics: EMNLP 2022, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 934–950. https://doi.org/10.18653/v1/2022.findings-emnlp.66

[65] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. Advances in neural information processing systems (2016).

[66] 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. https://doi.org/10.1016/j.inffus.2025.102963

[67] 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] https://arxiv.org/abs/2302.13425

[68] 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] https://arxiv.org/abs/2009.03300

[69] Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2023. Why So Gullible? Enhancing the Robustness of Retrieval-Augmented Models against Counterfactual Noise. arXiv preprint arXiv:2305.01579 (2023).

[70] Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2024. SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 4067–4082. https://doi.org/10.18653/v1/2024.naacl-long.226

[71] Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. 2023. Semstamp: A semantic watermark with paraphrastic robustness for text generation. arXiv preprint arXiv:2310.03991 (2023).

[72] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2024. Large Language Models for Software Engineering: A Systematic Literature Review. ACM Trans. Softw. Eng. Methodol. 33, 8, Article 220 (Dec. 2024), 79 pages. https://doi.org/10.1145/3695988

[73] 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] https://arxiv.org/abs/2410.07589

[74] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, Andre Freitas, and Mustafa A. Mustafa. 2023. A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation. arXiv:2305.11391 [cs.AI] https://arxiv.org/abs/2305.11391

[75] 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).

[76] 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 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 1940–1945.

[77] 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 preprint arXiv:2411.14110 (2024). https://doi.org/10.48550/arXiv.2411.14110

[78] Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Suhang Wang, Yu Meng, and Jiawei Han. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. arXiv preprint arXiv:2404.07103 (2024).

[79] 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 Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 1601–1611. https://doi.org/10.18653/v1/P17-1147

[80] Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. 2024. Ward: Provable RAG Dataset Inference via LLM Watermarks. arXiv preprint arXiv:2410.03537 (2024).

[81] 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).

[82] 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).

[83] 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.

[84] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2023. On the reliability of watermarks for large language models. arXiv preprint arXiv:2306.04634 (2023).

[85] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2024. On the Reliability of Watermarks for Large Language Models. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=DEJIDCmWOz

[86] 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. https://doi.org/10.1007/978-3-540-30115-8_22

[87] 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 (New Orleans, LA, USA) (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 3131, 20 pages.

[88] 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] https://arxiv.org/abs/2305.18404

[89] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. Transactions of the Association for Computational Linguistics 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[90] 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] https://arxiv.org/abs/2312.03718

[91] Gregory Kang Ruey Lau, Xinyuan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. 2024. Waterfall: Framework for robust and scalable text watermarking. In ICML 2024 Workshop on Foundation Models in the Wild.

[92] Peter Lee, Sebastien 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. https://doi.org/10.1056/NEJMsr2214184

[93] Alessandro Giaj Levra, Mauro Gatti, Roberto Mene, Dana Shiffer, Giorgio Costantino, Monica Solbiati, Raffaello Furlan, and Franca Dipaola. 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 131 (2025), 113–120. https://doi.org/10.1016/j.ejim.2024.09.017

[94] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33 (2020), 9459–9474.

[95] 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] https://arxiv.org/abs/2110.01167

[96] 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).

[97] Jinming 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] https://arxiv.org/abs/2304.03879

[98] Shuo Li et al. 2023. TRAQ: Trustworthy Retrieval Augmented Question Answering via Conformal Prediction. arXiv preprint arXiv:2307.04642 (2023). https://arxiv.org/abs/2307.04642

[99] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022).

[100] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out. 74–81.

[101] 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.

[102] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil K. Jain, and Jiliang Tang. 2021. Trustworthy AI: A Computational Perspective. arXiv:2107.06641 [cs.AI] https://arxiv.org/abs/2107.06641

[103] Mingrui Liu, Sixiao Zhang, and Cheng Long. 2024. Mask-based Membership Inference Attacks for Retrieval-Augmented Generation. arXiv preprint arXiv:2410.20142 (2024). https://doi.org/10.48550/arXiv.2410.20142

[104] 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] https://arxiv.org/abs/2403.19889

[105] Yang Liu et al. 2024. Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv preprint arXiv:2308.05374 (2024). https://arxiv.org/abs/2308.05374

[106] 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).

[107] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Machine unlearning in generative ai: A survey. arXiv preprint arXiv:2407.20516 (2024).

[108] 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).

[109] Zheyuan Liu, Chunhui Zhang, Yijun Tian, Erchi Zhang, Chao Huang, Yanfang Ye, and Chuxu Zhang. 2023. Fair graph representation learning via diverse mixture-of-experts. In Proceedings of the ACM Web Conference 2023. 28–38.

[110] 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] https://arxiv.org/abs/2402.13532

[111] 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] https://arxiv.org/abs/2309.17050

[112] 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).

[113] 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).

[114] 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.

[115] 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.

[116] Behrooz Mansouri and Ricardo Campos. 2023. FALQU: Finding Answers to Legal Questions. arXiv:2304.05611 [cs.IR] https://arxiv.org/abs/2304.05611

[117] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, and W. Cheungpasitporn. 2024. Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications. Medicina (Kaunas) 60, 3 (Mar 2024), 445. https://doi.org/10.3390/medicina60030445

[118] Nighat Mir. 2014. Copyright for web content using invisible text watermarking. Computers in Human Behavior 30 (2014), 648–653.

[119] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In International Conference on Machine Learning. PMLR, 24950–24962.

[120] 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).

[121] Elizabeth A Mullins, Adrian Portillo, Kristalys Ruiz-Rohena, and Aritran Piplai. 2024. Enhancing classroom teaching with LLMs and RAG. arXiv:2411.04341 [cs.LG] https://arxiv.org/abs/2411.04341

[122] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. Comput. Surveys 55, 13s (2023), 1–42.

[123] Chee Ng and Yuen Fung. 2024. Educational Personalized Learning Path Planning with Large Language Models. arXiv:2407.11773 [cs.CL] https://arxiv.org/abs/2407.11773

[124] 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).

[125] 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 109, 12 (12 2023), 4382–4384. https://doi.org/10.1097/JS9.0000000000000719

[126] 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).

[127] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (Philadelphia, Pennsylvania) (ACL '02). Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[128] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. BBQ: A hand-built bias benchmark for question answering. arXiv preprint arXiv:2110.08193 (2021).

[129] 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). https://doi.org/10.48550/arXiv.2411.01705

[130] 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).

[131] 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).

[132] Nicholas Pipitone and Ghita Houir Alami. 2024. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. arXiv:2408.10343 [cs.AI] https://arxiv.org/abs/2408.10343

[133] Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. 2020. Regularizing black-box models for improved interpretability. Advances in Neural Information Processing Systems 33 (2020), 10526–10536.

[134] Sayantan Polley. 2022. Towards Explainable Search in Legal Text. In European Conference on Information Retrieval. Springer, 528–536.

[135] 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.

[136] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. arXiv preprint arXiv:1904.07531 (2019).

[137] 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] https://arxiv.org/abs/2306.10193

[138] P Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016).

[139] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

[140] Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias?. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.

[141] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 1135–1144.

[142] Stefano Giovanni Rizzo, Flavio Bertini, and Danilo Montesi. 2016. Content-preserving text watermarking through unicode homoglyph substitution. In Proceedings of the 20th International Database Engineering & Applications Symposium. 97–104.

[143] 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.

[144] N. Rohani and C. Eslahchi. 2019. Drug-Drug Interaction Predicting by Neural Network Using Integrated Similarity. Scientific Reports 9 (2019), 13645. https://doi.org/10.1038/s41598-019-50121-3

[145] Joel Rorseth, Parke Godfrey, Lukasz Golab, Divesh Srivastava, and Jaroslaw Szlichta. 2024. RAGE Against the Machine: Retrieval-Augmented LLM Explanations. arXiv preprint arXiv:2405.13000 (2024).

[146] Pouria Rouzrokh, Shahriar Faghani, Cooper U. Gamble, Moein Shariatnia, and Bradley J. Erickson. 2024. CONFLARE: CONFormal LArge language model REtrieval. arXiv preprint arXiv:2404.04287 (2024). https://arxiv.org/abs/2404.04287

[147] Mark Roy, Baichuan Sun, Nihir Chadderwala, Derrick Choo, Mani Khanuja, and Frank Winkler. 2024. Use RAG for drug discovery with Amazon Bedrock Knowledge Bases. Amazon Web Services. https://aws.amazon.com/blogs/machine-learning/use-rag-for-drug-discovery-with-amazon-bedrock-knowledge-bases/

[148] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys 16 (2022), 1–85.

[149] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. arXiv preprint arXiv:2304.11406 (2023).

[150] Ryoma Sato, Yuki Takezawa, Han Bao, Kenta Niwa, and Makoto Yamada. 2023. Embarrassingly simple text watermarks. arXiv preprint arXiv:2310.08920 (2023).

[151] Johannes Schneider. 2024. Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda. Artificial Intelligence Review 57, 11 (2024), 289.

[152] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. International journal of computer vision 128 (2020), 336–359.

[153] 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. https://doi.org/10.1016/j.compbiomed.2023.107441

[154] Glenn Shafer and Vladimir Vovk. 2007. A tutorial on conformal prediction. arXiv:0706.3188 [cs.LG] https://arxiv.org/abs/0706.3188

[155] 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] https://arxiv.org/abs/2310.10844

[156] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "Nice try, kiddo": Investigating Ad Hominems in Dialogue Responses. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, 750–767.

[157] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal Biases in Language Generation: Progress and Challenges. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, 4275–4293.

[158] Pengfei He Yue Xing Yiding Liu Han Xu Jie Ren Shuaiqiang Wang Dawei Yin Yi Chang Jiliang Tang Shenglai Zeng, Jiankun Zhang. 2024. The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG). ACL (2024). https://arxiv.org/abs/2402.16893

[159] 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.

[160] Jaspreet Singh, Megha Khosla, Wang Zhenye, and Avishek Anand. 2021. Extracting per query valid explanations for blackbox learning-to-rank models. In Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. 203–210.

[161] K. Singhal, S. Azizi, T. Tu, et al. 2023. Large language models encode clinical knowledge. Nature 620 (Aug 2023), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[162] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617 [cs.CL] https://arxiv.org/abs/2305.09617

[163] 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] https://arxiv.org/abs/2403.01216

[164] Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2776–2780.

[165] 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] https://arxiv.org/abs/2409.12524

[166] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model with Knowledge Graph. In International Conference on Learning Representations. arXiv:2307.07697 [cs.CL]

[167] Zhensu Sun, Xiaoning Du, Fu Song, and Li Li. 2023. Codemark: Imperceptible watermarking for code datasets against neural code completion models. In Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 1561–1572.

[168] 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.

[169] 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.

[170] Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. arXiv preprint arXiv:1803.06643 (2018).

[171] Sule Tekkesinoglu and Lars Kunze. 2024. From Feature Importance to Natural Language Explanations Using LLMs with RAG. arXiv preprint arXiv:2407.20990 (2024).

[172] 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).

[173] 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.

[174] 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 Proceedings of the 8th workshop on Multimedia and security. 164–174.

[175] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] https://arxiv.org/abs/2302.13971

[176] Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. 2022. Plex: Towards Reliability using Pretrained Large Model Extensions. arXiv preprint arXiv:2207.07411 (2022). https://arxiv.org/abs/2207.07411

[177] 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 Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1517–1542. https://doi.org/10.18653/v1/2024.acl-long.83

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

[179] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion 76 (2021), 89–106.

[180] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards axiomatic explanations for neural ranking models. In Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. 13–22.

[181] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. arXiv preprint arXiv:2310.09219 (2023). https://doi.org/10.48550/arXiv.2310.09219 Accepted to EMNLP 2023 Findings.

[182] 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).

[183] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Manias Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: a comprehensive assessment of trustworthiness in GPT models. In Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 1361, 108 pages.

[184] 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] https://arxiv.org/abs/2403.18105

[185] 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 on Artificial Intelligence, Vol. 38. 19206–19214.

[186] 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).

[187] Ziqiu Wang, Jun Liu, Shengkai Zhang, and Yang Yang. 2024. Poisoned LangChain: Jailbreak LLMs by LangChain. arXiv:2406.18122 [cs.CL] https://arxiv.org/abs/2406.18122

[188] Pierre Wargnier, Samuel Benveniste, Pierre Jouvelot, and Anne-Sophie Rigaud. 2018. Usability Assessment of Interaction Management Support in LOUISE, an ECA-based User Interface for Elders with Cognitive Impairment. Technology and Disability 30, 3 (2018), 105–126. https://doi.org/10.3233/TAD-180189

[189] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: how does LLM safety training fail?. In Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 3508, 32 pages.

[190] Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne L@inproceedingscohen2018understanding, title=Understanding the representational power of neural retrieval models using NLP tasks, author=Cohen, Daniel and O'Connor, Brendan and Croft, W Bruce, booktitle=Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, pages=67–74, year=2018 @inproceedingscohen2018understanding, title=Understanding the representational power of neural retrieval models using NLP tasks, author=Cohen, Daniel and O'Connor, Brendan and Croft, W Bruce, booktitle=Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, pages=67–74, year=2018 iret, and Bruno Fleisch. 2023. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. arXiv preprint arXiv:2312.10997 (2023).

[191] Caesar Wu, Yuan-Fang Lib, and Pascal Bouvry. 2023. Survey of Trustworthy AI: A Meta Decision of AI. arXiv:2306.00380 [cs.AI] https://arxiv.org/abs/2306.00380

[192] 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).

[193] Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12060–12075. https://doi.org/10.18653/v1/2023.emnlp-main.740

[194] Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably Robust RAG against Retrieval Corruption. arXiv:2405.15556 [cs.LG] https://arxiv.org/abs/2405.15556

[195] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. arXiv:2402.13178 [cs.CL] https://arxiv.org/abs/2402.13178

[196] Hengyuan Xu, Liyao Xiang, Xingjun Ma, Borui Yang, and Baochun Li. 2024. Hufu: A Modality-Agnositc Watermarking System for Pre-Trained Transformers via Permutation Equivariance. arXiv preprint arXiv:2403.05842 (2024).

[197] Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 1163–1180. https://aclanthology.org/2022.acl-long.83

[198] 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] https://arxiv.org/abs/2406.00944

[199] Xiaojun Xu, Yuanshun Yao, and Yang Liu. 2024. Learning to Watermark LLM-generated Text via Reinforcement Learning. arXiv:2403.10553 [cs.LG] https://arxiv.org/abs/2403.10553

[200] 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).

[201] 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. Advances in Neural Information Processing Systems 36 (2024).

[202] Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2023. Backdooring instruction-tuned large language models with virtual prompt injection. arXiv preprint arXiv:2307.16888 (2023).

[203] Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023. Prca: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. arXiv preprint arXiv:2310.18347 (2023).

[204] 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. https://doi.org/10.1038/s44401-024-00004-1

[205] 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 on Artificial Intelligence, Vol. 36. 11613–11621.

[206] 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. https://doi.org/10.1016/j.hcc.2024.100211

[207] 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] https://arxiv.org/abs/2401.12794

[208] Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial Report Chunking for Effective Retrieval Augmented Generation. arXiv:2402.05131 [cs.CL] https://arxiv.org/abs/2402.05131

[209] 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 Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Berlin, Germany, 201–206.

[210] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. Advances in neural information processing systems 32 (2019).

[211] Kenji Yokotani, Gen Takagi, and Kobun Wakashima. 2018. Advantages of Virtual Agents over Clinical Psychologists during Comprehensive Mental Health Interviews Using a Mixed Methods Design. Computers in Human Behavior 85 (2018), 135–145. https://doi.org/10.1016/j.chb.2018.03.045

[212] KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust Multi-bit Natural Language Watermarking through Invariant Features. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2092–2115. https://doi.org/10.18653/v1/2023.acl-long.117

[213] KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2024. Advancing Beyond Identification: Multi-bit Watermark for Large Language Models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 4031–4055. https://doi.org/10.18653/v1/2024.naacl-long.224

[214] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. arXiv:2310.01558 [cs.CL] https://arxiv.org/abs/2310.01558

[215] Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-Driven Retrieval Augmented Response Generation for Online Misinformation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 5628–5643. https://doi.org/10.18653/v1/2024.naacl-long.313

[216] Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Kevi@inproceedingscohen2018understanding, title=Understanding the representational power of neural retrieval models using NLP tasks, author=Cohen, Daniel and O'Connor, Brendan and Croft, W Bruce, booktitle=Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, pages=67–74, year=2018 n Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Joanna Nelson, and William Hiesinger. 2023. Almanac: Retrieval-Augmented Language Models for Clinical Medicine. arXiv preprint arXiv:2312.10997 (2023).

[217] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? ACL 2019 (2019). https://doi.org/10.48550/arXiv.1905.07830 arXiv:1905.07830 [cs.CL]

[218] Shenglai Zeng, Jiankun Zhang, Pengfei He, Jie Ren, Tianqi Zheng, Hanqing Lu, Han Xu, Hui Liu, Yue Xing, and Jiliang Tang. 2024. Mitigating the Privacy Issues in Retrieval-Augmented Generation (RAG) via Pure Synthetic Data. arXiv preprint arXiv:2406.14773 (2024). https://doi.org/10.48550/arXiv.2406.14773

[219] He Zhang, Bang Wu, Xingliang Yuan, Shirui Pan, Hanghang Tong, and Jian Pei. 2024. Trustworthy Graph Neural Networks: Aspects, Methods, and Trends. arXiv preprint arXiv:2205.07424 (2024). https://arxiv.org/pdf/2205.07424

[220] 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 Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14447–14465. https://doi.org/10.18653/v1/2023.acl-long.808

[221] 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 33rd USENIX Security Symposium (USENIX Security 24). 1813–1830.

[222] 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.

[223] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv preprint arXiv:2305.14551 (2023).

[224] Yongfeng Zhang, Jiaxin Mao, and Qingyao Ai. 2019. Www'19 tutorial on explainable recommendation and search. In Companion Proceedings of The 2019 World Wide Web Conference. 1330–1331.

[225] 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 on Research & Development in Information Retrieval. 1411–1413.

[226] 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.

[227] 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 Proceedings of the 41st International Conference on Machine Learning (Vienna, Austria) (ICML'24). JMLR.org, Article 2473, 24 pages.

[228] 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. https://doi.org/10.1145/3637870

[229] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning Retrieval Corpora by Injecting Adversarial Passages. arXiv:2310.19156 [cs.CL] https://arxiv.org/abs/2310.19156

[230] Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in Retrieval-Augmented Generation Systems: A Survey. arXiv preprint arXiv:2409.10102 (2024).

[231] 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).

[232] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. arXiv:2402.07867 [cs.CR] https://arxiv.org/abs/2402.07867