Article

Consistency Gap under Retrieval Corruption: Stress-Testing RAG Robustness with Adversarial Evidence

Juyi Yang [1,*]

[1] University of California, Los Angeles, Los Angeles, USA
* Correspondence: Juyi Yang, University of California, Los Angeles, Los Angeles, USA

Abstract: Retrieval-Augmented Generation (RAG) significantly enhances the factual answering capability and contextual awareness of large language models by dynamically incorporating external knowledge sources into the generation process. However, the overall reliability of these systems critically depends on the accuracy and quality of the documents provided during the initial retrieval stage. When retrieval results are inadvertently or maliciously contaminated with adversarial evidence—defined as documents that are highly relevant to the user's query but contain fundamentally incorrect information—the model's generated content frequently exhibits systematic drift. This phenomenon forms what is formally termed a "consistency gap," wherein the model struggles to reconcile its internal parametric knowledge with the flawed external context. This paper investigates retrieval contamination as a primary research setting and constructs a comprehensive stress-testing framework based on adversarial evidence to systematically examine how erroneous documents affect the logical consistency of model outputs. Extensive empirical evaluations reveal several critical findings: under conditions of retrieval contamination, models exhibit a highly asymmetric sensitivity to incorrect evidence, often prioritizing flawed retrieved text over accurate internal knowledge. Furthermore, different decoding strategies demonstrate highly differentiated robustness characteristics, suggesting that generation parameters can mitigate or exacerbate the issue. Finally, the magnitude of the observed consistency gap is significantly correlated with an increased risk of severe hallucination. Ultimately, this work provides new analytical perspectives and robust testing methodologies for evaluating, benchmarking, and improving the resilience of retrieval-augmented generation systems in real-world applications.

Keywords: retrieval-augmented generation; robustness testing; consistency gap; adversarial evidence; retrieval contamination

Received: 06 February 2026
Revised: 27 March 2026
Accepted: 09 April 2026
Published: 13 April 2026

1. Introduction

Retrieval-augmented generation technology has emerged as a transformative approach in enhancing the capabilities of large language models, particularly in tasks that require extensive knowledge. This innovative method combines the strengths of retrieval systems with generative models, aiming to improve the accuracy and relevance of the information provided. The fundamental premise of this approach is that the retrieval phase can supply precise and pertinent documents, which in turn allows the generative model to formulate responses based on solid evidence. However, in real-world applications, the retrieval process may inadvertently include documents that, while semantically similar to the query, contain incorrect or misleading information. This presents a significant challenge: when the generative model is exposed to such adversarial evidence, its ability to produce outputs that align with accurate knowledge becomes a critical factor in assessing the system's robustness. To address this issue, this paper employs the analytical framework of the consistency gap to conduct a comprehensive stress test. This test evaluates the robustness of retrieval-augmented generation systems when faced with retrieval contamination, thereby providing insights into their reliability and effectiveness in maintaining consistency with correct information [1].

2. Foundations of Vulnerability in Retrieval-Augmented Generation Systems

2.1. Sources of Uncertainty and Contamination Risk in the Retrieval Stage

The reliability of a retrieval-augmented generation system is fundamentally dependent on the quality of documents retrieved during the initial retrieval stage. In practical deployment scenarios, retrieval modules encounter numerous sources of uncertainty. The knowledge base's completeness is inherently constrained; no knowledge base can comprehensively encompass all valid information within a specific domain. This limitation results in coverage blind spots within retrieval outcomes [2]. Additionally, the relevance ranking mechanisms employed by retrieval algorithms are prone to inherent errors. For instance, vector similarity-based retrieval methods may prioritize documents that are semantically similar but factually incorrect. Furthermore, discrepancies in the authority of external knowledge sources exacerbate fluctuations in retrieval quality, as open-domain knowledge bases frequently contain factual inaccuracies, outdated information, or biased perspectives. These factors collectively contribute to the challenges faced by retrieval-augmented generation systems in maintaining reliability and accuracy.

When multiple sources of uncertainty converge, the likelihood of adversarial evidence appearing in retrieval results increases substantially. Adversarial evidence is characterized by its high similarity to correct knowledge while being deceptive in key factual aspects. This makes it challenging for generative models to differentiate based on surface-level semantics. Retrieval contamination refers to the inclusion of such adversarial evidence within retrieval results, posing a fundamental challenge to the robustness of retrieval-augmented generation systems. The presence of adversarial evidence can significantly undermine the system's ability to generate accurate and reliable outputs, as it introduces misleading information that can be difficult to filter out in subsequent stages of the generation process.

These uncertainties do not operate in isolation; rather, they accumulate and reinforce each other, creating systemic risks. The incompleteness of the knowledge base, coupled with the inherent errors in ranking algorithms, establishes initial pathways for adversarial evidence to infiltrate retrieval results. Variations in source authority further obscure the boundaries of evidence quality. When adversarial evidence appears prominently in retrieval results with high similarity, the generative model struggles to effectively exclude it in later stages. This systemic vulnerability, originating from the retrieval stage, represents a distinct risk type for retrieval-augmented generation systems compared to traditional generative models. It affects the entire generation process, as the presence of adversarial evidence can lead to the propagation of inaccuracies throughout the system [2, 3]. Addressing these vulnerabilities requires a comprehensive approach that considers the interplay of various uncertainties and implements robust mechanisms to enhance the reliability and accuracy of retrieval-augmented generation systems.

2.2. Dependency Mechanism of Generative Models on Retrieved Content

Generative models engage in a sophisticated process that goes beyond merely replicating retrieved content. These models employ intricate mechanisms to comprehend, filter, and integrate information. Retrieved documents serve as contextual inputs, and through the use of attention mechanisms, the model assesses the relevance of various segments of these documents to the task at hand. This allows the model to extract essential information and incorporate it into its generative process [4]. While this dependency mechanism enhances the factual accuracy of the output, it also introduces potential vulnerabilities. The model's reliance on retrieved content can lead to challenges in maintaining consistency and accuracy, particularly when the retrieved information is not entirely reliable.

Biases in the allocation of attention can cause the model to disproportionately focus on information that is locally relevant, potentially at the expense of maintaining global factual consistency [5]. In scenarios where adversarial evidence is present, the model may mistakenly regard such evidence as credible, especially if it does not directly contradict

the model's existing knowledge. The extent to which the model trusts retrieved content is also shaped by the way prompts are engineered. When prompts are designed to emphasize reliance on retrieved documents, the model's dependence on this information increases, which can amplify the misleading effects of adversarial evidence. This highlights the importance of carefully crafting prompts to guide the model's focus appropriately.

Furthermore, when the model integrates information from multiple retrieved documents, it may exhibit what is known as the "collaborative misinformation effect." This occurs when incorrect information from a single document gains unwarranted credibility because it appears to be corroborated by multiple sources [6]. This complex dependency mechanism creates a paradox where improvements in factual accuracy coexist with vulnerabilities. The interplay of attention biases, prompt reinforcement, and collaborative misinformation makes it challenging to predict the model's behavior when faced with contaminated retrievals. This underscores the necessity for systematic stress testing to accurately assess the model's robustness and reliability in various scenarios. Such testing can help identify potential weaknesses and guide the development of strategies to mitigate these vulnerabilities.

2.3. Theoretical Connotation and Measurement Dimensions of the Consistency Gap

The concept of the consistency gap is pivotal in understanding how model outputs can vary when there are changes in the retrieval inputs. This gap is theoretically significant as it highlights the model's sensitivity and its potential vulnerability to incorrect information present in the retrieval inputs. The consistency gap emerges due to an inherent tension within the model's knowledge representation system. On one hand, the model is required to effectively utilize the contextual information it retrieves, while on the other hand, it must ensure that its outputs remain consistent with its inherent parametric knowledge. When there is a conflict between these two aspects, the model's inherent preferences play a crucial role in determining the consistency of its outputs. This dynamic is essential for understanding the model's behavior and its ability to handle conflicting information [7].

From a measurement standpoint, the consistency gap can be assessed through three distinct dimensions. The first dimension, content, is concerned with the changes in factual accuracy [8]. It involves comparing the frequency of incorrect facts in the model's outputs when subjected to clean versus contaminated retrieval conditions. The second dimension, structural, examines the stability of the model's reasoning chain. This involves assessing whether adversarial evidence has the potential to alter the fundamental reasoning framework or pathway of the model. The third dimension, confidence, evaluates the changes in the model's self-assessed confidence levels. This dimension analyzes whether the model demonstrates appropriate levels of uncertainty when confronted with conflicting information. These dimensions collectively provide a comprehensive framework for understanding the consistency gap.

The interrelation of these dimensions forms a holistic characterization of the consistency gap. A large consistency gap indicates a high sensitivity to erroneous inputs and suggests a lack of robustness in the model. Conversely, a small consistency gap signifies effective resistance to retrieval contamination, indicating a robust model. The multi-dimensional measurement framework allows for a precise diagnosis of the sources of vulnerability within the model. This can occur at various levels, such as fact selection, reasoning organization, or self-assessment. By identifying these specific areas, targeted improvements can be made to enhance the model's robustness. This approach provides clear objectives for strengthening the model's ability to withstand retrieval contamination and maintain output consistency.

3. Design Principles of Adversarial Evidence and Stress Testing Framework

3.1. Construction Methods and Classification of Adversarial Evidence

The construction of adversarial evidence is guided by two fundamental principles: high similarity and high deceptiveness. These principles ensure that the evidence can successfully pass retrieval ranking systems while effectively misleading the model. There are three primary types of adversarial evidence. Fact-replacement adversarial evidence involves systematically altering key facts within correct knowledge, such as changing the time, location, or entities involved. This type tests the model's sensitivity to factual details [6]. Logic-distortion adversarial evidence maintains the factual elements but modifies the logical relationships, such as reversing causality, to test the model's reasoning capabilities. Opinion-biased adversarial evidence introduces subjective interpretations into factual statements, challenging the model's ability to differentiate between facts and opinions. Each type of adversarial evidence serves a unique purpose in evaluating the robustness and accuracy of models, highlighting potential vulnerabilities in their processing and interpretation of information.

Adversarial evidence can be further categorized based on the intensity of its misleading nature, classified as either weak or strong. Weak adversarial evidence impacts minor details without significantly altering the overall framework of the answer. This type of evidence is subtle and may only slightly mislead the model, testing its ability to maintain accuracy despite minor discrepancies. On the other hand, strong adversarial evidence directly contradicts core facts, posing a more significant challenge to the model's integrity and accuracy. This type of evidence is designed to test the model's resilience against substantial misinformation. Each type of adversarial evidence influences models through distinct mechanisms, necessitating separate evaluations to understand their effects comprehensively. By examining these different types and intensities of adversarial evidence, researchers can gain insights into the strengths and weaknesses of models, ultimately contributing to the development of more robust and reliable systems.

3.2. Simulation of Retrieval Contamination Scenarios and Testing Procedures

The simulation of retrieval contamination is a complex process that involves carefully controlling both the proportion and the positioning of adversarial evidence within the retrieval system. This must be done while ensuring that the retrieval processes remain realistic and reflective of actual conditions. The testing procedure is designed to be progressive, beginning with a baseline stage that utilizes only correct documents to establish a reference point for performance. This stage is crucial as it sets the standard against which all subsequent stages are measured. Following this, the single-document contamination stage is introduced, where one adversarial document is added to the mix. This stage is designed to evaluate how the system responds to isolated conflicts and discrepancies. Finally, the multi-document contamination stage is implemented, which increases the amount of adversarial evidence to assess the cumulative and collaborative effects of multiple conflicting documents. This progression allows for a comprehensive evaluation of the system's robustness and adaptability in the face of varying levels of contamination.

The positioning of adversarial evidence within the retrieval results is also a critical factor that influences the allocation of attention by users. Evidence placed at the top of the retrieval results can have a significantly different impact compared to evidence placed at the bottom. This is because users are more likely to focus on information presented earlier in the list. To ensure the reliability of the results, each condition must undergo repeated trials to control for randomness and variability. During these trials, it is essential to record outputs and confidence distributions meticulously for subsequent analysis. The testing should encompass a variety of question types, including factual, reasoning, and explanatory questions. This comprehensive approach is necessary to thoroughly evaluate the system's robustness across different tasks, as each task type may exhibit unique dependency patterns and consistency gap characteristics. By covering a broad spectrum

of question types, the evaluation can provide a more complete picture of the system's strengths and weaknesses.

3.3. The Modulating Role of Decoding Strategies on Robustness

Decoding strategies play a crucial role in shaping the behavior of models when they encounter conflicting information. Greedy decoding, which involves selecting the highest-probability token at each step, is highly sensitive to input signals. This sensitivity makes it particularly vulnerable to adversarial influences, as it can easily be swayed by misleading inputs. On the other hand, sampling-based decoding introduces an element of randomness into the process. This randomness can help reduce the model's reliance on any single input signal, thereby potentially enhancing its robustness. However, this approach also increases the variability of the output, which can be a drawback in scenarios where consistency is desired. Beam search, another popular strategy, maintains multiple candidate paths simultaneously and selects the optimal one based on a global evaluation. This method can effectively filter out local conflicts, but it comes at the cost of higher computational demands [9, 10]. Each of these strategies has its own strengths and weaknesses, and the choice of which to use often depends on the specific requirements of the task at hand, such as the need for robustness versus computational efficiency.

Adversarial decoding is an emerging strategy that seeks to enhance model robustness by considering both the internal knowledge distribution of the model and the content retrieved during the decoding process. This approach aims to suppress the model's over-reliance on adversarial evidence, which can be misleading. While adversarial decoding offers increased robustness, it is inherently complex and may lead to a reduction in performance under clean conditions where adversarial influences are absent. The trade-off between robustness and efficiency is a critical consideration in the design of decoding strategies. Developers must carefully balance these factors to ensure that the model performs optimally across a range of scenarios. The complexity of adversarial decoding arises from its need to integrate multiple sources of information and make nuanced decisions about which signals to prioritize. This complexity can be a barrier to its widespread adoption, but for applications where robustness is paramount, it offers a promising avenue for further exploration and development.

4. Empirical Measurement and Behavioral Analysis of the Consistency Gap

4.1. Quantitative Characteristics of Output Drift

Under conditions of retrieval contamination, the phenomenon of output drift can be observed to follow distinct quantifiable patterns. Directionally, this drift tends to be asymmetric, with models more frequently transitioning from correct answers to those that are adversarially suggested and incorrect. In terms of magnitude, the drift is not a simple binary shift but rather a continuous spectrum, ranging from minor errors in detail to the complete adoption of incorrect frameworks. Regarding stability, outputs can vary significantly across different runs, suggesting that the decision-making processes of these models are non-deterministic. These characteristics are not uniform but vary depending on the specific architectures of the models and the training regimes they have undergone. This variability highlights the complexity of model behavior under different conditions and underscores the importance of understanding the underlying mechanisms that contribute to such inconsistencies. By analyzing these patterns, researchers can gain insights into improving model robustness and reliability.

4.2. Cumulative Effects and Threshold Phenomena of Erroneous Evidence

The impact of adversarial evidence on models is complex and cannot be simply summed up in a linear fashion. Instead, it demonstrates cumulative and threshold effects. Initially, when there is minimal contamination, models are able to rely on their parametric knowledge, maintaining a high level of accuracy. However, as the level of contamination increases, the quality of the output begins to fluctuate. This fluctuation is not random but

follows a pattern where, beyond a certain critical threshold, a significant change occurs. This change, often referred to as a "cognitive flip," results in models systematically adopting incorrect information. This phenomenon highlights the importance of understanding the limits of model robustness and the conditions under which they can be compromised. It also underscores the need for developing strategies to mitigate the effects of adversarial evidence, ensuring that models remain reliable even in the presence of such challenges.

The threshold at which this cognitive flip occurs is influenced by several factors, including the semantic similarity between the adversarial evidence and the model's existing knowledge, the internal consistency of the adversarial evidence itself, and the complexity of the queries being processed. The presence of these thresholds indicates that model behavior undergoes state transitions rather than a gradual adaptation process [11, 12]. This insight is crucial for understanding the boundaries of model robustness and for designing systems that can withstand adversarial attacks. By identifying and analyzing these thresholds, researchers can gain a deeper understanding of how models process information and where their vulnerabilities lie. This knowledge is essential for improving the design and implementation of models, ensuring they can operate effectively even when faced with challenging and potentially misleading information.

4.3. Inconsistency as an Early Warning Signal for Hallucination Risk

The concept of a consistency gap is crucial in understanding the risk of hallucinations in models. A significant consistency gap indicates a higher susceptibility to hallucinations, even when the input data is clean and seemingly reliable. This vulnerability arises because models with larger gaps tend to overly depend on the input context, which can lead to misinterpretations or errors when the input is ambiguous or misleading. Additionally, these models exhibit a diminished reliance on their internal parametric knowledge, which should ideally serve as a stabilizing factor. On the other hand, models that exhibit a smaller consistency gap are generally more adept at balancing internal knowledge with external input. This balance allows them to process information more accurately and reduces the likelihood of hallucinations. By understanding and addressing the consistency gap, developers can enhance model reliability and performance, ensuring that the models are better equipped to handle a variety of inputs without succumbing to hallucinations [13].

Evaluating models based on the consistency gap offers a valuable complement to traditional confidence-based methods for detecting hallucinations. While confidence metrics provide an indication of a model's certainty in its outputs, they do not necessarily ensure the correctness of those outputs. The consistency gap, however, offers insight into how sensitive a model is to variations in input data. By combining both consistency gap analysis and confidence metrics, researchers and developers can achieve a more comprehensive understanding of a model's risk profile regarding hallucinations. This dual approach allows for more accurate predictions of potential hallucination occurrences, particularly in real-world applications where input data can be unpredictable and varied. Stress testing models under different conditions can yield practical insights into their expected behavior, enabling developers to refine and optimize models for deployment in diverse environments. This holistic evaluation strategy is essential for advancing the development of robust and reliable models that can operate effectively in complex, real-world scenarios.

5. Systematic Examination of Robustness Enhancement Strategies

5.1. Contamination Filtering at the Retrieval Level

Enhancing the robustness of information retrieval systems is a multifaceted process that begins with effective filtering at the retrieval stage. This involves employing multi-path retrieval techniques, which compare outputs from various systems to identify the most reliable documents. By doing so, it becomes possible to cross-verify information and

reduce the likelihood of relying on erroneous data. Additionally, re-ranking techniques are employed to further refine the selection of documents [9, 10]. These techniques utilize validation models that assess the quality of documents based on several criteria, including their authority, consistency, and alignment with established knowledge. This ensures that the information being retrieved is not only accurate but also credible. Furthermore, the curation of knowledge bases plays a crucial role in maintaining data quality at the source. By meticulously organizing and updating these databases, the integrity of the information is preserved. Metadata annotation, such as timestamps, authorship details, and credibility scores, also supports dynamic prioritization of information, allowing for more informed decision-making. Although these strategies do not completely eliminate the presence of adversarial evidence, they significantly mitigate its impact, thereby enhancing the overall reliability of the system.

5.2. Evidence Balancing Mechanisms at the Generation Level

During the generation stage, evidence balancing mechanisms play a crucial role in enhancing the reliability and accuracy of models. These mechanisms allow models to effectively compare and weigh multiple sources of information. One key aspect is the credibility assessment, which involves evaluating the relevance of information, ensuring internal consistency, and aligning it with the model's existing knowledge base. This process helps in filtering out unreliable data and focusing on more credible sources. Additionally, conflict detection is an essential capability that enables models to identify contradictions within the data, allowing them to select the most reliable evidence. Source attribution is another important feature that enhances transparency by clearly indicating the origin of information, thereby encouraging cautious and informed generation of content. Furthermore, confidence calibration is vital as it ensures that models express uncertainty appropriately, which is crucial for maintaining the integrity of the generated content. These capabilities can be significantly improved through targeted fine-tuning using datasets rich in conflicts, thereby enhancing the model's ability to handle complex and contradictory information effectively.

5.3. Evaluation Systems and Continuous Monitoring Frameworks

A comprehensive evaluation system is essential for ensuring the robustness and reliability of technological frameworks. Such a system should encompass metrics for adversarial robustness, domain generalization, and output stability. Adversarial robustness refers to the system's ability to withstand and function correctly under malicious attacks or unexpected inputs. Domain generalization ensures that the system performs well across various environments and conditions, not just those it was specifically trained on. Output stability metrics assess the consistency and reliability of the system's outputs over time. Continuous monitoring transforms evaluation from a one-time assessment into an ongoing process. This involves regularly sampling real-world outputs and using the results to inform system improvements. When vulnerabilities are detected, they can be addressed through adversarial training, which involves exposing the system to challenging scenarios to enhance its resilience. Anomaly detection mechanisms are crucial in this framework, as they trigger alerts when there are significant fluctuations in consistency metrics. This closed-loop system is designed to ensure that the system's robustness improves continuously over time, adapting to new challenges and maintaining high performance standards [6].

6. Conclusion and Future Directions

6.1. Findings and Theoretical Contributions

This paper systematically studies the consistency gap under retrieval contamination and proposes a stress-testing framework based on adversarial evidence. The research reveals asymmetric vulnerabilities, continuous and threshold-based drift patterns, and significant differences across decoding strategies. The consistency gap is shown to be

predictive of hallucination risk, highlighting the importance of understanding these dynamics in the context of retrieval-augmented generation (RAG) systems. By identifying these vulnerabilities, the study provides a foundation for developing more robust systems that can withstand adversarial conditions. The findings emphasize the need for a comprehensive approach to evaluating system performance, one that goes beyond static measures of correctness to include dynamic behavioral stability. This approach is crucial for anticipating and mitigating potential risks associated with hallucinations, which can undermine the reliability of AI systems. The insights gained from this study are not only theoretical but also have practical implications for the design and deployment of RAG systems.

Theoretically, the consistency gap framework expands robustness evaluation from static correctness to dynamic behavioral stability. This shift in focus is significant as it acknowledges the complex and evolving nature of AI systems. The classification and construction of adversarial evidence provide reusable methodological tools that can be applied across different contexts, enhancing the versatility and applicability of the framework. The stress-testing framework reflects a shift toward multi-dimensional and dynamic evaluation, which is essential for capturing the nuanced behaviors of AI systems under various conditions. By incorporating these elements, the framework offers a more comprehensive understanding of system robustness, paving the way for future research that can build on these foundational concepts. This theoretical advancement is crucial for developing AI systems that are not only accurate but also resilient to adversarial influences, ensuring their reliability and trustworthiness in real-world applications.

6.2. Practical Implications and Future Research

Practically, robustness should be treated as equally important as accuracy in system design. This perspective necessitates the implementation of defensive mechanisms at both retrieval and generation stages to safeguard against potential vulnerabilities. Consistency gap testing should be incorporated into pre-deployment evaluation to ensure that systems are adequately prepared to handle adversarial conditions. This proactive approach is vital for maintaining the integrity and reliability of AI systems in diverse applications. By prioritizing robustness, developers can create systems that are better equipped to withstand the challenges posed by adversarial evidence, ultimately enhancing their performance and trustworthiness. This emphasis on robustness is particularly important as AI systems become increasingly integrated into critical domains, where the consequences of failure can be significant. Therefore, a robust evaluation framework is essential for ensuring the long-term success and reliability of AI technologies.

Future research may explore automated adversarial evidence generation, cross-lingual and multimodal robustness, and long-term real-world validation. These areas represent promising directions for advancing the field of AI robustness. Automated adversarial evidence generation could streamline the process of identifying vulnerabilities, making it easier to test and improve system resilience. Cross-lingual and multimodal robustness are critical for ensuring that AI systems can perform reliably across different languages and modalities, broadening their applicability and effectiveness. Long-term real-world validation is essential for assessing the sustained performance of AI systems in dynamic environments, providing valuable insights into their long-term reliability. As RAG systems become widely adopted, continued research on robustness will be essential for building reliable and trustworthy knowledge-enhanced AI systems. This ongoing research will help to address emerging challenges and ensure that AI technologies continue to evolve in a manner that prioritizes safety and reliability.

6.3. Conclusion

This paper examines the consistency gap of retrieval-augmented generation systems under retrieval contamination. Results show asymmetric drift when adversarial evidence is introduced, significant variation across decoding strategies, and a strong link between

consistency gaps and hallucination risk. The cumulative effect of adversarial evidence reveals a threshold-driven cognitive flip phenomenon, providing important insights into the robustness boundaries of RAG systems. These findings underscore the complexity of AI system behavior under adversarial conditions and highlight the need for comprehensive evaluation frameworks that can capture these dynamics. By understanding the factors that contribute to hallucination risk, researchers and developers can devise strategies to mitigate these risks, enhancing the reliability and trustworthiness of AI systems. The study's insights into the cognitive flip phenomenon offer a new perspective on how adversarial evidence can impact system performance, paving the way for future research that can further explore these dynamics and develop more robust AI technologies.

References

1. S. Amirshahi, A. Bigdeli, C. L. Clarke, and A. Ghenai, "Evaluating the Robustness of Retrieval-Augmented Generation to Adversarial Evidence in the Health Domain," arXiv preprint arXiv:2509.03787, 2025.
2. F. Fang, Y. Bai, S. Ni, M. Yang, X. Chen, and R. Xu, "Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Aug. 2024, pp. 10028-10039.
3. Y. Tu, W. Su, Y. Zhou, Y. Liu, and Q. Ai, "Robust Fine-tuning for Retrieval Augmented Generation against Retrieval Defects," in Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 2025, pp. 1272-1282.
4. M. V. D. S. de Oliveira, J. D. A. Silva, and A. D. L. Fontao, "Fairness Testing in Retrieval-Augmented Generation: How Small Perturbations Reveal Bias in Small Language Models," arXiv preprint arXiv:2509.26584, 2025.
5. Y. Tu, W. Su, Y. Zhou, Y. Liu, and Q. Ai, "RbFT: Robust Fine-tuning for Retrieval-Augmented Generation against Retrieval Defects," arXiv preprint arXiv:2501.18365, 2025.
6. H. Zhou, K. H. Lee, Z. Zhan, Y. Chen, Z. Li, Z. Wang, et al., "TrustRAG: enhancing robustness and trustworthiness in retrieval-augmented generation," arXiv preprint arXiv:2501.00879, 2025.
7. Y. Zeng, T. Cao, D. Wang, X. Zhao, Z. Qiu, M. Ziyadi, et al., "Rare: Retrieval-aware robustness evaluation for retrieval-augmented generation systems," arXiv preprint arXiv:2506.00789, 2025.
8. C. Sharma, "Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers," arXiv preprint arXiv:2506.00054, 2025.
9. T. Sun, A. Somalwar, and H. Chan, "Multimodal retrieval augmented generation evaluation benchmark," in 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), Jun. 2024, pp. 1-5.
10. S. Perçin, X. Su, Q. S. Syed, P. Howard, A. Kuvshinov, L. Schwinn, and K. U. Scholl, "Investigating the robustness of retrieval-augmented generation at the query level," in Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²), Jul. 2025, pp. 439-457.
11. Y. Zhou, Y. Liu, X. Li, J. Jin, H. Qian, Z. Liu, et al., "Trustworthiness in retrieval-augmented generation systems: A survey," arXiv preprint arXiv:2409.10102, 2024.
12. J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 16, Mar. 2024, pp. 17754-17762.
13. S. Yang, J. Wu, W. Ding, N. Wu, S. Liang, M. Gong, et al., "Quantifying the robustness of retrieval-augmented language models against spurious features in grounding data," arXiv preprint arXiv:2503.05587, 2025.