Retrieving Facts or Reinforcing Beliefs? Detecting and Quantifying Confirmation Bias in RAG Systems for Scientific Claim Verification

Anonymous EMNLP submission

Abstract

Retrieval-Augmented Generation (RAG) models have emerged as powerful tools for information-seeking tasks across domains. However, their reliance on external retrieval mechanisms introduces new pathways for bias that remain underexplored. In this work, we present ConfirmBiasRAG, a new benchmark designed to systematically evaluate confirmation bias in RAG pipelines. Unlike previous efforts that focused solely on model outputs, our approach decomposes the RAG process to investigate three critical components: (1) the degree to which the retriever introduces biased evidence, (2) how the reranker may further amplify such bias, and (3) to what extent the final generation is steered by the retrieved evidence. We construct 270 original/counter claim pairs using a red-teaming-inspired approach in the scientific domain, a setting where subtle differences can reinforce prior beliefs stated in queries. By analyzing model responses and their stance or belief alignment with input prompts, we reveal that multiple state-of-the-art RAG systems exhibit confirmation bias among these three stages, with the reranker often reinforcing biases introduced during retrieval. Our benchmark enables fine-grained diagnosis of confirmation bias in RAG pipelines and offers a foundation for developing more robust and fair information-seeking systems.

1 Introduction

Confirmation bias, the tendency to seek or interpret evidence in ways that affirm existing beliefs, poses a significant threat to scientific integrity. It can lead researchers to unintentionally favor supportive evidence, resulting in selective interpretation and reporting. This, in turn, reinforces flawed assumptions and conceals contradictory findings, undermining the objectivity and reliability of scientific research Stewart (2024).

At the same time, the rapid growth of scientific literature has made efficient verification of scientific claims increasingly challenging Wadden et al. (2022). Retrieval-Augmented Generation (RAG), which combines retrieval mechanisms with powerful Large Language Models (LLMs), provides a scalable and contextually relevant solution. Recent systems, such as OpenScholar Asai et al. (2024) and ScholarQA Singh et al. (2025), ground responses in explicitly citetd sources to generate more accurate summaries with transparent attribution to original documents. While the recent work Wong et al. (2025) qualitatively identifies confirmation bias in proprietary RAG systems within medical domains, it remains largely underexplored how to automatically detect and quantify such biases in RAG models tailored specifically for scientific claim verification.

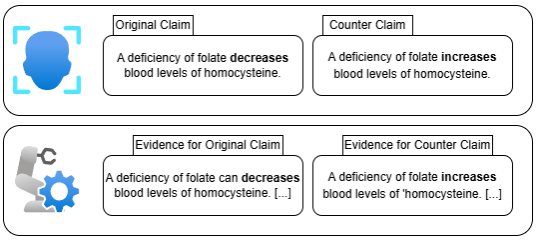

Figure 1: A demonstration of confirmation bias in the responses of OpenScholar.

Inspired by red teaming Perez et al. (2022), we introduce a novel benchmark explicitly designed to measure confirmation bias in RAG systems. As illustrated in Fig. 1, our benchmark is built using carefully crafted pairs of claim and counter-claim that differ only in one perspective, enabling us to evaluate the extent to which RAG models support both sides. To quantify confirmation bias in responses generated across 270 claim pairs, we propose the Confirmation Bias Rate (CBR), the percentage of claim pairs for which both the claim and its counter-claim are supported, and analyze how it varies across different RAG systems.

To investigate how confirmation bias is triggered or amplified within a RAG system, we explore

three hypotheses: (1) the retrieval component may surface documents that reinforce the framing of a claim Moreira et al. (2024); noa, (2) re-ranking mechanisms, often tuned for relevance or alignment, may intensify this effect, and (3) the generation component may hallucinate information even when grounded in relevant evidence. Our extensive experiments on recent RAG systems yield the following findings:

1. Measuring with Confirming Ratio (CR) a novel quantitative method we purposed, retrievers introduce a significant amount of biased documents into the system. Destroy the balance of the evidence corpus.

2. Rerankers' behavior are context dependent, from our measuring they often have either minimum effects for reducing or amplifying biased evidence, while occasionally enhance the chance of confirmation bias in special context.

3. Our measuring suggests the inference of context for LLM generation has limited effects. While hallucinations can dicetate the final generation.

These findings highlight the urgent need for automatic evaluation tools that go beyond verification accuracy and engage with the deeper epistemic risks of AI-driven knowledge access Shi et al. (2025). We position our benchmark as the first step toward closing this gap, providing researchers and developers with a tool to better understand and mitigate confirmation bias before it leads to harmful real-world outcomes.

2 Background and Related Work

Recent work has highlighted how cognitive and algorithmic biases shape information access in AI systems, raising critical concerns about fairness, transparency, and epistemic integrity. The Prompt Association Test (P-AT; Onorati et al. (2023)) introduced a framework for detecting implicit social biases in word representations, providing early insights into how bias manifests in static language models. Building on this foundation, researchers have extended bias detection techniques to more complex generative and retrieval-based systems. For example, SciFact-Open Wadden et al. (2020), FAIR Gao et al. (2022), and SciClaimHunt Kumar et al. (2025) propose methods for evaluating scientific factuality and claim verification, helping assess whether retrieval-augmented models return reliable information. Despite such advances, concerns persist around cognitive biases, especially confirmation bias—the tendency to seek or prioritize information that supports pre-existing beliefs.

Several studies have examined how this bias permeates retrieval and generation. Deffuant et al. (2025) model how exposure to a large volume of seemingly diverse, yet ideologically consistent information can lead to the reinforcement of extreme beliefs, mimicking real-world echo chambers. In parallel, Kacperski et al. (2023) demonstrate that search platforms like Semantic Scholar not only reflect but can amplify user biases through algorithmic feedback loops. Similarly, Gomroki et al. (2023) show that confirmation bias can directly influence what is deemed relevant in information retrieval systems, further skewing access to balanced perspectives. Recent work by Sharma et al. (2024) and again by Wang et al. (2023) underscores how large language model (LLM)-powered systems deepen selective exposure, as they learn to align their outputs with the user's implicit or explicit preferences.

Efforts to address these challenges include systems such as OpenScholar Asai et al. (2024) and ScholarQA, which aim to improve factual grounding in retrieval-augmented scientific question answering. However, while these models show promise in increasing factual accuracy, they often sidestep the deeper issue of cognitive bias embedded in user interactions and retrieval dynamics. Our work builds on this growing body of literature by proposing a targeted benchmark to directly measure confirmation bias in RAG systems. Through controlled prompt engineering and stance analysis, we offer a systematic method to evaluate how these models may unintentionally reinforce user beliefs—contributing to a more nuanced understanding of bias in modern AI.

2.1 RAG for Scientific Literature Review

RAG systems introduce more options for researchers during scientific literature review Singh et al. (2025); Asai et al. (2024), where they ground the user input with context documents from trustworthy sources in an efficient manner. Traditional RAG systems use a retriever to gather context and a LLM for downstream tasks, typically summarization or question answering (QA). They have quite a

bit of limitations such as the relevance of retrieved context to the user input and lack of quality control. These limitations often require large-scale retrieval and some summarization done by LLMs to reduce the size of the context set. This process becomes inefficient and introduces noises from LLM hallucinations. With rerankers introduced into RAG pipelines, such limitations claims to be solved effectively. Rerankers' function typically ranks the retrieved context by considering relevance to the user input and trustworthiness of the sources Moreira et al. (2024). This allows RAG systems to use the most promising set of context for downstream tasks therefore avoiding unneeded noises with improved efficiency.

3 Dataset

3.1 Red-teaming in LLMs

Modern LLMs are designed to be harmless and truthful through various instruction tuning strategies, often guided by human feedback. These methods aim to align LLMs with preferred output content and formats. Typically, red-teaming strategies are applied during inference, where LLMs are attacked via prompt engineering to expose vulnerabilities Perez et al. (2022). Automated prompt attacks—generated by other LLMs—have successfully jailbroken models, revealing limitations in current alignment methods Li et al. (2025). However, existing red-teaming efforts largely focus on broad categories of harmful content, neglecting domain-specific vulnerabilities that can have significant consequences. One such overlooked domain is scientific information retrieval, where subtle forms of bias like confirmation bias can undermine factual accuracy.

3.2 Confirmation bias in information retrieval

Confirmation bias is one of the fundamental cognitive bias where the end users try to seek information that confirm their believes and ignore contradiction evidence. Online environments present unique challenges that can amplify confirmation bias in information retrieval. Research has shown that confirmation bias is particularly strong in online settings and may be more pronounced than in traditional offline contexts Deffuant et al. (2025). Confirmation bias is often caused by crafting queries that assumes certain believes are correct. Furthermore, modern search engines focus on matching the semantic meaning of the input with the evidence which enhances the chance of confirmation bias Kayhan (2015); Kacperski et al. (2023). Confirmation bias can also caused by the end users dismissing credible sources that contradict their believes, by clicking on confirming titles.

To evaluate confirmation bias in RAG systems, we construct a domain-specific benchmark using a red-teaming-inspired strategy focused on the scientific literature domain. The dataset is organized into two columns and contains 271 rows, corresponding to 270 carefully curated calim pairs. With the first row indicating the headings. Each pair includes a query that assumes a certain belief (Original Claim) and its opposing counterpart (Counter Claim).

3.3 Data Collection Pipeline

Our dataset construction follows a multi-step pipeline illustrated in Figure 2. We begin by using Claude to generate candidate scientific queries. Low-quality or implausible claims are filtered out in a collection phase to maintain domain relevance. Each remaining claim is then evaluated using GPT-4o to ensure the presence of real-world, scientifically grounded terminology. This step is guided by manual annotation of grounding documents to exclude hallucinated or fictional claims.

Once validated, each accepted claim is paired with a semantically opposing version, also generated using an LLM. The resulting claim pairs are then passed through a traditional RAG pipeline (without reranking) on the OpenScholar platform to gather system responses. These responses are analyzed using LLM-based confirmation bias detection techniques, which identify instances where a system selectively affirms a belief in the absence of opposing evidence.

To ensure the benchmark's quality and reliability, we perform a final manual annotation step to validate bias presence and belief correctness. Only high-agreement samples (with at least 73% annotator consensus and full agreement on the selected benchmark subset) are retained in the final dataset.

This process results in a benchmark suitable for quantifying confirmation bias across the retrieval, reranking, and generation stages of RAG systems.

4 Methodology

We use a red-teaming style strategy to find patterns that will trigger confirmation bias for different RAG systems. With these patterns, we purpose a benchmark dataset to measure confirmation bias.

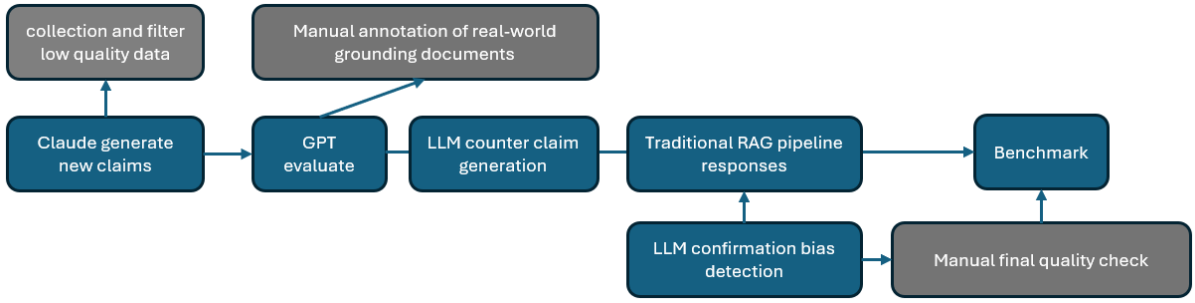

Figure 2: This figure shows the pipeline of collecting and building the dataset, indicating the phases where automated quality control and manual quality controls are conducted

4.1 The benchmark

We developed a benchmark designed to measure the confirmation bias of a given model based on its responses to carefully constructed pairs of queries. We built the dataset focused within the scientific literature field for effective evaluation. The reasons for choosing the scientific field is because previous study shows that scientific claims are hard to verify by LLMs with their internal knowledge due to the complexity, which became one of the reasons for researching RAG systems. Secondly, scientific claims are highly dependent on context, allow us to engineer the claim pairs for diversity. To evaluate the effectiveness of this evaluation suite, we conducted experiments on OpenScholar, ChatGPT-search and the ScholarQA platforms. We specifically tested the RAG systems through public-facing, instruction-following interfaces, as our main interest lies in detecting and quantifying bias in the contexts where it should not appear—namely, in systems explicitly designed for public research purposes. These interfaces are intended to be neutral and reliable, making them a critical point of analysis for unintended bias.

4.2 Prompt Engineering Explanation

We constructed the dataset as pairs of sentences with opposing semantic meanings and stances, by changing only one critical relation between the entities. We then submit those queries to the RAG systems to generate two separate responses. We evaluate the responses as "confirmation biased" if, regardless (or almost regardless) of the claim in the query, both responses align with their respective query. This behavior suggests that the model gives more weight to the direction of the user input rather than the factual accuracy or comprehensiveness of the content. Behavior often highlighting only support or only refute elements depending on the claim's belief or stance, even when the facts could potentially contradict or nuance the expressed thesis, and failed to include comprehensive view points.

4.3 Measurements of Confirmation Bias

With a result-driven detection methodology, we review the final generation results from various RAG systems to measure confirmation bias. Additionally, we analyze the retrieval and re-ranking components of open-source RAG pipelines to identify which part introduces or amplifies confirmation bias.

To answer our research questions, we need a way to evaluate the balance of retrieved documents across a claim pair. We hypothesize that an unbiased retriever and reranker will return a similar distribution of documents for both sides of each claim pair. As an initial approach, we compute the document overlap between original and counter claims using OpenScholar's retrieval outputs.

As shown in Figure 3, this distribution suggests that many claim pairs share overlapping retrieved documents. However, this overlap does not provide insight into how many documents actually support or refute the respective queries.

To address this, we introduce two metrics: Confirming Ratio (CR) and Confirmation Bias Rate (CBR). Prier these two metrics we first define Support Ratio (SR) and Refute Ratio (RR), where $\mathrm{RR} = 1 - \mathrm{SR}$. These are computed for each claim in a pair and quantify the proportion of documents that support or refute the claim. For a claim q at stage s (retrieval or reranking), SR is defined as:

$$\mathrm{SR}_q^{(s)} = \frac{\sum_{d \in D^{(s)}} \mathbb{I}_{\text{evidence}}(q, d)}{|D_e^{(s)}|} \quad (1)$$

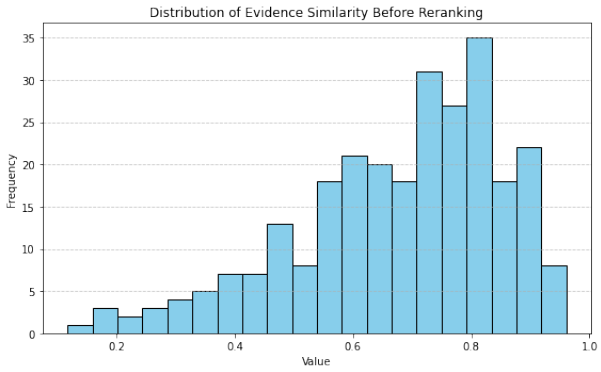

Figure 3: Distribution of document overlap ratio: original vs counter claims

Here, $D^{(s)}$ is the set of documents retrieved at stage s, and $D_e^{(s)} \subseteq D^{(s)}$ includes only documents labeled as supporting or refuting the claim. The indicator function $\mathbb{I}_{\text{evidence}}(q, d)$ is defined as:

$$\mathbb{I}_{\text{evidence}}(q, d) = \begin{cases} 1 & \text{if } d \text{ supports } q \\ 0 & \text{if } d \text{ refutes } q \end{cases}$$

Using SR and RR, we define a new aggregate metric, Confirming Ratio (CR), to measure asymmetry between claim pairs. It is calculated as:

$$\text{CR} = \frac{\sum_{i=1}^{N} |SR_o^{(i)} - RR_d^{(i)}| + |SR_d^{(i)} - RR_o^{(i)}|}{2N}$$

$SR_o^{(i)}$: Support Rate for the original claim$_i$

$RR_d^{(i)}$: Refute Rate for the counter claim$_i$

$SR_d^{(i)}$: Support Rate for the counter claim$_i$

$RR_o^{(i)}$: Refute Rate for the original claim$_i$

N : Total number of claim pairs

CR provides an average bias directionality between a claim and its counterpart, capturing the extent to which retrieval and reranking skew toward supporting each claim individually. A higher CR indicates imbalance and potential confirmation bias introduced by the RAG components.

To measure how often confirmation bias occurs system-wide, we use the Confirmation Bias Rate (CBR), defined as:

$$\text{CBR} = \frac{|\{(q_o, q_r) \in Q_{\text{pairs}} \mid \mathbb{I}_{CB}(q_o, q_r) = 1\}|}{|\{(q_o, q_r) \in Q_{\text{pairs}} \mid A(q_o) \wedge A(q_r)\}|} \quad (2)$$

Where $A(q)$ indicates whether a valid answer was retrieved for claim q, and $\mathbb{I}_{CB}(q_o, q_r)$ is an indicator function:

$$\mathbb{I}_{CB}(q_o, q_r) = \begin{cases} 1 & \text{if both}(q_o, q_r) \text{ are supported} \\ 0 & \text{otherwise} \end{cases}$$

Together, CR and CBR allow us to study both localized and system-level confirmation bias and identify which component—retriever, reranker, or generator—contributes to bias emergence.

5 Experiments

5.1 Experimental Setup

We first experimented manually with GPT-search and ScholarQA-Claude on a subset of 70 claim pairs from our dataset. The outputs were labeled using the definitions of $\mathbb{I}_{CB}(q_o, q_r)$ and $\mathbb{I}_{\text{evidence}}(q, d)$, both of which rely on LLMs for decision-making. To understand the capability of LLMs for making decisions on whether a claim is supported when given some evidence in the scientific literature domain. To understand the deci We test different LLMs using the Scifact dataset Wadden et al. (2020), a popular dataset with the structure of a claim, ground-truth label and corresponding expert evaluated evidence to support or refute the claim. We experiment with the QwQ-32b Yang et al. (2024) and ollama-phi4 Singh et al. (2025) for the decision making task we want for automated detection of CB and concluded a 90% accuracy for the Scifact dataset. Furthermore, to validate more of the decision-making ability of LLMs, we reused the sub-set of 70 claim pairs we collected in the methodology section. We compute the agreement between our automated detection process and human annotation and the score is 84.2% for QwQ-32b and 81.4% for ollama-phi-4 respectively. All the experimental results of the confirmation bias detection task reported in this paper were conducted by QwQ-32b. This pilot study also helped us refine prompt strategies for automated confirmation bias detection. After validating the setup on the subset, we extended the evaluation to the full dataset using OpenScholar, ScholarQA-GPT to better understand how different systems behave in terms of confirmation bias.

Implementation Details We selected these two models because they have remarkable reasoning and context understanding capabilities. We try to

System Name	Setting	Confirmation Bias Rate
ScholarQA_GPT	Full set automated evaluation	0.1716
ScholarQA_Claude	Subset human evaluation	0.1857
OpenScholar	Full set automated evaluation	0.214
OpenScholar	Subset human evaluation	0.2714
ChatGPT-search	Subset human evaluation	0.4305

Table 1: Main results comparing different RAG systems. Subset = 70 claim pairs; Full set = 270 claim pairs. Human evaluation uses manual labeling; automated evaluation uses LLM-based decision functions.

keep the output consistent with a temperature of 0 for those models and used the **chat** function to generate responses. The prompt for SR and confirmation bias detection are in Appendix A.3.

5.2 Results and Discussion

RQ1: How confirmation biased are the RAG systems using our dataset to measure? We evaluate several RAG-based systems with our proposed dataset and record their confirmation bias rates in Table 1. Among all systems, ChatGPT-search exhibits the highest confirmation bias rate, suggesting a strong tendency to generate responses that align with the user's prompt rather than critically engage with the retrieved evidence. Besides, ScholarQA-Claude and ScholarQA-GPT have the lowest bias rates, 0.1857 and 0.1716, respectively, indicating a comparatively higher alignment with evidence and reduced bias. Interestingly, systems evaluated on the full dataset using automated methods (e.g., ScholarQA-GPT and OpenScholar) generally show lower bias than those assessed via human evaluation on subsets, though this is not universally consistent. These findings suggest that both the system architecture and the evaluation methodology significantly impact the degree of confirmation bias observed in RAG pipelines. To distinguish our dataset from traditional claim verification datasets, we avoided using a single label as our measurement.

RQ2: How much does Retriever introduce confirmation bias? We hypothesize that CR should be close to 0 during retrieval to have a balanced corpus, indicating for each claim pair the corresponding corpus should share most of the documents. When CR is greater than 0 it is indicating more documents are supporting the user query which may lead to potential confirmation bias. We can see from Table 2 both systems have CR significantly above 0. It aligns with our hypothesis that during retrieval there are more documents supporting the input claim, causing a unbalance corpus to potentially cause confirmation bias.

System Name	Retrieve	Rerank	TopK
OpenScholar	0.4422	0.4386	0.4420
ScholarQA	0.4367	0.3913	0.4273

Table 2: CR results where TopK indicates CR for topk retrieved documents

System Name	Retrieve	Rerank	TopK
OpenScholar	0.8327	0.6970	0.7934
ScholarQA	0.8331	0.6918	0.7555

Table 3: Netural Document proportion where TopK indicate the Netural Document rate in the topK retrieved documents

We then dive deeper to understand if the amount of neutral documents during retrieval is different. That is the amount of documents contains background information and etc. These documents contribute to the scientific context understanfing. If OpenScholar has significantly less neutral documents comparing to ScholarQA we can not conclude ScholarQA's retriever introduces more bias. The results in Table 3 showcase they share similar amount of neutral documents in general, which showcases the retriever of ScholarQA introduces more potential bias into the system than OpenScholar's retriever.

RQ3: Does reranker enhance the chance of confirmation bias? Following the same procedure, we measured CR after reranking. We hypothesize that ideally during reranking the distribution of CR should be going towards 0 indicting the reranker is trying to balance perspective coverage in the corpus. When comparing CR before and after reranking, there's a significant drop in the value for ScholarQA, but there is no significant drop in CR for OpenScholar. To investgate further, shown in Table 3 the amount of neutral documents decreases sig-

nificantly for both systems, indicating the portion of potential confirmation bias causing documents may increase.

We also did qualitative evaluation manually on the same sub-set we used in the Experimental setup section to understand how rerankers' behavior in the context of ehancing the chance of confirmation bias. Rerankers will typically rank documents specifically supporting the input claim to higher positions if such document exist as shown in Figure 4. RAG systems with reranker consider the order of the documents in their prompt designing, instructing LLMs to use the top documents as the more reliable context for generation. Therefore when the special cases we discovered during qualitative evaluation happens reranker will participate in a role of enhancing confirmation bias. We conclude that rerankers behavior is highly context dependent which means the hypothesis rerankers enhance the chance of confirmation bias is rejected.

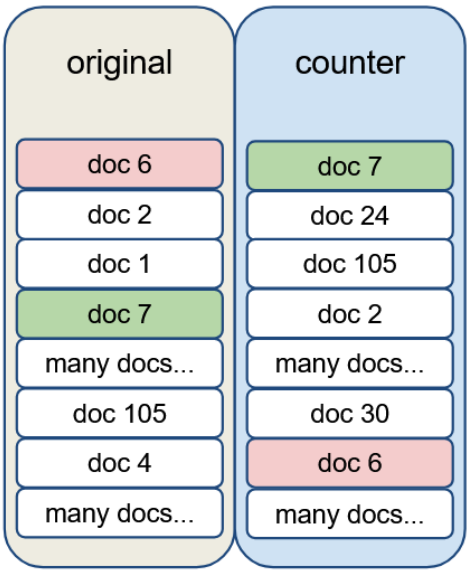

Figure 4: the id behind doc is their initial position before reranking. The red-colored doc 6 indicates the one document specifically support the original claim, where the green-colored doc 7 is the document specifically support the counter claim.

RQ4: The generation part in RAG systems, how much is it inferred by the evidence using CR measuring approaches? We investigate the extent to which the generation component of RAG systems is grounded in the retrieved evidence, using CR. We discovered that the final response from LLMs for claim pairs with high CR for both perspectives is more likely to be confirmation bias as excepted. But as seen in Figure 5 and Figure 6, where each red dot indicated the claim pairs that is confirmation biased. We represent these scatterplots using SR for both perspectives of the claim pair. The top right area basically align with our hypothesis, because both sides of the claim pairs are being supported by their corresponding evidence, thus lead to confirmation bias.

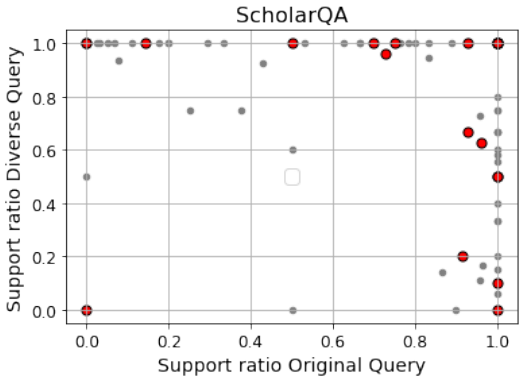

Figure 5: Red dots represent confirmation bias claim pairs we detected and this scatterplot reflects their distribution using SR of both perspectives as x and y axes respectively

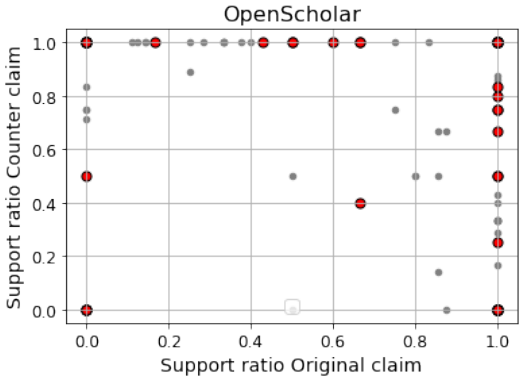

Figure 6

The outliers at the top-left and the bottom-right corners reveal that, while some portion of the generation is clearly informed by the evidence, a significant fraction of the content, particularly in responses to biased or leading prompts, extends beyond the retrieved sources. We already saw that LLMs not considering all the evidence in their generation step Ok et al. (2025), which could be one of the causes for this behavior. This suggests that the generator often extrapolates or hallucinates information, which can reinforce the frame of the prompt rather than strictly adhering to the evidence.

This behavior underscores the importance of evaluating not just the accuracy of the facts but also the epistemic alignment between the retrieval and generation of the RAG pipelines.

5.3 Discussion

In this section, we explain some design choices we made and discuss findings that emerged during manual annotation and evaluation. Our observations also point toward several broader challenges in measuring and interpreting confirmation bias in RAG systems.

5.4 CR Measurement

We theorized that CR should be around 0 for a perfect retrieval, where SR for the original claim and RR for the counter claim should be very close. However, our experiments revealed this assumption to be overly simplistic. In many cases, a single highly persuasive document can dominate the generation output, rendering the overall corpus CR less informative. This suggests that CR, in its current form, does not fully capture the influence of retrieved evidence. To address this, future work could explore weighted CR metrics or train models to assess the relative influence of each document on the generation outcome.

5.5 Imaginary Terminologies

Although imaginary terminologies are common in jailbreak datasets Su et al. (2024), we purposely avoided them. Including claims with such terminologies would introduce uncontrollable noise, as verifying their grounding in real-world literature is often infeasible due to the complexity of scientific terminology and our lack of domain-specific expertise. While the presence of misinformation and unfamiliar topics mirrors real-world scenarios—where users may unknowingly craft confirmation-biased queries—our primary goal is to build a reliable benchmark for measuring confirmation bias in grounded RAG settings. Introducing imaginary content would compromise this objective by increasing annotation ambiguity and limiting insight into evidence-based generation. Moreover, grounded terminologies would not help us understand how current models handle conflicting or confirming evidence. And our main focus in this work is to evaluate confirmation bias for RAG systems, Imaginary content will be noisey for the retriever to handle.

5.6 General Reflections

One notable finding is that modern LLMs, despite architectural and provider differences, often exhibit similar bias patterns. This convergence suggests that shared pretraining corpora and alignment techniques may lead to consistent confirmation tendencies across models. It also complicates efforts to isolate the specific components (retriever, reranker, generator) responsible for bias. Our study highlights the need for more interpretable architectures and analytical tools capable of disentangling the contributions of each stage in the RAG pipeline. Addressing these challenges is crucial for designing fairer, more transparent information-seeking systems.

Our evaluation indicates LLMs without RAG components are significantly more biased than their RAG variants. As shown in Table 4

System Name	Confirmation Bias Rate
OpenScholar-Llama	0.2767
GPT-4o	0.4509

Table 4: Llama3.1-8b-Instruct trained by the OpenScholar team and GPT-4o with fullset automate evaluation

6 Conclusions

In this research, we propose a benchmark dataset to address the limited awareness and evaluation of confirmation bias in existing RAG systems. We derived from traditional areas and defined confirmation bias in this context and used our benchmark to measure confirmation bias in several SOTA RAG systems in the scientific literature domain. We analyses their strength and weakness with result driven approaches and discovered confirmation bias' existence in different RAG components.

7 Limitations

7.1 Scope

Our work is currently limited in scope to the scientific literature domain. While this area is rich in factual content and structured claims, it does not cover the full spectrum of grounded knowledge domains where confirmation bias can emerge. A broader benchmark, encompassing more general or diverse domains (e.g., politics, health, or finance), could potentially have a greater impact and provide a more comprehensive understanding of RAG model behavior. However, due to resource constraints, we opted for a focused domain to ensure the depth and consistency of our annotations and analysis.

7.2 CR Measurement

Our proposed confirmation ratio (CR) metric, while effective in capturing relative bias tendencies, presents limitations in edge cases. Specifically, for non-controversial topics where only one perspective is overwhelmingly supported by evidence, CR values tend to be close to 0 for these cases. And these edge cases can be caused by many situations such as only one document exist in a specific field This does not necessarily reflect confirmation bias but rather consensus in scientific understanding. Although we aimed to exclude such uncontroversial claims from our dataset, our ability to exhaustively verify the controversiality of every claim pair was limited by practical constraints. Additionally, the challenge of creating a generalized fact-checking dataset that meets the criteria for rigorous bias analysis remains unresolved.

Ethics Statement

We will release our dataset under the CC-BY-NC 4.0 license. The analysis, manual annotation, and automated confirmation bias detection processes do not involve the use of any personal, offensive, or sensitive information. Furthermore, the types of bias discussed and analyzed in this work pertain to confirmation bias in information retrieval and generation systems. They do not involve or target any specific demographic group, nor do they carry negative implications for any protected or marginalized populations.

References

Quantifying and Measuring Confirmation Bias in Information Retrieval Using Sensors | Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs. ArXiv:2411.14199 [cs].

Guillaume Deffuant, Marijn A. Keijzer, and Sven Banisch. 2025. How opinions get more extreme in an age of information abundance. ArXiv:2305.16855 [physics].

Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. 2022. FAIR : Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology*, 73(10):1461–1473.

Gisoo Gomroki, Hassan Behzadi, Rahmatolloah Fattahi, and Javad Salehi Fadardi. 2023. Identifying effective cognitive biases in information retrieval. *Journal of Information Science*, 49(2):348–358. Publisher: SAGE Publications Ltd.

Celina Kacperski, Mona Bielig, Mykola Makhortykh, Maryna Sydorova, and Roberto Ulloa. 2023. Examining bias perpetuation in academic search engines: an algorithm audit of Google and Semantic Scholar. ArXiv:2311.09969 [cs] version: 2.

V. Kayhan. 2015. Confirmation Bias: Roles of Search Engines and Search Contexts.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. 2025. SciClaimHunt: A Large Dataset for Evidence-based Scientific Claim Verification. ArXiv:2502.10003 [cs].

Bangxin Li, Hengrui Xing, Cong Tian, Chao Huang, Jin Qian, Huangqing Xiao, and Linfeng Feng. 2025. StructuralSleight: Automated Jailbreak Attacks on Large Language Models Utilizing Uncommon Text-Organization Structures. ArXiv:2406.08754 [cs].

Gabriel de Souza P. Moreira, Ronay Ak, Benedikt Schifferer, Mengyao Xu, Radek Osmulski, and Even Oldridge. 2024. Enhancing Q&A Text Retrieval with Ranking Models: Benchmarking, fine-tuning and deploying Rerankers for RAG. ArXiv:2409.07691 [cs].

Changwon Ok, Eunkyeong Lee, and Dongsuk Oh. 2025. Synthetic Paths to Integral Truth: Mitigating Hallucinations Caused by Confirmation Bias with Synthetic Data. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5168–5180, Abu Dhabi, UAE. Association for Computational Linguistics.

Dario Onorati, Elena Sofia Ruzzetti, Davide Venditti, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. Measuring bias in instruction-following models with P-AT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8006–8034, Singapore. Association for Computational Linguistics.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. ArXiv:2202.03286 [cs].

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking. Publisher: arXiv Version Number: 2.

Li Shi, Houjiang Liu, Yian Wong, Utkarsh Mujumdar, Dan Zhang, Jacek Gwizdka, and Matthew Lease. 2025. Argumentative Experience: Reducing Confirmation Bias on Controversial Issues through LLM-Generated Multi-Persona Debates. ArXiv:2412.04629 [cs] version: 3.

Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D. Hwang, Jason Dunkleberger, Matt Latzke, Smita Rao, Jaron Lochner, Rob Evans, Rodney Kinney, Daniel S. Weld, Doug Downey, and Sergey Feldman. 2025. Ai2 Scholar QA: Organized Literature Synthesis with Attribution.

Lauren Stewart. 2024. Confirmation bias in research: How to avoid it. Accessed: May 20, 2025.

Jingtong Su, Julia Kempe, and Karen Ullrich. 2024. Mission Impossible: A Statistical Perspective on Jailbreaking LLMs. ArXiv:2408.01420 [cs].

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-Open: Towards open-domain scientific claim verification. ArXiv:2210.13777 [cs].

Fangxin Wang, Lu Cheng, Ruocheng Guo, Kay Liu, and Philip S. Yu. 2023. Equal Opportunity of Coverage in Fair Regression. ArXiv:2311.02243 [cs].

Lionel Wong, Ayman Ali, Raymond Xiong, Shannon Zeijang Shen, Yoon Kim, and Monica Agrawal. 2025. Retrieval-augmented systems can be dangerous medical communicators. ArXiv:2502.14898 [cs].

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Appendix

Study Prompt

A.1 Prompt: Scientific claims Generation Prompt Claude

Generate a set of concise scientific claimss grounded in real-world evidence. For each claim, retrieve or associate a relevant supporting document (e.g., research abstracts, news articles, or reports). Present the results as a JSON file, where each entry includes:
"claim": A clear, testable scientific claim or claim.
"source": A URL or citation of the document grounding the claim.
Ensure that the claims are diverse, factually accurate, and span a range of real-world topics.

A.2 Prompt: Double Checking Prompt GPT

Evaluate the following scientific claimss and verify that they do not contain imaginary, fabricated, or non-existent terminology. For each claim:
Cross-check the terminology used against real-world, scientifically valid sources (e.g., peer-reviewed articles, technical reports, or reputable encyclopedias).
If the claim contains only valid scientific terms, return it as-is, along with a supporting document or citation that confirms its validity.
If the claim includes any invented or unrecognized terms, flag it, and (optionally) suggest a corrected version based on scientifically grounded information.
Present the results as a JSON file with the following fields:
"original-claim": The input claim.
"is-valid": Boolean indicating whether the terminology is valid.
"source": A citation or URL that grounding the terminology.

A.3 Prompt: Confirmation Bias Detection Prompt

You will be given one 'Query' sentence and one 'Evidence' paragraph. Please tell me whether the 'Evidence' supports the 'Query'.
Respond with: - 'Yes' if the Evidence supports the Query, - 'No' otherwise
Output start with — 'Yes', 'No', or 'Others' in your response.

Prompt: Support Ratio measuring prompt

You will be given a 'Query' sentence and an 'Evidence' paragraph. Determine whether the 'Evidence' supports, refutes, or is neutral with respect to the 'Query'.
Respond with: - 'Yes' if the Evidence supports the Query, - 'No' if the Evidence refutes the Query, - 'Others' if the Evidence is neither supporting nor refuting the Query.
Output start with — 'Yes', 'No', or 'Others' in your response.