Evaluation of Attribution Bias in Generator-Aware Retrieval-Augmented Large Language Models

Amin Abolghasemi[1] **Leif Azzopardi**[2] **Seyyed Hadi Hashemi**[3]
Maarten de Rijke[4] **Suzan Verberne**[1]
[1]Leiden University, Netherlands [2]University of Strathclyde, UK
[3]eBay Inc., Netherlands [4]University of Amsterdam, Netherlands
{m.a.abolghasemi, s.verberne}@liacs.leidenuniv.nl
leif.azzopardi@strath.ac.uk shashemi@ebay.com m.derijke@uva.nl

Abstract

Attributing answers to source documents is an approach used to enhance the verifiability of a model's output in retrieval-augmented generation (RAG). Prior work has mainly focused on improving and evaluating the attribution quality of large language models (LLMs) in RAG, but this may come at the expense of inducing biases in the attribution of answers. We define and examine two aspects in the evaluation of LLMs in RAG pipelines, namely attribution sensitivity and bias with respect to authorship information. We explicitly inform an LLM about the authors of source documents, instruct it to attribute its answers, and analyze (i) how sensitive the LLM's output is to the author of source documents, and (ii) whether the LLM exhibits a bias towards human-written or AI-generated source documents. We design an experimental setup in which we use counterfactual evaluation to study three LLMs in terms of their attribution sensitivity and bias in RAG pipelines. Our results show that adding authorship information to source documents can significantly change the attribution quality of LLMs by 3 to 18%. We show that LLMs can have an attribution bias towards explicit human authorship, which can serve as a competing hypothesis for findings of prior work that shows that LLM-generated content may be preferred over human-written contents. Our findings indicate that metadata of source documents can influence LLMs' trust, and how they attribute their answers. Furthermore, our research highlights attribution bias and sensitivity as a novel aspect of the brittleness of LLMs.

1 Introduction

The goal of retrieval-augmented generation (RAG) is to generate an answer to a given question using a set of top-k retrieved documents as context (Lewis et al., 2020). Large language models (LLMs) have been a crucial part of RAG pipelines, mainly as the generator component (Asai et al., 2023; Jeong et al., 2024; Lee et al., 2024; Li et al., 2024a). Although the use of LLMs offers potential benefits, it also presents considerable risks, as they are prone to generate false or hallucinated claims (Ji et al., 2023). This is important as such claims may misguide users, particularly when they are being used in critical fields such as the legal or medical domain (Augenstein et al., 2023; Malaviya et al., 2024; Xiong et al., 2024).

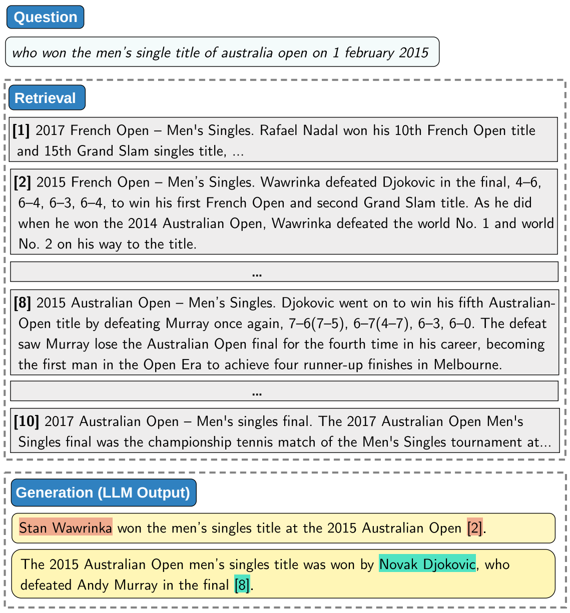

Figure 1: Retrieval-augmented answer/attribution generation using two LLMs. Together with the question, retrieval results are given to the LLMs in order to generate the answer.

Enabling LLMs to attribute their answer to the source of information has been proposed as a promising direction towards reducing the likelihood of such potential harms (Li et al., 2024b,c; Patel et al., 2024). This attribution could assist users in tracing and understanding the basis of the information that LLMs are generating (Gao et al., 2023; Huang and Chang, 2024). There are many prior studies on answer attribution in RAG pipelines (Bohnet et al., 2022; Hu et al., 2024; Li

Findings of the Association for Computational Linguistics: ACL 2025, pages 21105–21124
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2024b; Menick et al., 2022; Muller et al., 2023; Stolfo, 2024). As Figure 1 illustrates, LLMs are susceptible to making mistakes when attributing their answers to the input documents in RAG. Moreover, enabling LLMs in RAG to attribute their answer may come at the expense of inducing biases, as LLMs may carry potential biases (Esiobu et al., 2023; Ozeki et al., 2024; Xie et al.; Ziems et al., 2024). For instance, Tan et al. (2024) show that retrieval-augmented LLMs can be biased towards selecting their own generated text when this kind of content is present in their input. Inspecting these biases is of paramount importance as they can be leveraged for both positive and negative purposes.

In this paper, we study the performance of LLMs in terms of *attribution sensitivity* and *attribution bias* w.r.t. authorship information. When we explicitly inform LLMs about the authors of input documents, and instruct them to attribute their answers to the input documents (by providing citations to these documents), how sensitive are they to the authorship information of input documents? And are they biased towards either human or LLM authorship of input documents? To address these questions, we design a simulated evaluation setup in which we measure to what extent knowing the type of author of input documents affects the quality of attribution (citation) in LLMs.

Prior work has indicated that LLM-generated content may consistently outperform human-authored content in search rankings, which, in turn, results in reducing the presence and exposure of human contributions online (Chen et al., 2024; Dai et al., 2024). Inspired by these studies, we compare human-written documents against LLM-generated documents. We follow prior work in attribution generation by prompting LLMs to generate citations to the input documents (Gao et al., 2023; Yue et al., 2023). Furthermore, we use counterfactual evaluation (Abolghasemi et al., 2024a; Goldfarb-Tarrant et al., 2023; Howard et al., 2024; Huang et al., 2020; Xie et al., 2023) to measure both authorship sensitivity and authorship bias of LLMs in RAG pipelines. This approach can be used more generally to measure algorithmic sensitivity or bias in a model or method: using counterfactual scenarios to see if changing certain characteristics leads to different outcomes.

Our experimental results show that three LLMs (`Mistral`, `Llama3` and `GPT-4`) are sensitive to authorship information that is included in the input documents prior to the generation. Moreover, we show that these models carry a bias towards human authorship against LLM authorship: they are more likely to attribute their answers to documents that are explicitly labelled as having been written by humans (even if the documents are actually generated by LLMs).

We summarize our contributions as follows:

- We define and study attribution sensitivity and bias w.r.t. authorship information, as a novel aspect of trustworthiness and brittleness in retrieval-augmented LLMs.
- We propose a systematic evaluation framework for measuring attribution sensitivity and bias.
- We show that adding authorship information (as metadata) to source documents may lead to statistically significant changes in the attribution quality of retrieval-augmented LLMs.
- We show that LLMs may have an attribution bias towards explicit human *authorship*, which can serve as a competing hypothesis for findings of prior work that shows that LLM-generated *content* is preferred over human-written *content* by LLMs.[1]

2 Background

Retrieval-Augmented Generation. Given a question q and a set of top-k retrieved documents $\mathcal{D} = \{d_1, d_2, \ldots, d_k\}$ from a collection $\mathcal{C}$, the goal of retrieval-augmented generation (RAG) is to generate an answer for q using $\mathcal{D}$ as context. LLMs are currently an important component of RAG pipelines, acting as the generator. The generator is given q, $\mathcal{D}$, and an instruction prompt on how to generate the answer (Jeong et al., 2024; Lee et al., 2024; Li et al., 2024a). Using top-k retrieved documents helps LLMs to be exposed to information that it might not have been trained/fine-tuned with during development. These documents are commonly retrieved using an effective sparse and/or dense retriever (Lewis et al., 2020; Rau et al., 2024).

Attributive RAG. LLMs are prone to generate hallucinated (and even factually incorrect) answers (Ji et al., 2023; Rawte et al., 2023; Yue et al., 2024). Attributing answers in RAG with LLMs is an approach taken as a step towards ensuring the veracity of the output of these models (Bohnet et al., 2022; Hu et al., 2024; Kamalloo et al., 2023; Khalifa et al.; Li et al., 2024b). Menick et al. (2022) teach

[1]Our code is available at `https://github.com/aminvenv/attrieval`

language models to support answers with verified quotes. Ye et al. (2024) propose a learning-based framework in which they fine-tune LLMs to generate citations, as opposed to prompting or relying on post-hoc attribution. Stolfo (2024) analyzes whether every generated sentence in the output of LLMs is grounded in the retrieved documents or the LLM's pre-training data.

3 Methodology

We aim to measure the attribution sensitivity and bias of LLMs in RAG settings. We investigate to what extent the attribution quality of LLMs is affected by authorship information. To this end, we use counterfactual evaluation (Bottou et al., 2013; Gardner et al., 2020; Wang and Culotta, 2021). Counterfactual evaluation has been used across various natural language processing and information retrieval tasks (Abolghasemi et al., 2024a,b; Goldfarb-Tarrant et al., 2023; Howard et al., 2024; Huang et al., 2020). This approach evaluates how a model's predictions change when a specific feature or set of features is altered while keeping everything else constant. In our case, the change is to add authorship information to the input documents of an LLM in a RAG setting. By doing so, we can evaluate the model's reliance on, bias towards, or sensitivity to that feature. To this end, we first generate answers with LLMs in a RAG setting using three RAG modes, as shown in Figure 2.

3.1 RAG Modes

Given a query q and a set of top-k retrieved documents $\mathcal{D}_q$ for q, we define three modes, based on authorship information of these documents that we provide to the answer generator LLM.

Vanilla RAG. In this mode, each document in $\mathcal{D}$ is given to the LLMs without information about who the authors are. This is the plain input format for input documents as shown in the input prompt for *vanilla* answer/attribution generation in Figure 3.

Authorship Informed RAG. In this mode, we inform the LLM about the actual author of each document. We denote the authorship of either an LLM or a human using [LLM] and [Human] tokens as shown by Figure 7 in the Appendix.[2]

Counterfactual-Authorship Informed RAG. In this mode, we assign counterfactual authorship for each document. If a document is written by a human, the counterfactual authorship of this document is [LLM]. In contrast, if a document is generated by an LLM, its counterfactual authorship is [Human]. By doing so, we can investigate to what extent being written by either human or LLM affects the attribution quality of LLM. The prompt used for this mode is the same as the one for Authorship Informed RAG mode.

[2]In Section C in the Appendix, we study and provide results on replacing [Human] with a set of actual {firstname, lastname} as authors.

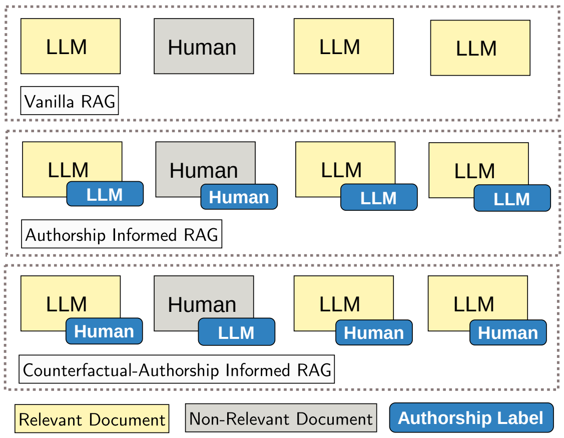

Figure 2: Three RAG modes (Section 3.1) for the setting with LLM actual authorship for relevant documents and Human actual authorship for non-relevant documents. The text in a rectangle denotes the actual generator (i.e., author) of each document. The text in the blue tags denotes the authorship label about which we inform the answer/attribution generator LLM.

Figure 2 shows the three RAG modes for a setting where the relevant documents are LLM-written and the non-relevant documents are human-written.

3.2 Answer/Attribution Generation

In order to generate answers with each of the aforementioned RAG modes, we experiment with three LLMs: `Mistral` (Jiang et al., 2023), `Llama3` (Dubey et al., 2024) and `GPT-4` (OpenAI, 2023). Figure 3 shows the prompt used for *vanilla* answer generation. Figure 7 in the Appendix shows the prompt used for *Authorship-Informed* and *Counterfactual-Authorship Informed* answer generation. We follow prior work (Gao et al., 2023) in curating our prompts for this task.

3.3 Evaluation Metrics

Attribution Quality. We use precision and recall for evaluating the quality of attribution, i.e., how well the LLMs cite the relevant input documents. Precision of attribution for a single query is the fraction of correct citations among all cited documents

```
Instruction: Write a concise answer for the given question (query) based on the provided search
result documents, and cite them properly using [0][1][2] etc.

Please take these strict considerations into account during answer generation:
1. Documents are retrieved by a search engine. As such, not all the documents are relevant to
the query. Only use and cite the relevant documents that contain the answer.
2. Do not analyze irrelevant documents.

Search Results:

Document [0]({text of Document [0]})
Document [1]({text of Document [1]})
...
Document [9]({text of Document [9]})

Question: {query}.
```

Figure 3: Prompt used for vanilla retrieval-augmented answer generation.

in the output of an LLM. Recall is the fraction of cited relevant documents out of all relevant documents (Djeddal et al., 2024). We use the queries that have only one relevant document containing the ground-truth answer in their top-k retrieved list of documents.

Attribution Sensitivity. In order to measure the sensitivity of LLMs in RAG pipelines towards knowing authors of input documents in comparison to not knowing it, we use counterfactual evaluation and define a metric called Counterfactually-estimated Attribution Sensitivity (CAS):

$$\text{CAS}(Q) = \frac{1}{|Q|}\sum_{q\in Q}|M^q_{\text{Informed}} - M^q_{\text{Vanilla}}|. \quad (1)$$

Here, M^q represents the precision and recall metrics for query q, i.e., attribution quality for query q. For a single query q, CAS measures the difference between a base setup (the vanilla RAG mode) and a counterfactual setup (the authorship informed RAG mode) for the same set of input documents.

Attribution Bias. In order to measure the attribution bias of LLMs in RAG pipelines we use counterfactual evaluation and define a metric called Counterfactually-estimated Attribution Bias (CAB):

$$\text{CAB}(Q) = \frac{\omega}{|Q|}\sum_{q\in Q} M^q_{\text{Informed}} - M^q_{\text{CF-informed}} \quad (2)$$

$$\omega = \begin{cases} 1, & \text{if } L_f(\mathcal{R})=\texttt{[Human]}, L_f(\mathcal{N})=\texttt{[LLM]} \\ -1, & \text{otherwise.} \end{cases} \quad (3)$$

Here, M^q represents the precision and recall metrics, i.e., attribution quality, for query q, given the set of retrieved relevant documents $\mathcal{R}$, and the set of retrieved non-relevant documents $\mathcal{N}$. $L_f(\mathcal{X})$ stands for the authorship label of the set of documents $\mathcal{X}$ in the first term of Eq. 2, i.e., corresponding to M^q_{Informed}. For example, if we use human-written version of relevant documents ($\mathcal{R}$), and LLM-written version of non-relevant document ($\mathcal{N}$), and we label them with their actual generators (authors), i.e., we use authorship-informed RAG mode, then $L_f(\mathcal{R})$ is equal to [Human], and $L_f(\mathcal{R})$ is equal to [LLM]. CAB measures the difference between metric values of a base setup (the Authorship Informed RAG mode) and a counterfactual setup (the Counterfactual-authorship Informed RAG mode) for the same set of input documents consisting of $\mathcal{R}$, and $\mathcal{N}$. ω determines the direction of bias towards either human or LLMs: if the set of relevant documents ($\mathcal{R}$) and non-relevant documents ($\mathcal{N}$) are respectively written by Human and LLM (i.e., $L_f(\mathcal{R})$ = [Human], $L_f(\mathcal{N})$ = [LLM]), for a single query, a positive difference ($M_{\text{Informed}} - M_{\text{CF-informed}}$) indicates bias towards human authorship, and a negative difference shows bias towards LLM authorship. In contrast, if the set of relevant documents ($\mathcal{R}$) and non-relevant documents ($\mathcal{N}$) are respectively written by LLM and Human (i.e., $L_f(\mathcal{R})$ = [LLM], $L_f(\mathcal{N})$ = [Human]), a negative difference ($M_{\text{Informed}} - M_{\text{CF-informed}}$) indicates a bias towards human authors, and a positive difference shows bias towards LLMs. We use ω to align these two conditions of actual authorship of input documents.

Attribution Confidence. To better explore the performance of LLMs in attribution generation, we analyze whether the LLMs are more confident when they attribute to certain types of document. To this aim, we look into the average probability of generation for attribution tokens, i.e., citation numbers (0, 1, ...):

$$\mathrm{AC}(\mathcal{S}) = \frac{\sum_{q\in Q}\sum_{c_i\in C_q} p(c_i|q,\mathcal{D}_q)\cdot \mathbb{1}[c_i\in\mathcal{S}]}{|\sum_{q\in Q}\sum_{c_i\in C_q}\mathbb{1}[c_i\in\mathcal{S}]|}, \quad (4)$$

where q is a query in the set of queries Q, $\mathcal{D}_q$ is the top-k retrieved documents for q, C_q stands for all attribution numbers in the answer to q_i, $c_i \in \{0, 1, \ldots, k\}$, $\mathcal{S}$ indicates a set of documents, e.g., the set of relevant documents for all queries, and $p(c_i|q, \mathcal{D}_q)$ shows the probability of generation for the attribution token c_i in the answer provided by LLM given query q and its top-k retrieved documents $\mathcal{D}_q$. $\mathbb{1}[c_i \in \mathcal{S}]$ equals 1 if $c_i \in \mathcal{S}$.

Answer Correctness. In order to evaluate the quality of the generated answer, we follow (Gao et al., 2023; Petroni et al., 2021) and use automatic evaluation. Following (Gao et al., 2023; Stolfo, 2024), we use the normalized human-generated answer in the benchmark as the ground-truth answer and adopt Exact Match (EM) (Siriwardhana et al., 2023; Wang et al., 2023) as the evaluation metric for answer correctness (see example in Figure 16).

4 Experimental Settings

Models. We use `gpt-4-0314` (OpenAI, 2023), `meta-llama/Meta-Llama-3-8B-Instruct`,[3] and `mistralai/Mistral-7B-Instruct-v0.3`[4] as answer generator LLMs in our RAG pipelines. We refer to these models as `GPT-4`, `Llama3`, and `Mistral`, respectively.

Benchmarks. We use two benchmarks in our experiments: Natural Questions (NQ) (Kwiatkowski et al., 2019) and MS MARCO Question Answering (Bajaj et al., 2016) (to which we refer as MS MARCO). For each benchmark, we randomly sample 500 queries. To retrieve top-k passages for each query in the NQ benchmark, we use BM25, a widely-used lexical matching retrieval model. For queries in the MS MARCO benchmark, we use passages that are extracted from relevant web documents using the state-of-the-art passage retrieval system at Bing (Bajaj et al., 2016). We note that we study the effect of different retrievers and different number of retrieved source documents in Section D and E in the Appendix, respectively.

Synthetic Collection. To construct LLM-written documents, we use `Llama3` to re-write a given document from our collections without adding/removing information to/from the document. Specifically, we use a low temperature close to 0 as it makes the LLM extremely restrictive, focusing only on the most probable tokens resulting in (highly) deterministic outputs. The reason for not generating the documents from scratch is to make sure we keep the relevance/non-relevance status of documents w.r.t a query. To ensure the quality of synthetic passages, we conduct a number of annotation steps using two expert annotators. This is detailed in Section A in the Appendix. Importantly, in Section 5, we show that even without using LLM-generated documents (i.e., only designating `[Human]` and `[LLM]` as authors of documents to the original input documents) findings are the same as when we use actual LLM-generated documents.

5 Experimental Results

In this section, we explore the performance of LLMs for attributing their answer to top-k retrieved source documents using the evaluation metrics introduced in Section 3. All significance tests in the result tables are according to a paired t-test with $p < 0.05$.

Attribution Quality. Table 1 shows the results of attribution by three LLMs, `Mistral`, `Llama3` and `GPT-4`, under different settings for NQ benchmark. Besides, Table 11 in the Appendix shows the same set of results for the MS MARCO benchmark. The two columns "Relevant documents" and "Non-relevant documents" indicate the actual generator (author) of these documents. The column "RAG mode" indicates how we inform the answer generator LLMs about the authorship label of relevant and non-relevant documents, as described in Section 3.1: in the "Vanilla" RAG mode, no information regarding the generator (author) of the input source documents is given to the LLM. In the "Informed" RAG mode the LLM is informed about the actual generator of the input source documents, and in the "CF-Informed" RAG mode the LLM is provided with counterfactual authorship information. As Table 1 shows, the three LLMs (`Mistral`, `Llama3` and `GPT-4`) fall short of perfectly grounding their answers to the relevant documents of a given question, which is in line with the findings of

[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[4]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Answer generator	Relevant documents	Non-relevant documents	RAG mode	Attribution quality		Correctness
				Precision	Recall	EM
NQ						
Mistral	LLM	Human	Vanilla	47.6	76.6	0.722
			Informed	42.1	68.2	0.730
			CF-informed	**52.7**†	**77.8**†	**0.738**
	Human	LLM	Vanilla	51.0	78.4	**0.776**
			Informed	**53.4**†	**77.8**†	0.774
			CF-informed	44.0	70.2	0.772
Llama3	LLM	Human	Vanilla	49.2	69.2	0.742
			Informed	45.4	69.6	0.730
			CF-informed	**57.2**†	**77.6**†	**0.748**
	Human	LLM	Vanilla	53.5	71.0	0.766
			Informed	**59.9**†	**77.8**†	**0.790**
			CF-informed	44.8	69.2	0.762
GPT-4	LLM	Human	Vanilla	63.3	68.8	0.736
			Informed	59.7	64.6	0.740
			CF-informed	**65.9**†	**72.2**†	**0.742**
	Human	LLM	Vanilla	64.1	68.8	0.760
			Informed	**66.1**	**72.2**†	**0.776**
			CF-informed	60.3	65.0	0.758

Table 1: Quality of attribution and answer correctness. The columns “Relevant Documents” and “Non-relevant Documents” refer to the actual authorship of input documents. Informed refers to the authorship-informed RAG and CF-informed refers to counterfactual-authorship informed RAG (Section 3.1). † indicates statistically significant improvements over the two other RAG Modes in each combination of “Relevant” and “Non-relevant” documents.

prior work (Djeddal et al., 2024; Gao et al., 2023; Li et al., 2024c).

Answer generator	Relevant documents	Non-relevant documents	CAS	
			ΔPrecision	ΔRecall
NQ				
Mistral	LLM	Human	16.2†	17.2†
	Human	LLM	20.1	17.0
Llama3	LLM	Human	13.2†	14.8
	Human	LLM	17.7†	16.0†
GPT-4	LLM	Human	9.7†	10.2†
	Human	LLM	8.7	9.0†
MS MARCO				
Mistral	LLM	Human	10.9	21.4†
	Human	LLM	12.9†	16.6
Llama3	LLM	Human	12.9†	20.4†
	Human	LLM	17.8†	19.6†
GPT-4	LLM	Human	8.2†	9.6†
	Human	LLM	10.9	15.8†

Table 2: Attribution sensitivity (CAS) results. Values range from 0 (minimum sensitivity) to 100 (maximum sensitivity). † indicates statistically significant values.

Attribution Sensitivity and Bias. Table 3 shows the attribution bias results in terms of CAB (Eq. 2). All three LLMs, `Mistral`, `Llama3`, and `GPT-4`, carry a bias towards human authorship in the input documents. Moreover, on both datasets, NQ and MS MARCO, `Mistral` and `Llama3` have higher bias values than `GPT-4`. Besides, Table 2 shows the attribution sensitivity results in terms of CAS (Eq. 1). All three LLMs, `Mistral`, `Llama3`, and `GPT-4`, show sensitivity towards the inclusion of authorship information for the input documents of LLMs. Moreover, similar to the attribution bias values in Table 3, `Mistral` and `Llama3` carry a higher attribution sensitivity than `GPT-4`. We note that we conducted experiments using different prompts and observed that the findings remained consistent across multiple runs.

Mixed RAG Mode. To better disentangle the effect of LLM generated text qualities (e.g., a potential implicit bias of LLMs towards LLM-written documents (Tan et al., 2024)) from the impact of authorship information, we now use the same set of documents in the input of LLM in the RAG, and only use different authorship labels for relevant and non-relevant documents. For this new setup, to which we refer as the Mixed RAG mode, we evaluate both a complete set of synthetic documents (i.e., for both relevant and non-relevant) and a complete set of human-written documents.

Answer generator	Relevant documents	Non-relevant documents	CAB ΔPrecision	CAB ΔRecall
NQ				
Mistral	LLM	Human	+10.6†	+9.6†
	Human	LLM	+9.4†	+7.6†
Llama3	LLM	Human	+11.8†	+8.0†
	Human	LLM	+15.1†	+8.6†
GPT-4	LLM	Human	+6.2†	+7.6†
	Human	LLM	+5.8†	+7.2†
MS MARCO				
Mistral	LLM	Human	+9.5†	+13.8†
	Human	LLM	+8.0†	+12.4†
Llama3	LLM	Human	+15.6†	+18.2†
	Human	LLM	+15.1†	+16.4†
GPT-4	LLM	Human	+6.1†	+9.0†
	Human	LLM	+5.4†	+10.8†

Table 3: Attribution Bias (CAB) results. Values range from -100 (completely biased towards LLM authorship) to +100 (completely biased towards human authorship). † indicates statistically significant bias values.

Figure 4 shows an example of Mixed RAG mode for the setting where we have human-written documents, with different authorship labels for relevant and non-relevant documents. The CAB (Eq. 2) for Mixed RAG mode is reformulated as follows:

$$\mathrm{CAB}(Q) = \frac{\omega}{|Q|}\sum_{q\in Q} M^q_{\text{Informed/CF-Informed}} - M^q_{\text{CF-Informed/Informed}}, \tag{5}$$

where X and Y in $M^q_{X/Y}$ stand for the RAG mode for the set of relevant documents and the set of non-relevant documents, respectively. The results of attribution quality for Mixed-RAG modes are shown in Table 4.[5] We see that, similar to Table 1, across different settings, when the relevant documents are labeled with human-authorship and non-relevant ones are labeled with LLM-authorship, the attribution quality is higher than the other way around. Moreover, Table 5 illustrates the attribution bias for Mixed RAG modes. Similar to the results in Table 3, there is a bias towards human authorship in all three LLMs. This indicates the existence of authorship bias regardless of the origin of the input documents, i.e., the actual author of the input documents. Furthermore, similar to the results in Table 3, `Mistral` and `Llama3` show higher rates of attribution bias than `GPT-4`. Additionally, we find that when we have the same authorship label on both relevant and non-relevant documents (rows with the same RAG mode for relevant and non-relevant documents in Tables 14 and 15 in the Appendix), we do not see consistent patterns as to how LLMs attribute the answers to the input documents. Finally, we note that in Section C of the Appendix, we show additional results using real-world names as authors which further indicates the presence of attribution bias and sensitivity in LLMs towards authorship information.

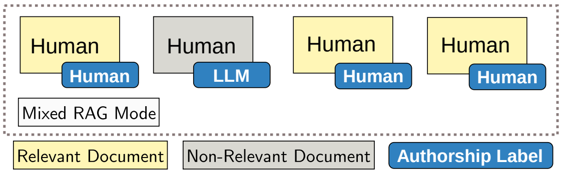

Figure 4: Mixed RAG mode for the setting where we use original human-authored documents. In this example, we have "Informed" mode for relevant documents and "CF-Informed" for non-relevant documents.

Attribution Confidence. Using Eq. 4, we compute the confidence of LLMs when they attribute their answer to an input document. Table 6 shows the attribution confidence of LLMs for relevant and non-relevant documents.[6] Across the majority of RAG modes over different origins for relevant and non-relevant documents, the confidence of all three LLMs for attributing to relevant documents is higher than for attributing to non-relevant ones. We can also see that authorship labels (i.e., using different RAG modes) do not affect this outcome. In other words, it is being relevant or not that makes the difference here. These results indicate a promising direction for improving attribution in LLMs: low confidence of LLMs in attributing to a specific document might be a useful signal for the relevance of that document to a given query.

Frequency of Attribution. In Table 1, across the majority of the settings, `GPT-4` outperforms `Mistral` and `Llama3` in terms of precision of results. In contrast, in terms of recall, it is `Mistral` and `Llama3` which outperform `GPT-4`. To better explore this difference, we examine the average number of relevant citations and total citations for the three models. Figure 5 shows the average number of total citations[7] for each model. In comparison to `Mistral` and `Llama3`, `GPT-4` tends to cite

[5] See Tables 14 and 15 (Appendix) for the complete set of results.

[6] Table 10 in the Appendix shows the results on MS MARCO.

[7] Tables 12 and 13 in the Appendix show both the average number of relevant citations and the total citations.

Answer generator	Relevant documents	Non-relevant documents	Mixed RAG mode		Attribution quality		Correctness
			Relevant	Non-relevant	Precision	Recall	EM
NQ							
Mistral	Human	Human	CF-informed	Informed	44.8	71.8	0.772
			Informed	CF-informed	**52.3**†	**77.2**†	**0.780**
	LLM	LLM	CF-informed	Informed	**48.7**†	**74.6**†	0.718
			Informed	CF-informed	42.9	69.4	**0.742**
Llama3	Human	Human	CF-informed	Informed	45.7	69.6	0.784
			Informed	CF-informed	**57.4**†	**77.6**†	**0.808**
	LLM	LLM	CF-informed	Informed	**59.3**†	**77.8**†	**0.744**
			Informed	CF-informed	44.7	68.4	0.726
GPT-4	Human	Human	CF-informed	Informed	65.8	70.6	**0.794**
			Informed	CF-informed	**69.1**†	**74.0**†	0.784
	LLM	LLM	CF-informed	Informed	**66.1**†	**71.2**†	**0.730**
			Informed	CF-informed	61.7	66.8	0.716

Table 4: Quality of attribution and answer correctness for Mixed RAG mode. The columns "Relevant Documents" and "Non-relevant Documents" refer to the actual authorship of input documents. † indicates statistically significant improvements over the other Mixed RAG mode in each combination of relevant and non-relevant documents.

Answer generator	Relevant documents	Non-relevant documents	CAB	
			ΔPrecision	ΔRecall
NQ				
Mistral	Human	Human	+7.5†	+5.4†
	LLM	LLM	+5.8†	+5.2†
Llama3	Human	Human	+11.7†	+8.0†
	LLM	LLM	+14.6†	+9.4†
GPT-4	Human	Human	+3.3†	+3.4†
	LLM	LLM	+4.4†	+4.4†
MS MARCO				
Mistral	Human	Human	+8.6†	+14.8†
	LLM	LLM	+8.7†	+13.8†
Llama3	Human	Human	+12.6†	+10.4†
	LLM	LLM	+9.7†	+9.8†
GPT-4	Human	Human	+7.4†	+9.4†
	LLM	LLM	+5.4†	+5.2†

Table 5: Attribution Bias (CAB) results for Mixed RAG modes. Positive values indicate a bias towards human. † indicates statistically significant bias values. Values range from -100 (completely biased towards LLM authorship) to +100 (completely biased towards human authorship).

fewer documents as supporting documents for its generated answer. This is in line with the previous results, where `Mistral` and `Llama3` score higher on recall.

Answer Correctness. Table 1 and 4 show that when the relevant documents are labeled with human-authorship and non-relevant ones are labeled with LLM-authorship, the answer correctness is higher than in the reverse case, across the majority of settings. Although this improvement is not significant and consistent across all settings, similar to attribution quality, it could indicate a bias towards human authorship. Nevertheless, we note that the automatic evaluation of answer correctness without human evaluation is not an ideal method (Bojic et al., 2023; Chiang and Lee, 2023; Zhang et al., 2019).We leave this aspect for future work as the focus of this paper is on the performance of LLMs in how frequently they tend to cite and attribute their output on documents with either human or LLM authorship.

Answer generator	Rel. Docs.	Non-rel. docs.	RAG mode	Confidence (AC)	
				Rel.	Non-rel.
NQ					
Mistral	LLM	Human	Vanilla†	0.9647	0.9284
			Informed†	0.9656	0.9257
			CF-informed†	0.9737	0.9401
	Human	LLM	Vanilla†	0.9678	0.9355
			Informed†	0.9707	0.9400
			CF-informed†	0.9638	0.9434
Llama3	LLM	Human	Vanilla†	0.9060	0.8145
			Informed†	0.8960	0.8260
			CF-informed†	0.9235	0.8282
	Human	LLM	Vanilla†	0.9088	0.7985
			Informed†	0.9163	0.8160
			CF-informed†	0.8908	0.8238
GPT-4	LLM	Human	Vanilla†	0.9807	0.9042
			Informed†	0.9796	0.9130
			CF-informed†	0.9834	0.9094
	Human	LLM	Vanilla†	0.9819	0.9238
			Informed†	0.9778	0.9205
			CF-informed†	0.9776	0.9346

Table 6: The attribution confidence (AC) of LLMs in relevant and non-relevant documents for NQ dataset. † indicates a statistically significant difference between the AC values of relevant and non-relevant documents.

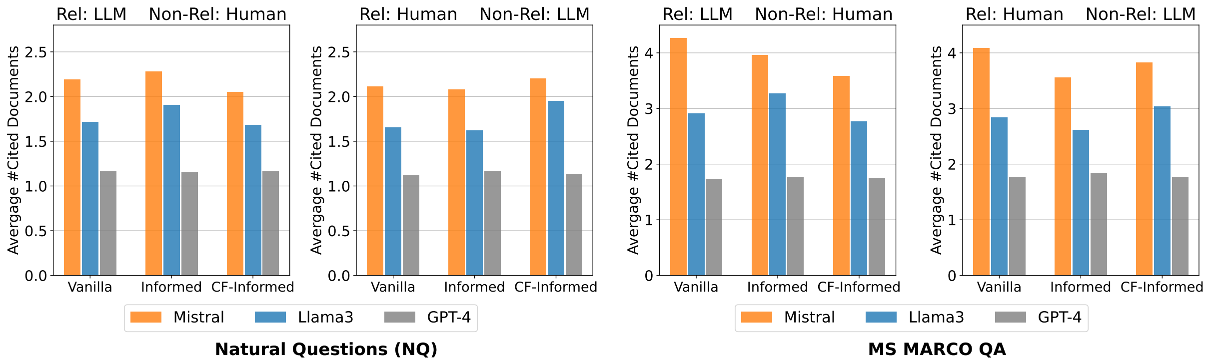

Figure 5: The average total number of cited documents by `Mistral`, `Llama3`, and `GPT-4` across various RAG settings on NQ and MS MARCO benchmarks.

6 Conclusion and Future Work

In this paper, we have defined and studied attribution sensitivity and bias with respect to authorship information of source documents in RAG with LLMs. We have proposed a systematic evaluation framework based on counterfactual evaluation. Our results indicate that by adding authorship information to source documents, the attribution quality of LLMs may significantly change by 3% to 18%. Moreover, our results on three LLMs indicate that they have an attribution bias towards explicit human *authorship*, in contrast to previous studies that show that LLM-generated *content* may consistently be preferred over human-authored *content* by LLMs.

As to broader implications of our work, while understanding the roots and causes of the observed sensitivity and bias requires access to the implementation, training, and fine-tuning of these models (which is beyond the scope of this paper), our findings highlight a critical aspect of how LLMs operate. Our results show the brittleness of LLMs for attributing their answers. Such brittleness can be used for both constructive and harmful purposes, e.g., one can bias the output of an LLM towards their own content by incorporating authorship information in their documents.

While we only focused on human versus LLM authorship as metadata in this work, in future work our systematic evaluation method can be used to investigate the sensitivity and bias towards other metadata of source documents (e.g., gender and race of authors). Furthermore, our evaluation methodology can be incorporated in trustworthiness benchmarks used for the evaluation of LLMs such as DecodingTrust (Wang et al., 2024). Finally, our proposed methodology for the evaluation of sensitivity and bias is adaptable to other metrics for measuring the quality of attribution, i.e., metrics other than precision and recall can be used as M in Eq. 1, 2, and 5.

Limitations

In this work we do not propose or explore solutions for mitigating the observed bias as our focus is on uncovering the brittleness of LLMs when being used for retrieval-augmented generation. Besides, we have evaluated three LLMs in our experimental setup, two of which are open-source and the other closed-source. Consequently, investigating the same attribution sensitivity and bias on other LLMs is of interest for future studies. Furthermore, in our experiments, we used queries that have only one relevant document containing the ground-truth answer in their top-k retrieved list of documents. We do this to ensure the traceability of the correct attribution. However, we acknowledge the limitation of this evaluation method in capturing the fine-grained attribution support of input documents. Finally, it is important to mention that our current research is limited to datasets and prompts in English. Therefore, we point out the need to expand our evaluation and analysis to include datasets in other languages.

References

Amin Abolghasemi, Leif Azzopardi, Arian Askari, Maarten de Rijke, and Suzan Verberne. 2024a. Measuring bias in a ranked list using term-based representations. In *European Conference on Information Retrieval*, pages 3–19. Springer.

Amin Abolghasemi, Zhaochun Ren, Arian Askari, Mohammad Aliannejadi, Maarten de Rijke, and Suzan Verberne. 2024b. CAUSE: Counterfactual assessment of user satisfaction estimation in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14623–14635, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, and 1 others. 2023. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, and 1 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Iva Bojic, Jessica Chen, Si Yuan Chang, Qi Chwen Ong, Shafiq Joty, and Josip Car. 2023. Hierarchical evaluation framework: Best practices for human evaluation. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 11–22.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11).

Xiaoyang Chen, Ben He, Hongyu Lin, Xianpei Han, Tianshu Wang, Boxi Cao, Le Sun, and Yingfei Sun. 2024. Spiral of silences: How is large language model killing information retrieval?–a case study on open domain question answering. *arXiv preprint arXiv:2404.10496*.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. Neural retrievers are biased towards LLM-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 526–537.

Hanane Djeddal, Pierre Erbacher, Raouf Toukal, Laure Soulier, Karen Pinel-Sauvagnat, Sophia Katrenko, and Lynda Tamine. 2024. An evaluation framework for attributed information retrieval using large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)*. Association for Computing Machinery.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. ROBBIE: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, and 1 others. 2020. Evaluating models' local decision boundaries via contrast sets. *Findings of Empirical Methods in Natural Language Processing*.

Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. 2023. Bias beyond English: Counterfactual tests for bias in sentiment analysis in four languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4458–4468.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2024. SocialCounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11975–11985.

Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Sheng Bi, Tongtong Wu, and Jeff Z. Pan. 2024. Benchmarking large language models in complex question answering attribution using knowledge graphs. *arXiv preprint arXiv:2401.14640*.

Jie Huang and Kevin Chang. 2024. Citation: A key to building responsible and accountable large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 464–473.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7029–7043.

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

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A human-LLM collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. Source-aware training enables knowledge attribution in language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Myeonghwa Lee, Seonho An, and Min-Soo Kim. 2024. PlanRAG: A plan-then-retrieval augmented generation for generative large language models as decision makers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6537–6555.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. 2024a. TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3799–3821.

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2024b. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 493–516, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024c. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14919–14935, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807*.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Benjamin Muller, John Wieting, Jonathan H Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and modeling attribution for cross-lingual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 144–157.

OpenAI. 2023. GPT-4 technical report. *OpenAI*.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2024. Exploring reasoning biases in large language models through syllogism: Insights from the

NeuBAROCO dataset. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16063–16077, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Nilay Patel, Shivashankar Subramanian, Siddhant Garg, Pratyay Banerjee, and Amita Misra. 2024. Towards improved multi-source attribution for long-form answer 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)*, pages 3906–3919.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, and 1 others. 2021. KILT: A benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Vassilina Nikoulina, and Stéphane Clinchant. 2024. Bergen: A benchmarking library for retrieval-augmented generation. *arXiv preprint arXiv:2407.01102*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Alessandro Stolfo. 2024. Groundedness in retrieval-augmented long-form generation: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1537–1552.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain QA? *arXiv preprint arXiv:2401.11911*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, and 1 others. 2024. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. *Advances in Neural Information Processing Systems*, 36.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14024–14031.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Zhongbin Xie, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2023. Counter-gap: Counterfactual bias evaluation through gendered ambiguous pronouns. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3761–3773.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation 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)*, pages 6237–6251.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635.

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)*, pages 5628–5643.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Caleb Ziems, William Held, Jane Dwivedi-Yu, and Diyi Yang. 2024. Measuring and addressing indexical bias in information retrieval. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12860–12877, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Appendix

A Synthetic Document Generation

Prompt. Figure 6 shows the prompt used for rewriting passages for the two benchmarks of NQ and MS MARCO.

```
Instruction: Please write a high-quality
paraphrase for the given passage.
Keep the length approximately the same. Do
not add any new information.

Passage: {input passage}
```

Figure 6: Prompt used for generating synthetic documents.

Data Quality. In order to ensure the quality of synthetic passages, we conduct the following annotation steps using two expert annotators: (i) for each of the queries in our two benchmarks, we provide the annotators the quadruple of (query q, original relevant passage p_r, synthetic relevant $\hat{p}_r$ passage, answer a). We then ask the annotators to determine whether the synthetic passage $\hat{p}_r$ is still relevant to the query and includes the answer a to the query q. (ii) In order to ensure that non-relevant passages are still non-relevant after being rewritten by an LLM, for each query, we provide the annotators the quadruple of (query q, original non-relevant passage p_n, synthetic non-relevant $\hat{p}_n$ passage, answer a). We then ask the annotators to determine whether the synthetic non-relevant passage $\hat{p}_n$ is still non-relevant to the query and does not include the answer a to the query q. Due to the large number of non-relevant passages for each query, we randomly select 10% of queries, i.e., 50 queries out of 500 queries. Our annotation results indicate a perfect performance in keeping the relevance and non-relevance status of synthetic documents with respect to their corresponding queries.

B Authorship Informed Answer/Attribution Generation Prompt

Figure 7 shows the prompt used for authorship-informed answer/attribution generation with all three LLMs, `Mistral`, `Llama3`, and `GPT-4`.

C Extended Set of Authorship Labels

So far, we have used `[LLM]` and `[Human]` as the authorship labels for the source documents. In this section, we discuss and provide results using an extended set of authorship labels. Specifically, we use `[AI]` as the label for denoting the synthetic (LLM) authorship. For human authorship, on the other hand, we analyze the use of real-world names to indicate the authors of documents. This reflects a more realistic setting of authorship indication on documents. To create this set of names, we prompt `GPT-4` to randomly generate a pool of 100 (first name, last name) pairs. Figure 8 shows the prompt we use for this task. We then randomly sample one instance of (first name, last name) from this pool when labeling human authorship for each document in the list of top-k source documents of a query (instead of using `[Human]` as the authorship label).

Table 7 shows the attribution sensitivity results using the extended set of authorship labels. As we can see, all three LLMs `Mistral`, `Llama`, and `GPT-4` are sensitive to adding the authorship information similar to the attribution sensitivity results with `[Human]` and `[LLM]` authorship labels (Table 2). In addition, we see that `GPT-4` shows a lower level of sensitivity than `Mistral` and `Llama`. Moreover, Table 8 shows the attribution bias results using the extended set of authorship labels. Similar to the attribution bias results with `[Human]` and `[LLM]` authorship labels (Table 3), all three LLMs `Mistral`, `Llama`, and `GPT-4` show an attribution bias towards human authorship, i.e., they are biased towards documents that are labeled with human author names. This indicates the robustness of our analysis against changes in labels.

D Effect of the Number of Source Documents

To study the effect of the number of source documents, i.e., the length of the retrieved ranked list of documents given to the answer generator LLM, we evaluate the attribution sensitivity and bias using varying number of source documents. To this end, we use 4 ranking cut-offs for the ranked list of source documents (k): 2, 5, 8, 10. To ensure the existence of the relevant document as the input, we randomly put the relevant document in the top-k ($k \in \{2, 5, 8, 10\}$). For this set of experiments we use human-generated versions of both relevant and non-relevant documents. Furthermore, we use the

```
Instruction: Write a concise answer for the given question (query) based on the provided search
result documents, and cite them properly using [0][1][2] etc.

Please take these strict considerations into account, during answer generation:
1. Documents are retrieved by a search engine. As such, not all the documents are relevant to the
query. Only use and cite the relevant documents that contain the answer.
2. Do not analyze irrelevant documents.
3. Some documents are authored by humans, while others are generated by a Large Language Model (LLM).
We denote this distinction for each document.

Search Results:

Document [0]({text of Document [0]}) (written by [Author])
Document [1]({text of Document [1]}) (written by [Author])
...
Document [9]({text of Document [9]}) (written by [Author])

Question: {query}.
```

Figure 7: Prompt used for authorship-informed answer/citation generation with LLM. `[Author]` is filled with one instance from either {`Human`, `Person`, `Individual`} or {`AI`, `LLM`, `Machine`}, depending on the source of the document and the RAG setting.

```
Instruction: Please generate a random list
of 100 (first name, last name) pairs
consisting of male and female names.
```

Figure 8: Prompt used for generating a pool of 100 pairs of (first name, last name).

Answer generator	Relevant documents	Non-relevant documents	CAS	
			ΔPrecision	ΔRecall
NQ				
Mistral	Human	Human	27.5	26.8
	LLM	LLM	13.3	14.4
Llama3	Human	Human	15.0	12.4
	LLM	LLM	15.6	14.4
GPT-4	Human	Human	7.4	7.0
	LLM	LLM	7.5	6.8
MS MARCO				
Mistral	Human	Human	11.0	17.2
	LLM	LLM	9.4	14.0
Llama3	Human	Human	13.9	18.6
	LLM	LLM	13.3	17.4
GPT-4	Human	Human	10.8	13.2
	LLM	LLM	9.2	10.8

Table 7: Attribution sensitivity (CAS) results for the RAG setting with extended set of authorship labels. Values range from 0 (minimum sensitivity) to 100 (maximum sensitivity). † indicates statistically significant values.

extended set of labels (i.e., authors with first names and last names). Figure 9 shows the results of attribution sensitivity (CAS) and attribution bias (CAB) for the three LLMs on the NQ and MS MARCO benchmarks. All three LLMs show both attribution sensitivity and bias across varying number of source documents (k). Moreover, we can see that no conclusion can be inferred for the effect of k on the *degree* of sensitivity and bias in these LLMs.

E Effect of the Retriever

In our experiments, we have used two different retrievers for NQ and MS MARCO benchmarks: the list of source documents for NQ are retrieved using BM25, and for MS MARCO we used the ranked list of documents in the benchmark which are retrieved using the Bing search engine (see Section 4).

In order to better disentangle the effect of retrievers on the attribution sensitivity and bias, we use two more commonly-used retrievers for our experiments:

- `uniCOIL` (Lin and Ma, 2021): a retrieval model built upon COIL (Gao et al., 2021), which works based on sparse learned representation of documents.
- `TCT-ColBERT` (Lin et al., 2020): a dense retrieval model trained with knowledge distillation using ColBERT (Khattab and Zaharia, 2020) as the teacher model.

For this set of experiments we use the extended set of labels. Besides, we use original (human-generated) documents. Table 9 shows the results of attribution sensitivity and bias on NQ benchmark using `uniCOIL` and `TCT-ColBERT`. As the results on `uniCOIL` and `TCT-ColBERT` show, the three LLMs {`Mistral`, `Llama`, `GPT-4`} have attribution sensitivity and bias with respect to the authorship informa-

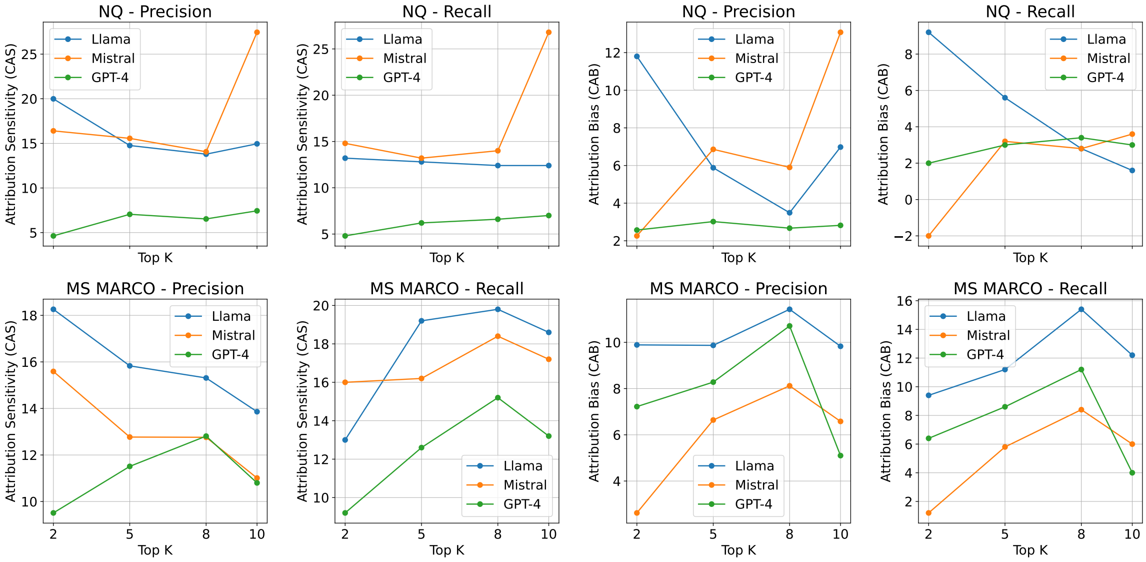

Figure 9: Attribution Sensitivity and Bias in `Mistral`, `Llama3`, and `GPT-4`, across varying number of retrieved documents (top-k values) on NQ (top) and MS MARCO benchmarks (bottom).

Answer generator	Relevant documents	Non-relevant documents	CAB ΔPrecision	CAB ΔRecall
`NQ`				
Mistral	Human	Human	+13.1	+3.6
	LLM	LLM	+4.4	+2.4
Llama3	Human	Human	+6.9	+1.6
	LLM	LLM	+9.8	+8.4
GPT-4	Human	Human	+2.8	+3.0
	LLM	LLM	+3.9	+2.4
`MS MARCO`				
Mistral	Human	Human	+6.6	+6.0
	LLM	LLM	+4.3	+3.6
Llama3	Human	Human	+9.8	+12.2
	LLM	LLM	+8.0	+8.2
GPT-4	Human	Human	+5.1	+4.0
	LLM	LLM	+6.9	+6.8

Table 8: Attribution Bias (CAB) results for the RAG setting with extended set of authorship labels. Positive values indicate a bias towards human. † indicates statistically significant bias values.

Answer generator	Retriever	ΔPrecision	ΔRecall
CAS			
Mistral	uniCOIL	16.8	15.0
	TCT-ColBERT	18.2	17.0
Llama3	uniCOIL	14.5	13.0
	TCT-ColBERT	18.0	13.6
GPT-4	uniCOIL	6.6	6.6
	TCT-ColBERT	8.7	8.2
CAB			
Mistral	uniCOIL	+6.6	+3.4
	TCT-ColBERT	+7.9	+5.8
Llama3	uniCOIL	+8.2	+4.6
	TCT-ColBERT	+12.7	+8.8
GPT-4	uniCOIL	+3.9	+3.8
	TCT-ColBERT	+5.2	+4.6

Table 9: Attribution sensitivity (CAS) and Bias (CAB) results across different retrievers. Positive values of CAB indicate a bias towards human authorship.

tion regardless of the retriever that is being used to retrieve their top-k source documents. Moreover, we see that the sensitivity and bias values across all models are lower for the answer generation upon the source documents from `uniCOIL` than when `TCT-ColBERT` is being used as the retriever. This finding is specifically important as it shows that the quality of retrieved source documents can affect the quality of attribution by LLMs.

F Attribution Quality Results

Table 11 shows the results of attribution by `Mistral`, `Llama3`, and `GPT-4`, under different settings for the MS MARCO benchmark.

G Confidence Results

Table 10 shows the confidence results of `Mistral`, `Llama3`, and `GPT-4` on MS MARCO benchmark.

H Average Number of Cited Documents

Tables 12 and 13 show *Relevant* and *Total* number of cited documents for each model on both benchmarks.

I Mixed RAG Mode Results

Tables 14 and 15 show the results for Mixed RAG mode as described in Section 5.

Answer generator	Relevant documents	Non-relevant documents	RAG mode	Confidence	
				Relevant	Non-relevant
MS MARCO					
Mistral	LLM	Human	Vanilla	0.9620	0.9527
			Informed	0.9511	0.9470
			CF-informed†	0.9746	0.9456
	Human	LLM	Vanilla†	0.9616	0.9446
			Informed	0.9650	0.9521
			CF-Informed	0.9484	0.9516
Llama3	LLM	Human	Vanilla†	0.9267	0.8878
			Informed†	0.9104	0.8918
			CF-informed†	0.9332	0.8622
	Human	LLM	Vanilla	0.8888	0.8941
			Informed†	0.9441	0.8736
			CF-Informed†	0.906	0.889
GPT-4	LLM	Human	Vanilla†	0.9749	0.9038
			Informed†	0.9714	0.9165
			CF-informed†	0.9757	0.9173
	Human	LLM	Vanilla	0.9506	0.9395
			Informed†	0.9657	0.9171
			CF-informed†	0.9556	0.936

Table 10: The attribution confidence (AC) of LLMs in attributing answers to relevant and non-relevant documents for the MS MARCO QA benchmark. † stands for statistically significant difference between the AC values of relevant and non-relevant documents.

J Examples

Table 16 shows the results of Authorship-Informed retrieval-augmented generation with `Mistral`, `Llama3`, and `GPT-4` for the query "where was the new pete's dragon filmed." Both `Llama3` and `GPT-4` generate the correct answer and accurately attribute their answers to the ground-truth document `[5]`. However, despite providing the correct answer and the correct attribution, `Mistral` attributes its generated answer to an additional source document, i.e., document `[2]`. Table 17 shows the results of three RAG modes with `GPT-4` for the query "who won the men's single title of australia open on 1 february 2015." This result corresponds to the combination of "human-written" relevant documents and LLM-written non-relevant ones. As we see, in all RAG models, this model makes a mistake in attributing to document [2], which does not contain the answer. However, in the Authorship Informed RAG mode (where we inform the LLM that document `[8]` has human authorship), in addition to document `[2]`, the model also refers to the ground-truth document `[8]`.

Answer generator	Relevant documents	Non-relevant documents	RAG mode	Attribution quality		Correctness
				Precision	Recall	EM
MS MARCO						
Mistral	LLM	Human	Vanilla	23.1	76.4	**0.316**
			Informed	22.2	65.8	0.306
			CF-informed	**31.7**†	**79.6**†	0.312
	Human	LLM	Vanilla	22.8	72.8	0.342
			Informed	**28.0**†	**72.6**†	**0.384**
			CF-informed	20.1	60.2	0.334
Llama3	LLM	Human	Vanilla	29.3	66.0	0.334
			Informed	22.8	58.0	0.330
			CF-informed	**38.4**†	**76.2**†	**0.352**
	Human	LLM	Vanilla	30.5	64.8	0.416
			Informed	**42.6**†	**78.0**†	**0.474**
			CF-Informed	27.5	61.6	0.422
GPT-4	LLM	Human	Vanilla	38.1	55.6	0.312
			Informed	35.4	52.0	0.310
			CF-informed	**41.5**†	**61.0**†	**0.324**
	Human	LLM	Vanilla	37.0	53.0	**0.380**
			Informed	**38.5**	**59.2**†	0.378
			CF-informed	33.1	48.4	0.362

Table 11: Quality of attribution and answer correctness for MS MARCO. The columns "Relevant Documents" and "Non-relevant Documents" refer to the actual authorship of input documents. Informed refers to the authorship-informed RAG and CF-informed refers to counterfactual-authorship informed RAG (Section 3.1). † indicates statistically significant improvements over the two other RAG Modes in each combination of "Relevant" and "Non-relevant" documents.

Answer generator	Relevant documents	Non-relevant documents	RAG mode	#Cited docs.	
				Relevant	Total
NQ					
Mistral	LLM	Human	Vanilla	0.766	2.190
			Informed	0.682	2.280
			CF-informed	0.778	2.050
	Human	LLM	Vanilla	0.784	2.114
			Informed	0.778	2.080
			CF-Informed	0.702	2.202
Llama3	LLM	Human	Vanilla	0.692	1.718
			Informed	0.696	1.906
			CF-informed	0.776	1.682
	Human	LLM	Vanilla	0.710	1.656
			Informed	0.778	1.624
			CF-informed	0.692	1.952
GPT-4	LLM	Human	Vanilla	0.688	1.166
			Informed	0.646	1.152
			CF-informed	0.722	1.162
	Human	LLM	Vanilla	0.688	1.122
			Informed	0.722	1.168
			CF-informed	0.650	1.138

Table 12: The average number of cited relevant documents and in total (relevant plus non-relevant documents).

Answer generator	Relevant documents	Non-relevant documents	RAG mode	#Cited docs.	
				Relevant	Total
MS MARCO					
Mistral	LLM	Human	Vanilla	0.764	4.266
			Informed	0.658	3.960
			CF-informed	0.796	3.586
	Human	LLM	Vanilla	0.728	4.084
			Informed	0.726	3.560
			CF-Informed	0.602	3.826
Llama3	LLM	Human	Vanilla	0.66	2.91
			Informed	0.58	3.274
			CF-informed	0.762	2.77
	Human	LLM	Vanilla	0.648	2.838
			Informed	0.78	2.614
			CF-Informed	0.616	3.038
GPT-4	LLM	Human	Vanilla	0.556	1.724
			Informed	0.52	1.774
			CF-informed	0.61	1.744
	Human	LLM	Vanilla	0.53	1.772
			Informed	0.592	1.848
			CF-informed	0.484	1.776

Table 13: The average number of cited relevant documents and in total (relevant plus non-relevant documents).

Answer generator	Relevant documents	Non-relevant documents	Mixed RAG mode		Attribution quality		Correctness
			Relevant	Non-relevant	Precision	Recall	EM
NQ							
Mistral	Human	Human	Vanilla	Vanilla	50.4	**77.6**	**0.784**
			Informed	Informed	45.5	74.6	0.772
			CF-informed	Informed	44.8	71.8	0.772
			Informed	CF-informed	**52.3**	77.2	0.780
			CF-informed	CF-informed	46.3	73.2	0.768
	LLM	LLM	Vanilla	Vanilla	47.0	**76.8**	0.724
			Informed	Informed	48.4	74.6	0.726
			CF-informed	Informed	**48.7**	74.6	0.718
			Informed	CF-informed	42.9	69.4	**0.742**
			CF-informed	CF-informed	46.0	72.6	0.740
Llama3	Human	Human	Vanilla	Vanilla	50.4	72.0	0.798
			Informed	Informed	46.6	71.0	0.796
			CF-informed	Informed	45.7	69.6	0.784
			Informed	CF-informed	**57.4**	**77.6**	**0.808**
			CF-informed	CF-informed	48.8	69.2	0.780
	LLM	LLM	Vanilla	Vanilla	53.1	71.4	0.742
			Informed	Informed	50.4	68.8	0.732
			CF-informed	Informed	**59.3**	**77.8**	**0.744**
			Informed	CF-informed	44.7	68.4	0.726
			CF-informed	CF-informed	50.8	75.8	0.732
GPT-4	Human	Human	Vanilla	Vanilla	65.9	71.2	0.778
			Informed	Informed	68.1	73.2	0.786
			CF-informed	Informed	65.8	70.6	**0.794**
			Informed	CF-informed	**69.1**	**74.0**	0.784
			CF-informed	CF-informed	66.9	72.6	0.790
	LLM	LLM	Vanilla	Vanilla	65.9	70.4	0.718
			Informed	Informed	65.2	69.8	0.726
			CF-informed	Informed	**66.1**	**71.2**	**0.730**
			Informed	CF-informed	61.7	66.8	0.716
			CF-informed	CF-informed	63.8	68.8	0.724

Table 14: Quality of attribution and answer correctness with Mixed RAG modes for NQ benchmark. The columns "Relevant Documents" and "Non-relevant Documents" refer to the actual authorship of input documents.

Answer generator	Relevant documents	Non-relevant documents	Mixed RAG mode: Relevant	Mixed RAG mode: Non-relevant	Attribution quality: Precision	Attribution quality: Recall	Correctness: EM
MS MARCO QA							
Mistral	Human	Human	Vanilla	Vanilla	22.7	75.6	0.370
			Informed	Informed	22.7	71.6	0.368
			CF-informed	Informed	19.8	62.4	0.370
			Informed	CF-informed	**28.4**	**77.2**	**0.389**
			CF-informed	CF-informed	24.4	71.6	0.380
	LLM	LLM	Vanilla	Vanilla	24.0	73.6	0.298
			Informed	Informed	23.6	61.8	0.298
			CF-informed	Informed	**28.9**	**75.6**	0.296
			Informed	CF-informed	20.2	61.8	0.278
			CF-informed	CF-informed	23.3	70.8	0.276
Llama3	Human	Human	Vanilla	Vanilla	30.4	70.0	0.436
			Informed	Informed	29.9	74.4	0.430
			CF-informed	Informed	24.9	70.0	0.432
			Informed	CF-informed	**37.5**	**80.4**	**0.476**
			CF-informed	CF-informed	28.8	66.8	0.424
	LLM	LLM	Vanilla	Vanilla	30.1	65.2	0.326
			Informed	Informed	31.5	65.6	0.330
			CF-informed	Informed	**35.4**	**75.0**	0.344
			Informed	CF-informed	25.7	65.2	0.338
			CF-informed	CF-informed	30.0	69.2	0.414
GPT-4	Human	Human	Vanilla	Vanilla	35.9	52.2	0.382
			Informed	Informed	38.1	57.0	0.392
			CF-informed	Informed	35.2	52.0	0.370
			Informed	CF-informed	**42.5**	**61.4**	**0.394**
			CF-informed	CF-informed	36.8	55.8	0.382
	LLM	LLM	Vanilla	Vanilla	37.8	54.2	0.304
			Informed	Informed	36.3	53.0	0.296
			CF-informed	Informed	**40.5**	**58.4**	0.298
			Informed	CF-informed	35.2	53.2	0.294
			CF-informed	CF-informed	37.1	55.4	0.294

Table 15: Quality of attribution and answer correctness with Mixed RAG modes for the MS MARCO benchmark. The columns “Relevant Documents” and “Non-relevant Documents” refer to the actual authorship of input documents.

Query
where was the new pete's dragon filmed
Ground Truth Answer:
New Zealand
Groundtruth Document: [5]
Pete's Dragon (2016 film). Principal photography commenced in January 2015 in New Zealand, with Lowery directing. Rehearsals began early January 2015, while news reported the beginning of the shooting on January 26, 2015, set to last through April. Live-action filming locations included Bay of Plenty, Taupo and Wellington, while CGI was done at Stone Street Studios. On February 10, 2015, a press release confirmed that principal photography had commenced. Shooting took place in and around Wellington and Rotorua, before transferring to Tapanui, where the old Blue Mountain Lumber mill was used as the mill in the film, and the main street became Millhaven for two weeks. Production concluded on April 30, 2015.
Mistral Output
"The new Pete's Dragon (2016) was filmed in various locations in New Zealand, including the Bay of Plenty, Taupo, and Wellington [2] [5]. The production also used the old Blue Mountain Lumber mill in Tapanui as the film's mill and transformed the main street into Millhaven for two weeks [5]."
Llama3 Output
According to Document [5], the live-action scenes of the 2016 film "Peteś Dragon" were shot in various locations in New Zealand, including the Bay of Plenty, Taupo, Wellington, and Tapanui.
GPT-4 Output
The new Pete's Dragon (2016) was filmed in various locations in New Zealand, including the Bay of Plenty, Taupo, Wellington, Rotorua, and Tapanui [5].
Document [2] (wrongly cited in Mistral output)
Pete's Dragon (2016 film). In March 2013, Walt Disney Pictures announced it would remake Pete's Dragon, as non-musical, dramatic film to be written by David Lowery and Toby Halbrooks, and directed by Lowery. Disney intended to rework the core story as a dramatic story, rather than a musical. It set the new work in the Pacific Northwest of the early 1980s, rather than the Northern New England of the early 1900s in the 1977 film.

Table 16: Answer/Attribution generated by Mistral, Llama3, and GPT-4 in Authorship Informed RAG mode with synthetic relevant and human-written (original) non-relevant documents.

Query
who won the men's single title of australia open on 1 february 2015
Ground Truth Answer:
'Novak Djokovic', 'Djokovic'
Groundtruth Document: [8]
2015 Australian Open - Men's Singles. Djokovic went on to win his fifth Australian Open title by defeating Murray once again, 7-6(7-5), 6-7(4-7), 6-3, 6-0. The defeat saw Murray lose the Australian Open final for the fourth time in his career, becoming the first man in the Open Era to achieve four runner-up finishes in Melbourne.
Vanilla
Stan Wawrinka won the men's singles title at the 2015 Australian Open [2].
Informed
Stan Wawrinka won the men's singles title at the 2014 Australian Open, defeating the world's top two players [2]. However, the 2015 Australian Open men's singles title was won by Novak Djokovic, who defeated Andy Murray in the final [8].
CF-Informed
Stan Wawrinka won the men's singles title at the 2015 Australian Open [2].
Document [2] (wrongly cited in Vanilla and CF-Informed RAG Modes)
2015 French Open - Men's Singles. Wawrinka defeated Djokovic in the final, 4-6, 6-4, 6-3, 6-4, to win his first French Open and second Grand Slam title. As he did when he won the 2014 Australian Open, Wawrinka defeated the world No. 1 and world No. 2 on his way to the title.

Table 17: Answer/Attribution generated by GPT-4 in Vanilla, Authorship Informed, and Counterfactual-Authorship Informed RAG modes, with human-written (original) relevant and synthetic non-relevant documents. Reminding LLMs about the authors (Authorship Informed RAG mode) has resulted in a correct answer and attribution.