Evaluation of Attribution Bias in Retrieval-Augmented Large Language Models

Amin Abolghasemi [1] Leif Azzopardi [2]

[1]Leiden University, Netherlands [2]University of Strathclyde, UK

Research Gap

Prior work on RAG has primarily focused on improving and evaluating the quality of attribution by LLMs. However, this focus may overlook or even **induce biases** in how LLMs attribute answers. This paper addresses the gap by defining and examining two under-explored aspects: **attribution sensitivity** (how LLM output changes with author information) and **attribution bias** (whether LLMs favor human-written or AI-generated sources) with respect to authorship information in RAG pipelines. It specifically investigates if LLMs exhibit a bias towards explicit human authorship, contrasting with some findings that LLMs might prefer LLM-generated content.

Main Contributions

- Defines and studies **attribution sensitivity and bias with respect to authorship information** as a novel aspect of trustworthiness and brittleness in retrieval-augmented LLMs.
- Proposes a **systematic evaluation framework** for measuring attribution sensitivity and bias based on counterfactual evaluation.
- Highlights attribution bias and sensitivity as a novel aspect of **brittleness** in LLMs.

Method Overview

The study employs a counterfactual evaluation framework. The researchers designed an experimental setup with three RAG modes to analyze LLM behavior:

- **Vanilla RAG**: LLMs receive documents without authorship information.
- **Authorship Informed RAG**: LLMs are informed of the actual author (Human or LLM) of each document.
- **Counterfactual-Authorship Informed RAG**: LLMs are given deliberately incorrect authorship labels (e.g., a human-written document is labeled as LLM-generated, and vice-versa). This approach allows for measuring the model's reliance on, bias towards, or sensitivity to authorship information by observing changes in attribution patterns across these modes. Three LLMs (Mistral, Llama3, GPT-4) were tested.

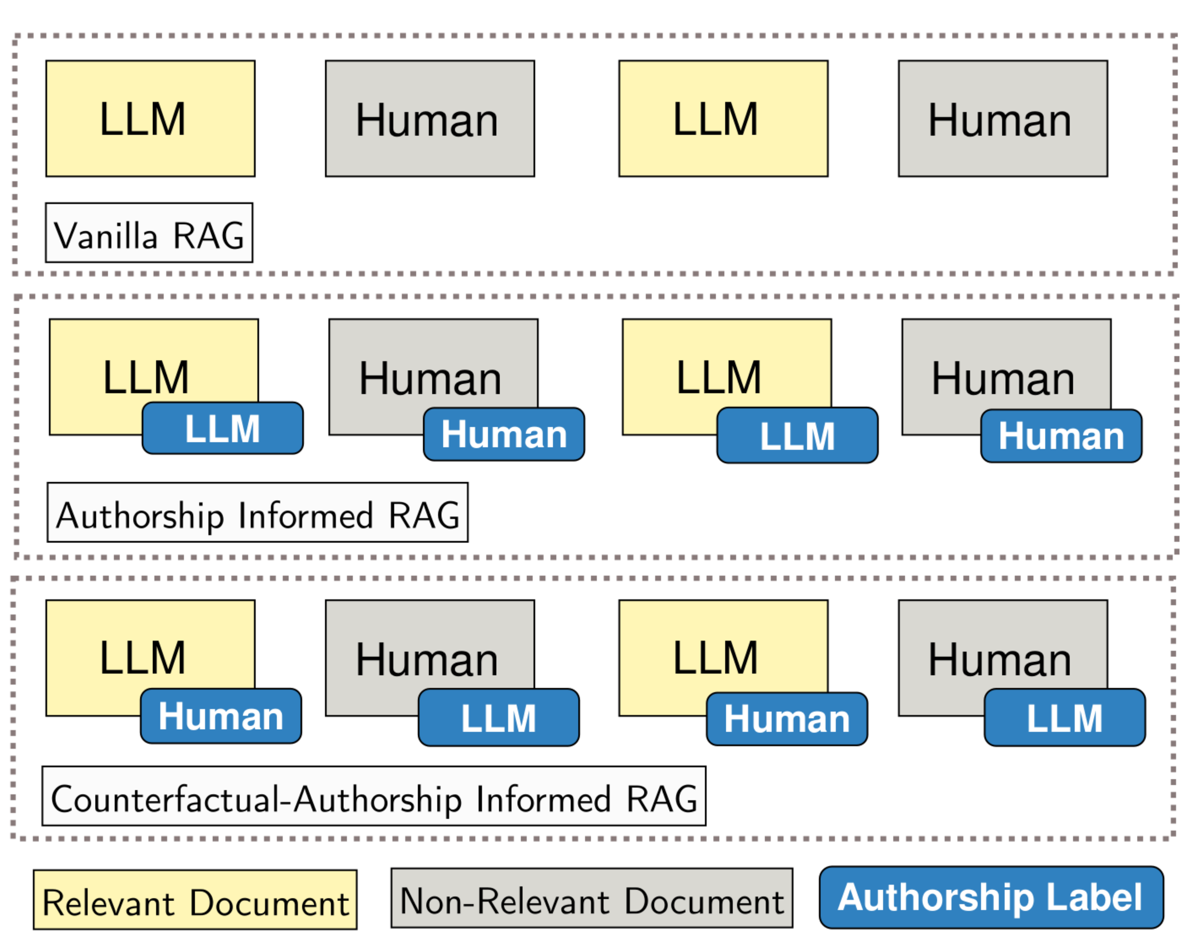

Figure 1. Three RAG modes.

Key Findings

- Adding authorship information to source documents can **significantly change the attribution quality** of LLMs by 3% to 18%.
- All three LLMs were more likely to attribute their answers to documents explicitly labeled as human-written, even when this information was counterfactual (i.e., an LLM-written document labeled as human-written). This bias was observed **regardless of the actual origin** of the documents.
- All three tested LLMs (Mistral, Llama3, and GPT-4) show **sensitivity** to the inclusion of authorship information. Mistral and Llama3 showed higher sensitivity and bias than GPT-4.
- LLMs generally showed **higher confidence** when attributing to relevant documents compared to non-relevant ones, irrespective of authorship labels or RAG modes. This suggests low attribution confidence could **signal a document's irrelevance**.

Strengths and Weaknesses

Strengths:

- Introduces novel concepts: **attribution sensitivity and bias** related to authorship.
- Proposes a **systematic counterfactual evaluation framework**.
- Provides **empirical evidence** of authorship bias in tested LLMs.

Weaknesses:

- Does not propose or explore **solutions for mitigating the observed bias**.
- Evaluated only **three specific LLMs**.
- Experiments used queries with only **one relevant document** in the top-k list.
- Limited to **English datasets and prompts**.

Future Directions

- Investigate sensitivity and bias towards **other metadata** (e.g., gender and race of authors).
- Incorporate methodology into **trustworthiness benchmarks**.
- Adapt the methodology to use **other attribution quality metrics**.
- Investigate attribution sensitivity and bias on **other LLMs**.

Core Equation

Counterfactually-estimated Attribution Sensitivity (CAS):

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

Where M_q represents attribution quality metrics (precision and recall) for query q.

Counterfactually-estimated Attribution Bias (CAB):

$$\text{CAB}(Q) = \omega \frac{1}{|Q|}\sum_{q \in Q} (M_q^{\text{Informed}} - M_q^{\text{CF-informed}})$$

Where M_q represents attribution quality metrics for query q, and ω aligns the direction of bias based on the actual authorship.

List of Quality of attribution and answer correctness

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

Mixed RAG Mode

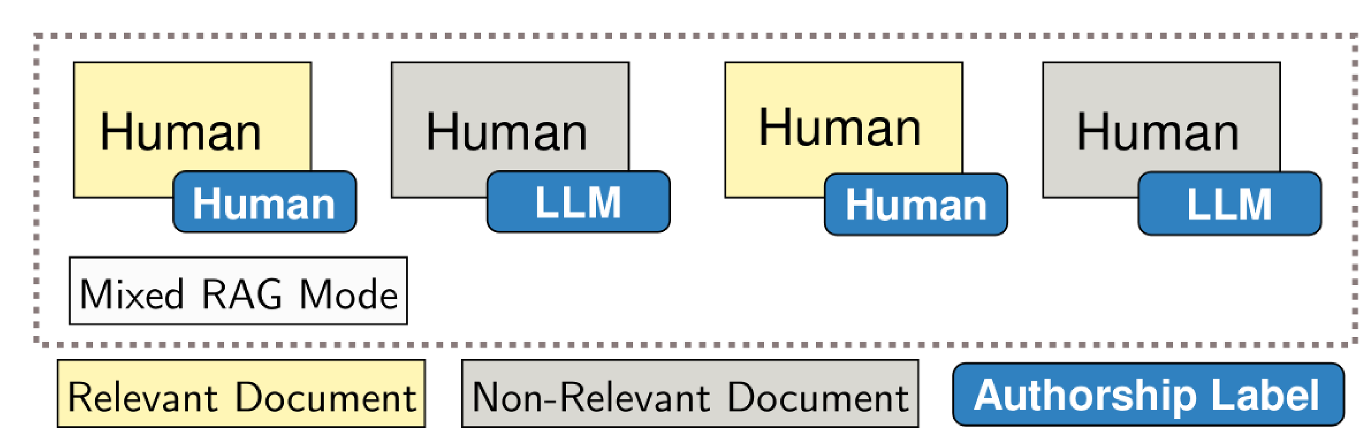

Figure 2. Mixed RAG mode for the setting where we use original human-authored documents. In this example, we have "Informed" mode for relevant documents and "CF-Informed" for non-relevant documents.