CrAM: Credibility-Aware Attention Modification in LLMs for Combating Misinformation in RAG

Boyi Deng[1], Wenjie Wang[2*], Fengbin Zhu[2*], Qifan Wang[3], Fuli Feng[1]
[1]University of Science and Technology of China,
[2]National University of Singapore,
[3]Meta AI
dengboyi@mail.ustc.edu.cn, wqfcr@fb.com,
{wenjiewang96, zhfengbin, fulifeng93}@gmail.com

Abstract

Retrieval-Augmented Generation (RAG) can alleviate hallucinations of Large Language Models (LLMs) by referencing external documents. However, the misinformation in external documents may mislead LLMs' generation. To address this issue, we explore the task of "credibility-aware RAG", in which LLMs automatically adjust the influence of retrieved documents based on their credibility scores to counteract misinformation. To this end, we introduce a plug-and-play method named **Cr**edibility-aware **A**ttention **M**odification (CrAM). CrAM identifies influential attention heads in LLMs and adjusts their attention weights based on the credibility of the documents, thereby reducing the impact of low-credibility documents. Experiments on Natual Questions and TriviaQA using Llama2-13B, Llama3-8B, and Qwen1.5-7B show that CrAM improves the RAG performance of LLMs against misinformation pollution by over 20%, even surpassing supervised fine-tuning methods.

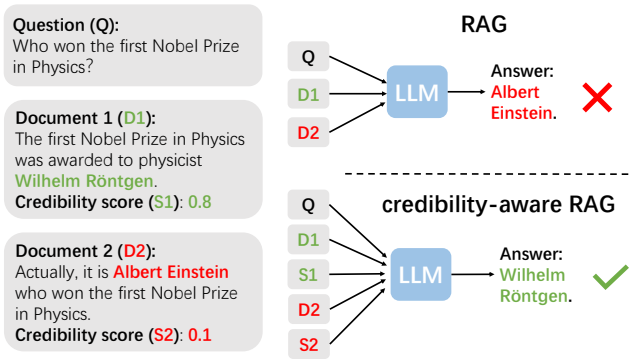

Figure 1: A comparison between RAG and credibility-aware RAG. Credibility-aware RAG considers credibility to reduce the impact of low-credibility documents.

1 Introduction

Retrieval-Augmented Generation (RAG) (Gao et al. 2024; Zhu et al. 2021) is a representative approach to mitigate hallucination issues of Large Language Models (LLMs) (Zhang et al. 2023) by retrieving and referencing relevant documents from an external corpus. Despite its effectiveness, most RAG works overlook a crucial issue: misinformation pollution in the external corpus (Pan et al. 2023b; Dufour et al. 2024). The maliciously generated misinformation may mislead LLMs to produce unfaithful responses. For instance, Microsoft's Bing can be misled by misinformation on the internet to generate incorrect information for Bing users (Vincent 2023). Besides, Pan et al. (2023b) and Pan et al. (2023a) demonstrated that inserting LLM-generated misinformation into the RAG corpus can significantly degrade LLMs' performance. Therefore, addressing the misinformation pollution for RAG is essential.

A straightforward and common idea to address this misinformation pollution issue is misinformation detection and filtering. Extensive misinformation detection works focus on measuring the *credibility* of documents, *i.e.,* the probability of the document not containing misinformation. And these works have achieve significant results (Kaliyar, Goswami, and Narang 2021; Pelrine et al. 2023; Quelle and Bovet 2024; Li et al. 2024). Once we obtain the credibility of each retrieved document, we can exclude those with credibility below a certain threshold before using them in RAG. However, directly discarding certain documents may result in the loss of relevant and important information, leading to performance degradation (Yoran et al. 2024)[1]. Therefore, given the remarkable advancements in the measurement of credibility scores and the relatively underdeveloped mechanisms for utilizing these scores, it is essential to explore how these scores can be effectively utilized by LLMs, assuming that high-quality credibility scores are accessible.

To achieve this, we focus on a task named "credibility-aware RAG" as shown in Figure 1. Specifically, given a user query x with a list of relevant documents $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ and $\mathcal{D}$'s credibility scores $\mathcal{S} = \{s_1, s_2, ..., s_n\}$, credibility-aware RAG requests LLMs to automatically adjust the influence of documents in $\mathcal{D}$ on the generated output y based on their credibility scores in $\mathcal{S}$. Initial attempts on credibility-aware RAG adopted supervised fine-tuning (SFT) to teach LLMs to distinguish the importance of different documents in the prompt by their credi-

*Corresponding author.

[1]Our experimental results in Table 2 also confirm that directly excluding documents leads to inferior performance.

bility scores (Hong et al. 2024; Pan et al. 2024). However, SFT requires additional computational resources and well-designed training data, which limits the application scenarios. Therefore, we explore non-SFT method for LLMs to attain credibility-aware RAG.

Given that the attention mechanism serves as the central component for adjusting the significance of various input data, we consider manipulating attention weights of LLMs to achieve credibility-aware RAG. In particular, we adjust attention weights according to credibility scores in the inference stage of LLMs. In this way, we can regulate LLMs to pay less "attention" to less credible documents by decreasing the corresponding attention weights. Moreover, previous studies (Clark et al. 2019; Elhage et al. 2021; Voita et al. 2019) have indicated that different attention heads exhibit distinct patterns and functions, resulting in varying impacts on LLMs' outputs. In this context, the key lies in identifying a subset of influential attention heads for attention weight modification.

In this work, we propose a plug-and-play method named **Cr**edibility-aware **A**ttention **M**odification (CrAM), which identifies the influential attention heads and then modifies their attention weights *w.r.t.* different document tokens to reduce the impact of low-credibility documents. Specifically, *1) influential head identification:* we select top-ranked attention heads according to an extended causal tracing method (Meng et al. 2022) that estimates the contribution of each attention head to generating incorrect answers over a small dataset. *2) Attention weight modification:* we scale down the attention weights of the retrieved documents based on their normalized credibility scores.

We conduct extensive experiments on two open-domain Question Answering (QA) datasets, Natual Questions (NQ) (Kwiatkowski et al. 2019) and TriviaQA (Joshi et al. 2017), using three open-source LLMs: Llama2-13B (Touvron et al. 2023), Llama3-8B (Meta 2024), and Qwen1.5-7B (Bai et al. 2023). The results show that CrAM significantly alleviates the influence of misinformation documents on RAG, in terms of both ideal credibility scores and GPT-generated credibility scores. It is worth noting that CrAM even outperforms the SFT-based method CAG (Pan et al. 2024) in most scenarios, demonstrating the superiority of CrAM. We release our code at https://github.com/Aatrox103/CrAM.

In summary, our main contributions are:

- We explore the task of credibility-aware RAG without fine-tuning LLMs to alleviate the misinformation pollution issue.
- We develop a plug-and-play method, CrAM, which identifies influential attention heads and modifies their attention weights to equip LLMs with credibility-aware RAG capabilities.
- We conduct extensive experiments with two QA datasets on three LLMs using ideal credibility scores and GPT-generated credibility scores, validating the superiority of CrAM.

2 Credibility-Aware RAG

Given a user query x, RAG retrieves a set of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ relevant to x through a retriever (Gao et al. 2024). Then the relevant documents $\mathcal{D}$ are evaluated by a credibility estimator[2], obtaining their credibility scores $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, which represents the probability of each document not containing misinformation.

Credibility-Aware RAG. Given an LLM L, a user query x, and relevant documents $\mathcal{D}$ associated with credibility scores $\mathcal{S}$, the objective of credibility-aware RAG is to enable LLMs to automatically adjust the influence of these documents on the generated output y based on their credibility scores $\mathcal{S}$. This can be formally defined as:

$$\max \ \mathrm{Metric}(\mathrm{Combine}(L, x, \mathcal{D}, \mathcal{S})),$$

where $\mathrm{Combine}(\cdot)$ represents the method or mechanism to integrate credibility scores into the generation process of L. For example, Pan et al. (2024) employ SFT to fine-tune LLMs to capture the credibility difference of documents more effectively, denoted as $\mathrm{Combine}(L, x, \mathcal{D}, \mathcal{S}) = L_{SFT}(x, \mathcal{D}, \mathcal{S})$. Additionally, $\mathrm{Metric}(\cdot)$ is a function that assesses whether documents with different credibility scores have varying impacts on the output of L. Indeed, we can utilize the performance of generating factual answers to measure $\mathrm{Metric}(\cdot)$. For instance, we use the accuracy of QA tasks to approximate $\mathrm{Metric}(\cdot)$ in this work. The rationality is that if the impact of low-credibility documents decreases, the accuracy of QA tasks should increase accordingly.

3 CrAM

CrAM first identifies influential attention heads, and then modifies the attention weights of these identified heads to reduce the impact of low-credibility documents as shown in Figure 2. Since influential attention heads identification process involves attention weight modification, we first explain the procedure of attention weight modification in Section 3.1, and then describe influential attention heads identification in Section 3.2. Finally, we summarize the overall CrAM workflow in Section 3.3.

3.1 Attention Weight Modification

As defined in Section 2, the objective of credibility-aware RAG is to reduce the impact of low-credibility documents on the generated output of LLMs. Intuitively, it requires LLMs to pay less "attention" to low-credibility documents. To this end, a natural approach is scaling down the corresponding attention weights of low-credibility documents.

For RAG, a user query x and a set of relevant documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ should be concatenated and tokenized into a token sequence $\mathcal{T}(x, \mathcal{D}) = \{t_1, t_2, \ldots, t_m\}$, where t_k denotes the k-th token. Given the credibility scores for each document $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, the normalized credibility score for token t_k can be calculated as follows:

$$\bar{s}_k = \begin{cases} \frac{s_i - \min(\mathcal{S})}{\max(\mathcal{S}) - \min(\mathcal{S})} & \text{if } t_k \text{ belongs to } d_i \\ 1 & \text{otherwise} \end{cases},$$

[2]Recent worked on this task has achieved promising performance (Kaliyar, Goswami, and Narang 2021; Pelrine et al. 2023).

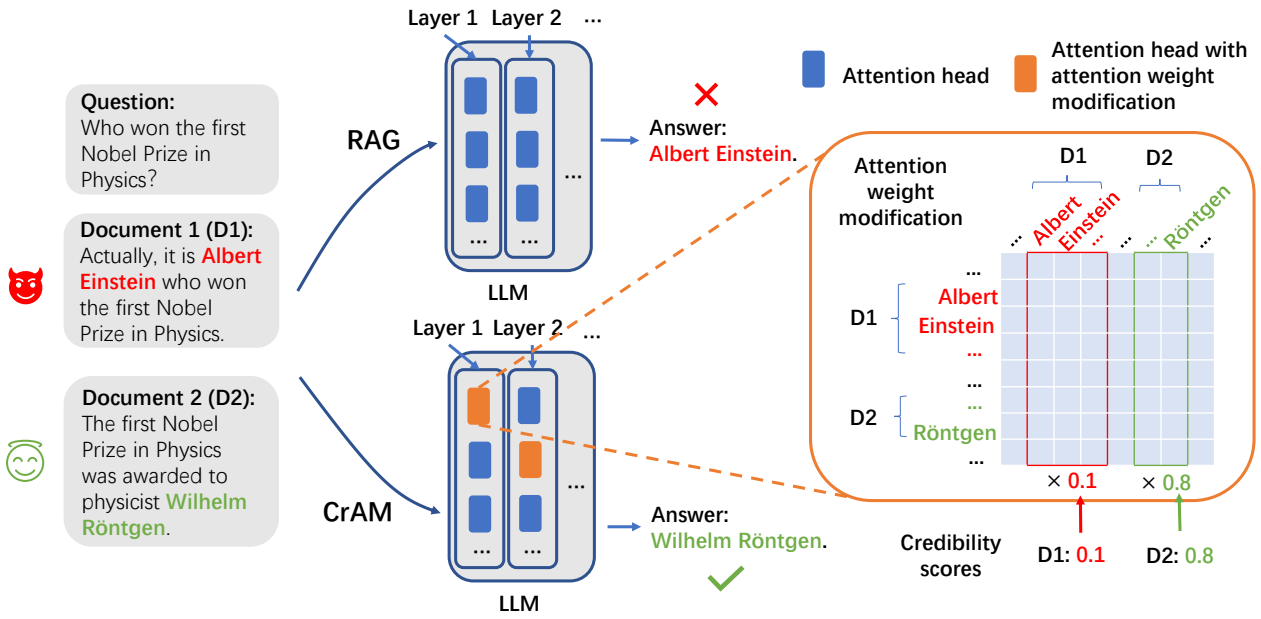

Figure 2: Illustration of CrAM. Compared to RAG, CrAM first identifies influential attention heads and then modifies their attention weights based on the credibility scores of each document.

where s_i is subtracted by $\min(\mathcal{S})$, and then scaled down by $1/(\max(\mathcal{S}) - \min(\mathcal{S}))$ to ensure all credibility scores are normalized to $[0, 1]$. Besides, we define $\bar{\mathbf{s}} = [\bar{s}_1, \ldots, \bar{s}_m] \in \mathbb{R}^{1\times m}$ as the normalized credibility scores of the whole token sequence $\mathcal{T}(x, \mathcal{D})$.

For each attention head h in LLM, $\mathbf{A}_h$ represents its attention weights matrix[3]. Let $(\mathbf{A}_h)_k$ represent the k-th row vector[4] of $\mathbf{A}_h$, we can obtain the modified attention weight matrix $\mathbf{A}_h^*$ by element-wise multiplying $\bar{\mathbf{s}}$ as follows:

$$(\mathbf{A}_h)_k^* = \text{Norm}((\mathbf{A}_h)_k \odot \bar{\mathbf{s}}), k \in \{1, \ldots, m\}, \quad (1)$$

where $\odot$ denotes the element-wise multiplication of vectors. The Norm function refers to ℓ_1 normalization, which ensures that the attention weights sum to one.

3.2 Influential Head Identification

Previous works Clark et al. (2019); Elhage et al. (2021); Voita et al. (2019) have found that different attention heads exhibit various patterns and functions, leading to different impacts on LLMs' output. As such, we hypothesize that some attention heads have a larger impact on using misinformation documents to generate incorrect answers. Previously, causal tracing (Meng et al. 2022) has been developed to quantify the contribution of each hidden state towards generating given answers. The contribution is measured by adding noises to each hidden state to compare the changes in the generation probability of the given answer. In light of this, CrAM revises causal tracing to evaluate the contribution of attention heads instead of hidden states. Utilizing attention weight modification, as detailed in Section 3.1, CrAM estimates the change in probability of generating incorrect answers to determine the contribution of each attention head. Thereafter, CrAM ranks all attention heads by contributions and identifies influential ones.

Specifically, the contribution of one attention head h can be obtained as follows:

- Given an LLM L, a user query x, a set of relevant documents $\mathcal{D} = \{d_{mis}, d_1, d_2, \ldots, d_n\}$ with one misinformation document d_{mis}, and an incorrect answer a_{wrong} to x that is supported by d_{mis}, we first calculate the generation probability of a_{wrong} with x and $\mathcal{D}$ by L. Formally, we have:
$$P_0 = P_L(a_{wrong} \mid x, \mathcal{D}).$$
- Next, we modify a specific attention head as described in Section 3.1 by using the credibility scores $\mathcal{S} = \{0, 1, 1, \ldots, 1\}$ of $\mathcal{D}$ and recalculate the generation probability of a_{wrong}:
$$P_1 = P_{L_h^*}(a_{wrong} \mid x, \mathcal{D}),$$
where L_h^* denotes the LLM L whose attention weight matrix of the attention head h is modified according to Equation (1).
- Finally, we quantify the contribution of head h towards generating the incorrect answer, *a.k.a.* the indirect effect (IE) (Meng et al. 2022):
$$\text{IE}_h = P_0 - P_1, \quad (2)$$
which can also be interpreted as the decrease in the generation probability of the incorrect answer a_{wrong} after modifying head h.

[3]The attention weights matrix is defined in Equation (3).

[4]$(\mathbf{A}_h)_k$ can be interpreted as the attention weight vector when using the k-th token as the query.

To improve the robustness of the contribution estimation, we utilize a small dataset $\{(x, a_{wrong}, \mathcal{D}, \mathcal{S}), \ldots\}$ that do not overlap with the test data to compute the average IE for each attention head (refer to Section 4.3 for robustness analysis). Thereafter, we can calculate IEs for all the attention heads and rank them to select the top-ranked ones with larger IEs for attention weight modification.

3.3 CrAM Workflow

The CrAM workflow is summarized as follows:

- First, we use a small dataset with misinformation-polluted documents to calculate the average IE for each attention head in an LLM as described in Section 3.2. Then, we rank all attention heads by their IEs in descending order and select the top-ranked heads as influential attention heads.
- Given any user query, along with the relevant documents and credibility scores, we modify the attention weights of influential attention heads using the method described in Section 3.1 to obtain the final answer, thereby significantly reducing the impact of low-credibility documents.

4 Experiments

4.1 Experimental Settings

Datasets, LLMs and Metrics. We conduct experiments over the Natural Questions (NQ) (Kwiatkowski et al. 2019) and TriviaQA (Joshi et al. 2017) datasets with three LLMs, i.e. Llama2-13B (Touvron et al. 2023), Llama3-8B (Meta 2024), and Qwen1.5-7B (Bai et al. 2023). We adopt Exact Match (EM) and F1 score as evaluation metrics, which are widely used in the QA setting (Karpukhin et al. 2020; Rajpurkar et al. 2016; Chen et al. 2017).

Document Preparation. We prepare both high-credibility and low-credibility documents (i.e., with misinformation) associated with the questions for evaluating the proposed method. 1) *High-credibility documents* are collected by retrieving the most relevant documents from the external corpus for each question. Specifically, we first employ `bge-large-en-v1.5`[5] to obtain a set of candidates from the Wikipedia dump on December 30, 2018 (Karpukhin et al. 2020). Then, we apply `bge-reranker-large`[6] to rank the retrieved candidates and select the top four documents. 2) *Low-credibility documents* are generated via prompting LLMs (i.e., gpt-3.5-turbo-0125), with misinformation included, similar to the practice in previous works (Pan et al. 2023a,b, 2024; Hong et al. 2024; Chen and Shu 2024). Specifically, given a question, we instruct the LLM to generate a news-style piece containing misinformation that supports an incorrect answer, which is regarded as one low-credibility document for the question. For each question, we collect three distinct low-credibility documents, all supporting the same incorrect answer. The prompts can be found in Appendix H.

[5]huggingface.co/BAAI/bge-large-en-v1.5.

[6]huggingface.co/BAAI/bge-reranker-large.

In our implementation, we combine generated low-credibility documents with retrieved high-credibility documents as input for the LLM. This approach avoids injecting low-credibility documents directly into the corpus, which can lead to inputs that are either overwhelmed by misinformation or completely devoid of it. In contrast, our method provides greater control, enabling us to effectively evaluate the impact of varying amounts of low-credibility documents on the LLM's performance.

Credibility Scores Generation. We adopt two different ways to assign credibility scores for each document. 1) *Ideal Setting.* After obtaining the high-credibility and low-credibility documents, we assign a score of 10 to each high-credibility document and a score of 1 to each low-credibility document. 2) *GPT Setting.* We employ GPT (i.e., gpt-3.5-turbo-0125) to directly generate the credibility score for each document. The prompts and the distribution of GPT-generated scores for all documents are provided in Figure 20 and Appendix C.

Compared Methods. We compare our CrAM model with four types of methods: 1) *Naive RAG.* The Naive RAG follows the standard RAG pipeline without any mechanisms against misinformation. 2) *Prompt Based.* This method directly informs the LLM of the credibility score via prompts, feeding the score and documents into the LLM without additional training. 3) *Exclusion.* This method excludes the documents with credibility scores below a threshold. This method will not be compared under the ideal setting due to the binary value of the ideal credibility score. 4) *CAG.* This method is proposed by Pan et al. (2024), which directly incorporates credibility scores and documents into prompts to fine-tune an LLM (i.e., Llama2-13B) to lift its understanding capabilities. Among them, Naive RAG, Prompt Based, and Exclusion are non-SFT methods, while CAG is an SFT-based method.

Hyperparameters. Unless otherwise specified, in the following experiments, we randomly select 100 data points from each dataset to calculate average IE for all the heads. And we use another validation set of 100 data points from each dataset to determine how many top-ranked heads should be included in the final modified set.

4.2 Main Results

Comparison with Non-SFT Methods. We first compare our CrAM model with Non-SFT methods, i.e., Naive RAG, Prompt Based, and Exclusion. Table 1 and Table 2 show the experimental results in the Ideal and GPT settings respectively. We make the following observations. 1) Table 1 demonstrates that our CrAM method significantly outperforms all compared methods across all three LLMs: Qwen1.5-7B, LLama2-13B, and LLama3-8B, on both NQ and TriviaQA datasets in the setting of 4 ✓+ 1 ✗ (i.e., four high-credibility documents plus one low-credibility document). For instance, our CrAM model surpasses the second-best method, i.e. Prompt Based, by 25.5%, 31.90% and 10.9% on Qwen1.5-7B, Llama2-13B and Llama3-8B in terms of EM on TriviaQA, demonstrating remarkable per-

Model	In-context corpus	Method	NQ		TriviaQA	
			EM	F1 score	EM	F1 score
Qwen1.5-7B	0 ✓	Naive LLM	7.20	16.41	28.00	38.23
	4 ✓	Naive RAG	27.60	39.08	55.30	66.85
	4 ✓ + 1 ✗	Naive RAG	10.50	20.71	25.00	35.63
		Prompt Based	12.20	22.26	27.40	37.98
		CrAM	**29.10** (+16.90)	**41.02** (+18.76)	**52.90** (+25.50)	**64.16** (+26.18)
Llama2-13B	0 ✓	Naive LLM	20.30	28.59	50.40	57.56
	4 ✓	Naive RAG	28.90	39.98	62.50	71.03
	4 ✓ + 1 ✗	Naive RAG	11.90	19.97	28.00	36.22
		Prompt Based	12.50	22.94	23.10	32.70
		CrAM	**33.60** (+21.10)	**44.62** (+21.68)	**59.90** (+31.90)	**67.11** (+30.89)
Llama3-8B	0 ✓	Naive LLM	20.60	30.58	55.70	62.67
	4 ✓	Naive RAG	33.10	45.66	64.30	73.68
	4 ✓ + 1 ✗	Naive RAG	16.00	26.16	36.80	47.09
		Prompt Based	29.90	39.69	53.50	63.01
		CrAM	**36.90** (+7.00)	**48.45** (+8.76)	**64.40** (+10.90)	**73.49** (+10.48)

Table 1: Main results under ideal setting. 0 ✓ indicates no document and the model directly prompted, 4 ✓ indicates all four documents retrieved from the Wikipedia dump, and 4 ✓ + 1 ✗ indicates four high-credibility documents (i.e., retrieved from external corpus) plus one low-credibility document (i.e., containing misinformation). In the 4 ✓ + 1 ✗ setting, the best performance is highlighted in **bold**. And the red part indicates the difference between CrAM and second best performance.

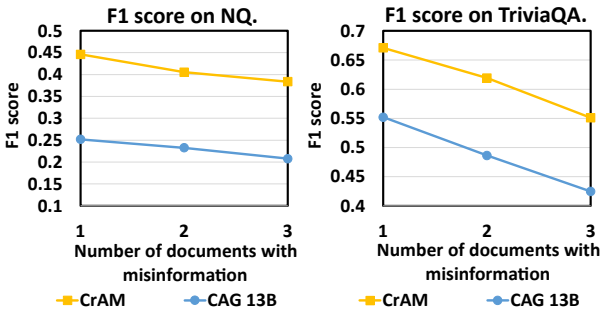

Figure 3: Performance comparison of CrAM and CAG-13B regarding the varying number of documents containing misinformation under ideal setting.

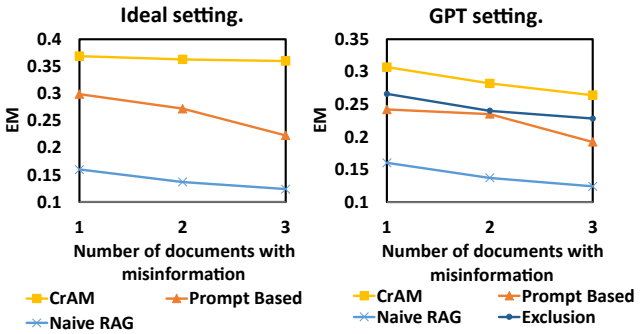

Figure 4: Performance change on NQ regarding the varying number of documents with misinformation.

formance gains. 2) With GPT-generated credibility scores, our CrAM model also outperforms all compared methods on all three LLMs over both NQ and TriviaQA datasets, as shown in Table 2, further highlighting its effectiveness. 3) Interestingly, we find that our CrAM model with 4 ✓ + 1 ✗ sometimes even outperforms the Naive RAG with 4 ✓ under ideal setting. This is likely because our generated misinformation includes both affirmations of incorrect information and denials of correct information, e.g."The first person to win the Nobel Prize in Physics was not Roentgen, but Einstein." This allows LLMs to reuse the correct information denied by the misinformation. To further validate this hypothesis, we conduct additional experiments and present the findings in Appendix F.

Comparison with SFT-based Method. For a fair comparison, we only compare our Llama2-13B based CrAM model with CAG-13B, because CAG-13B is trained on Llama2-13B. Moreover, to verify the robustness of our CrAM model, we perform comparisons using different numbers of low-credibility documents. As shown in Figure 3, our CrAM model consistently outperforms the CAG-13B model remarkably in terms of F1 score when the number of low-credibility documents ranges from 1 to 3. The results further prove the effectiveness of our CrAM model.

4.3 In-Depth Analysis

Effect of Number of Low-credibility Documents. In the following, we analyze the effect of varying the number of low-credibility documents fed into the LLM. We conduct experiments using Llama3-8B on the NQ dataset. Specifically, we vary the number of low-credibility documents from 1 to 3 while keeping the number of high-credibility documents constant, i.e., 4. We present the experimental results in Figure 4. From the figure, we make the following observations. 1) Our CrAM model consistently outperforms the compared models when changing the number of low-credibility documents from 1 to 3 in both ideal and GPT settings. 2) Compa-

Model	In-context corpus	Method	NQ		TriviaQA	
			EM	F1 score	EM	F1 score
Qwen1.5-7B	0 ✓	Naive LLM	7.20	16.41	28.00	38.23
	4 ✓	Naive RAG	27.60	39.08	55.30	66.85
	4 ✓ + 1 ✗	Naive RAG	10.50	20.71	25.00	35.63
		Prompt Based	12.50	22.98	29.70	40.18
		Exclusion	21.60	32.56	49.50	61.03
		CrAM	**23.10** (+1.50)	**34.84** (+2.28)	**52.10** (+2.60)	**63.76** (+2.73)
Llama2-13B	0 ✓	Naive LLM	20.30	28.59	50.40	57.56
	4 ✓	Naive RAG	28.90	39.98	62.50	71.03
	4 ✓ + 1 ✗	Naive RAG	11.90	19.97	28.00	36.22
		Prompt Based	11.20	21.62	20.50	30.09
		Exclusion	23.70	34.00	54.40	62.37
		CrAM	**25.10** (+1.40)	**35.56** (+1.56)	**56.20** (+1.80)	**64.03** (+1.66)
Llama3-8B	0 ✓	Naive LLM	20.60	30.58	55.70	62.67
	4 ✓	Naive RAG	33.10	45.66	64.30	73.68
	4 ✓ + 1 ✗	Naive RAG	16.00	26.16	36.80	47.09
		Prompt Based	24.20	34.10	49.50	58.59
		Exclusion	26.60	38.44	57.70	67.33
		CrAM	**30.70** (+4.10)	**41.71** (+3.27)	**62.20** (+4.50)	**70.70** (+3.37)

Table 2: Main results under GPT setting. 0 ✓ indicates no document and the model directly prompted, 4 ✓ indicates all four documents retrieved from the Wikipedia dump, and 4 ✓ + 1 ✗ indicates four high-credibility documents (i.e., retrieved from external corpus) plus one low-credibility document (i.e., containing misinformation). In the 4 ✓ + 1 ✗ setting, the best performance is highlighted in **bold**. The red part indicates the improvement of our CrAM compared to the second-best model.

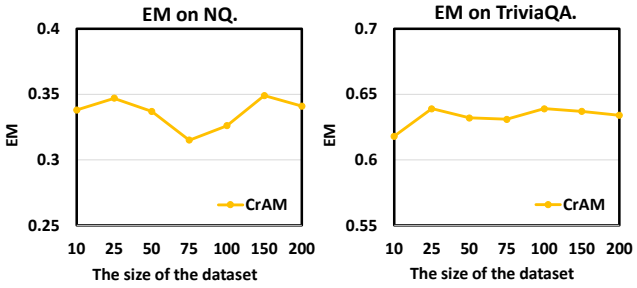

Figure 5: Performance on NQ and TriviaQA regarding the dataset size for determining the influential attention head changes.

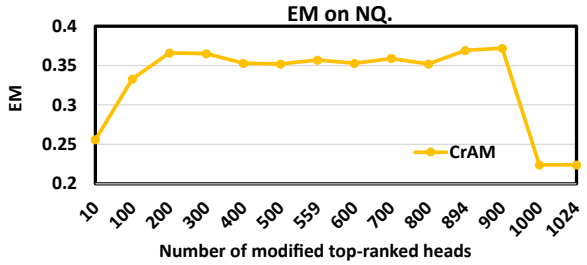

Figure 6: Performance on NQ in ideal setting regarding the varying number of selected attention heads.

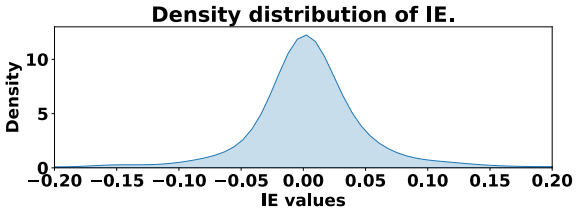

Figure 7: Density distribution of IE of all the attention heads in Llama3-8B.

rably, our CrAM model exhibits much smaller performance drops compared to other models when increasing the number of low-credibility documents. These results demonstrate the robustness of our proposed model to the varying number of low-credibility documents.

Effect of Dataset Size on Attention Heads Selection. As we described in Section 3.3, we randomly select 100 data points from each dataset to identify the influential attention heads. In the following, we vary the number of data points used for selecting these influential attention heads to analyze its impact on model performance. The experimental results are presented in Figure 5. Despite fluctuations in performance along with the changing dataset size, the variations are not substantial on both NQ and TriviaQA datasets, with a maximum difference of 4% in terms of EM. The results indicate that the number of data points has a minor impact on the final model performance.

Analysis on Number of Selected Attention Heads. In the following, we analyze the performance change when we adjust the number of selected attention heads. We present the results in Figure 6. We observe a sharp drop in model per-

Model	Method	NQ	TriviaQA
		EM	EM
Qwen1.5-7B	CrAM	29.10	52.90
	CrAM-all	27.20 (-1.90)	50.60 (-2.30)
	Naive RAG	10.50 (-18.60)	25.00 (-27.90)
Llama2-13B	CrAM	33.60	59.90
	CrAM-all	29.50 (-4.10)	59.50 (-0.40)
	Naive RAG	11.90 (-21.70)	28.00 (-27.90)
Llama3-8B	CrAM	36.90	64.40
	CrAM-all	22.40 (-14.50)	51.50 (-12.90)
	Naive RAG	16.00 (-20.90)	36.80 (-27.60)

Table 3: Results of ablation study under ideal setting with 4 ✓ + 1 ✗ (i.e., four high-credibility documents plus one low-credibility document).

formance when the number of selected attention heads is near either 0 or the maximum number of heads, i.e., 1024; comparably, it has a minor effect when the number of selected attention heads falls into the range of values in between. To investigate the underlying reasons, we further analyze the IE's density distribution using Llama3-8B, as shown in Figure 7. We find that the IE density distribution approximates a normal distribution centered around 0, with the majority of values concentrated near 0. It indicates that most attention heads have minor impact on model performance, and only when the attention heads with IE values far from zero, either positive or negative, are selected, the model performance will be affected significantly.

Ablation Study To better understand the rationality of our model design, we conduct ablation study and present the results in Table 3. First, we remove the selection of influential attention heads and apply attention weight modification on all attention heads in LLMs, and denote this variant model as CrAM-all. As shown in Table 3, we observe that the performance of the CrAM-all model has noticeable drops on all three LLMs. Among them, Llama3-8B based CrAM has the largest decrease on both NQ and TriviaQA, i.e., 14.5% and 12.9%. This indicates the necessity of identifying the influential attention heads before modifying the attention weights.

If we disable the attention weight modification mechanism in our model, it becomes the Naive RAG method. Table 3 shows that this results in a remarkable performance drop on all three LLMs compared to the CrAM model. For instance, the performance of all three LLMs decreases more than 27.5% on TriviaQA dataset. These results verify that it is necessary to modify the attention weight and meanwhile take into account the credibility scores of the documents.

5 Related Work

Misinformation Detection. Misinformation detection aims to identify false or misleading information from various data sources (Guo et al. 2019; Kaliyar and Singh 2019; Vaibhav, Mandyam, and Hovy 2019; Huang et al. 2024). It can be categorized into non-LLM-based methods and LLM-based methods. Non-LLM methods often involve training models to identify misinformation (Vaibhav, Mandyam, and Hovy 2019; Kaliyar, Goswami, and Narang 2021; Liu, Wang, and Li 2023; Goonathilake and Kumara 2020). For example, Kaliyar, Goswami, and Narang (2021) utilize BERT (Devlin et al. 2019) to score the credibility of documents, while Vaibhav, Mandyam, and Hovy (2019) use a graph neural network for misinformation detection. Comparably, LLM-based methods typically use LLMs without additional training (Pelrine et al. 2023; Quelle and Bovet 2024; Caramancion 2023; Hoes, Altay, and Bermeo 2023). For instance, Pelrine et al. (2023) adopt GPT-4 (OpenAI et al. 2024) for document credibility scoring, while Quelle and Bovet (2024) employ an LLM agent (Xi et al. 2023) for iterative verification of document credibility. In this study, we employ LLMs to obtain the credibility score for each document similar to the previous LLM-based methods (Pelrine et al. 2023; Hoes, Altay, and Bermeo 2023).

Combating Misinformation in RAG. Retrieval-Augmented Generation (RAG) enhance LLMs by retrieving relevant documents from external corpus (Lewis et al. 2020; Izacard and Grave 2021; Cai et al. 2024). However, prior works (Zou et al. 2024; Pan et al. 2023b,a) find that RAG is vulnerable to misinformation in its corpus, leading to undesired results. To combat misinformation in RAG, lots of studies have been conducted. For example, CAR (Weller et al. 2024) adopt a query augmentation scheme to retrieve a larger set of documents first and then apply a voting mechanism to mitigate the impact of misinformation. RobustRAG (Xiang et al. 2024) obtains the LLM response for each document independently and aggregates these responses through keyword-based and decoding-based algorithms to generate the final result. Hong et al. (2024) and Pan et al. (2024) assign each retrieved document a credibility score and fine-tune LLMs with the documents and their scores, enabling the LLMs to leverage these credibility scores when generating. CD^2 Jin et al. (2024) train two LLMs to generate truthful answers and misleading answers respectively to make it better distinguish the conflict information. However, CAR (Weller et al. 2024) and RobustRAG (Xiang et al. 2024) require multiple rounds of model inference, leading to inefficiency. The methods proposed by Hong et al. (2024), Pan et al. (2024), and Jin et al. (2024) require fine-tuning LLMs, which demands additional computational resources and well-designed training data, thereby limiting their application scenarios.

6 Conclusion

This work introduces CrAM, a plug-and-play method that enables RAG to automatically adjust the influence of retrieved documents on the output of LLMs based on document credibility. CrAM first identifies influential attention heads and then adjusts the attention weights of identified attention heads according to the credibility score of documents, regulating LLMs to pay less attention to the low-credibility documents. Empirical experiments demonstrate that, compared to vanilla RAG, CrAM improves EM performance by over 20% on two datasets and even outperforms the baseline with SFT, demonstrating CrAM's efficiency.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62272437).

References

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; et al. 2023. Qwen Technical Report. arXiv:2309.16609.

Cai, H.; Li, Y.; Wang, W.; Zhu, F.; Shen, X.; Li, W.; and Chua, T.-S. 2024. Large Language Models Empowered Personalized Web Agents. arXiv:2410.17236.

Caramancion, K. M. 2023. Harnessing the Power of ChatGPT to Decimate Mis/Disinformation: Using ChatGPT for Fake News Detection. In *2023 IEEE World AI IoT Congress (AIIoT)*, 0042–0046.

Chen, C.; and Shu, K. 2024. Can LLM-Generated Misinformation Be Detected? arXiv:2309.13788.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading Wikipedia to Answer Open-Domain Questions. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1870–1879. Vancouver, Canada: Association for Computational Linguistics.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In Linzen, T.; Chrupała, G.; Belinkov, Y.; and Hupkes, D., eds., *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286. Florence, Italy: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dufour, N.; Pathak, A.; Samangouei, P.; Hariri, N.; Deshetti, S.; et al. 2024. AMMeBa: A Large-Scale Survey and Dataset of Media-Based Misinformation In-The-Wild. arXiv:2405.11697.

Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1: 1.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; and Wang, H. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Goonathilake, M. D. P. P.; and Kumara, P. P. N. V. 2020. CNN, RNN-LSTM Based Hybrid Approach to Detect State-of-the-Art Stance-Based Fake News on Social Media. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 23–28.

Guo, B.; Ding, Y.; Yao, L.; Liang, Y.; and Yu, Z. 2019. The Future of Misinformation Detection: New Perspectives and Trends. arXiv:1909.03654.

Hoes, E.; Altay, S.; and Bermeo, J. 2023. Leveraging ChatGPT for Efficient Fact-Checking.

Hong, G.; Kim, J.; Kang, J.; Myaeng, S.-H.; and Whang, J. J. 2024. Why So Gullible? Enhancing the Robustness of Retrieval-Augmented Models against Counterfactual Noise. arXiv:2305.01579.

Huang, Y.; Zhu, F.; Tang, J.; Zhou, P.; Lei, W.; Lv, J.; and Chua, T.-S. 2024. Effective and Efficient Adversarial Detection for Vision-Language Models via A Single Vector. arXiv:2410.22888.

Izacard, G.; and Grave, E. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 874–880. Online: Association for Computational Linguistics.

Jin, Z.; Cao, P.; Chen, Y.; Liu, K.; Jiang, X.; Xu, J.; Qiuxia, L.; and Zhao, J. 2024. Tug-of-War between Knowledge: Exploring and Resolving Knowledge Conflicts in Retrieval-Augmented Language Models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 16867–16878. Torino, Italia: ELRA and ICCL.

Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611. Vancouver, Canada: Association for Computational Linguistics.

Kaliyar, R. K.; Goswami, A.; and Narang, P. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80(8): 11765–11788.

Kaliyar, R. K.; and Singh, N. 2019. Misinformation Detection on Online Social Media-A Survey. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. Online: Association for Computational Linguistics.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and

Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Li, M.; Wang, W.; Feng, F.; Zhu, F.; Wang, Q.; and Chua, T.-S. 2024. Think Twice Before Trusting: Self-Detection for Large Language Models through Comprehensive Answer Reflection. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 11858–11875. Association for Computational Linguistics.

Liu, H.; Wang, W.; and Li, H. 2023. Interpretable Multimodal Misinformation Detection with Logic Reasoning. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 9781–9796. Toronto, Canada: Association for Computational Linguistics.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and Editing Factual Associations in GPT. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 17359–17372. Curran Associates, Inc.

Meta. 2024. LlaMA 3.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pan, L.; Chen, W.; Kan, M.-Y.; and Wang, W. Y. 2023a. Attacking Open-domain Question Answering by Injecting Misinformation. In Park, J. C.; Arase, Y.; Hu, B.; Lu, W.; Wijaya, D.; Purwarianti, A.; and Krisnadhi, A. A., eds., *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 525–539. Nusa Dua, Bali: Association for Computational Linguistics.

Pan, R.; Cao, B.; Lin, H.; Han, X.; Zheng, J.; Wang, S.; Cai, X.; and Sun, L. 2024. Not All Contexts Are Equal: Teaching LLMs Credibility-aware Generation. arXiv:2404.06809.

Pan, Y.; Pan, L.; Chen, W.; Nakov, P.; Kan, M.-Y.; and Wang, W. 2023b. On the Risk of Misinformation Pollution with Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1389–1403. Singapore: Association for Computational Linguistics.

Pelrine, K.; Imouza, A.; Thibault, C.; Reksoprodjo, M.; Gupta, C.; Christoph, J.; Godbout, J.-F.; and Rabbany, R. 2023. Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6399–6429. Singapore: Association for Computational Linguistics.

Quelle, D.; and Bovet, A. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Austin, Texas: Association for Computational Linguistics.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Vaibhav, V.; Mandyam, R.; and Hovy, E. 2019. Do Sentence Interactions Matter? Leveraging Sentence Level Representations for Fake News Classification. In Ustalov, D.; Somasundaran, S.; Jansen, P.; Glavaš, G.; Riedl, M.; Surdeanu, M.; and Vazirgiannis, M., eds., *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, 134–139. Hong Kong: Association for Computational Linguistics.

Vincent, J. 2023. Google and Microsoft's chatbots are already citing one another's misinformation. *The Verge*. Accessed: 2023-06-05.

Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; and Titov, I. 2019. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5797–5808. Florence, Italy: Association for Computational Linguistics.

Weller, O.; Khan, A.; Weir, N.; Lawrie, D.; and Van Durme, B. 2024. Defending Against Disinformation Attacks in Open-Domain Question Answering. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 402–417. St. Julian's, Malta: Association for Computational Linguistics.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; et al. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. arXiv:2309.07864.

Xiang, C.; Wu, T.; Zhong, Z.; Wagner, D.; Chen, D.; and Mittal, P. 2024. Certifiably Robust RAG against Retrieval Corruption. arXiv:2405.15556.

Yoran, O.; Wolfson, T.; Ram, O.; and Berant, J. 2024. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. arXiv:2310.01558.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; et al. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219.

Zhu, F.; Lei, W.; Wang, C.; Zheng, J.; Poria, S.; and Chua, T.-S. 2021. Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering. arXiv:2101.00774.

Zou, W.; Geng, R.; Wang, B.; and Jia, J. 2024. PoisonedRAG: Knowledge Poisoning Attacks to Retrieval-Augmented Generation of Large Language Models. arXiv:2402.07867.