Not All Contexts Are Equal: Teaching LLMs Credibility-aware Generation

Ruotong Pan[1,2], **Boxi Cao**[1,2,*], **Hongyu Lin**[1], **Xianpei Han**[1],
Jia Zheng[1,*], **Sirui Wang**[3], **Xunliang Cai**[3], **Le Sun**[1]
[1]Chinese Information Processing Laboratory, Institute of Software,
Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Meituan
{panruotong2021, boxi2020, hongyu, xianpei, zhengjia}@iscas.ac.cn
{sunle}@iscas.ac.cn {wangsirui,caixunliang}@meituan.com

Abstract

The rapid development of large language models has led to the widespread adoption of Retrieval-Augmented Generation (RAG), which integrates external knowledge to alleviate knowledge bottlenecks and mitigate hallucinations. However, the existing RAG paradigm inevitably suffers from the impact of *flawed information* introduced during the retrieval phrase, thereby diminishing the reliability and correctness of the generated outcomes. In this paper, we propose Credibility-aware Generation (CAG), a universally applicable framework designed to mitigate the impact of flawed information in RAG. At its core, CAG aims to equip models with the ability to discern and process information based on its credibility. To this end, we propose an innovative data transformation framework that generates data based on credibility, thereby effectively endowing models with the capability of CAG. Furthermore, to accurately evaluate the models' capabilities of CAG, we construct a comprehensive benchmark covering three critical real-world scenarios. Experimental results demonstrate that our model can effectively understand and employ credibility for generation, significantly outperform other models with retrieval augmentation, and exhibit robustness despite the increasing noise in the context. [1]

1 Introduction

In recent years, Large Language Models (LLMs) (Brown et al., 2020; OpenAI et al., 2023; Touvron et al., 2023; Anil et al., 2023) have experienced significant growth and demonstrated excellent performance in multiple domains (Kojima et al., 2022; Thirunavukarasu et al., 2023; Ziems et al., 2023; Min et al., 2023). With the ascendancy of LLMs, Retrieval-Augmented Generation (RAG) has attracted significant interest. RAG mitigates the knowledge bottleneck of LLMs by incorporating externally retrieved documents into their generation process. This inclusion helps diminish the occurrences of hallucinations and misinformation during generation, thereby substantially enhancing the quality of output from LLMs (Petroni et al., 2021; Zhu et al., 2021; Mallen et al., 2023).

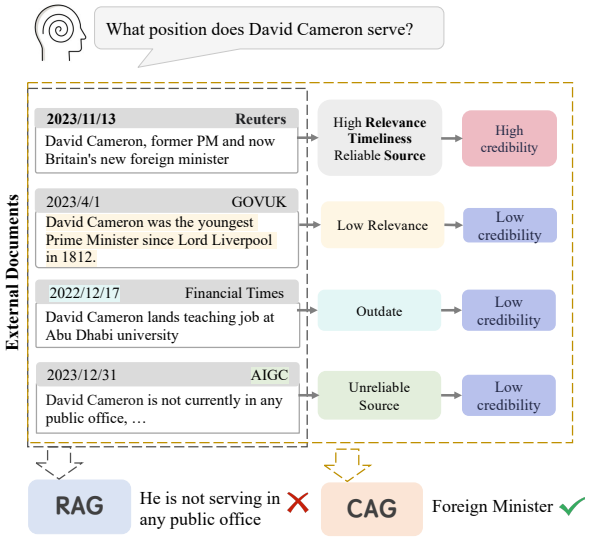


Figure 1: The comparison between Retrieval-Augmented Generation (RAG) and Credibility-aware Generation (CAG). Incorporating credibility into the model aids in mitigating errors caused by *flawed information* introduced from the retrieval process.

However, RAG for large language models remains significantly impacted by flawed information. This is mainly because the retrieval process often provides noisy, outdated, and incorrect contexts which adversely affects RAG, substantially reducing its effectiveness. Specifically, previous research (Shi et al., 2023a; Chen et al., 2023) has found that LLMs are highly sensitive to noise, which impacts LLMs' capacity to discern and trust accurate information, ultimately affecting the outcomes they generate. Furthermore, due to the temporal insensitivity of LLMs (Su et al., 2022; Zhao et al., 2024), these models struggle to discern out-

*Corresponding authors.
[1]Our code, benchmark, and models are available at https://github.com/panruotong/CAG

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19844–19863
November 12-16, 2024

dated information solely based on their internal knowledge. More critically, because LLMs are trained on extensive collections of historical text, there's an inherent risk that outdated information will align with the models' internal knowledge bases. This alignment can encourage LLMs to favor and perpetuate outdated information. Besides, the prevalence of misinformation on the current web poses a significant challenge for large models, which struggle to identify misinformation using only their inherent knowledge (Xie et al., 2023; Pan et al., 2023). This difficulty makes them susceptible to misinformation, leading to the generation of incorrect answers. Therefore, flawed information, characterized by noisy, outdated, and incorrect information, has substantial negative effects on RAG.

From a cognition perspective, a common approach humans adopt to combat flawed information is to assess the credibility of external information (Burgoon et al., 2000). For humans, credibility refers to the acceptability based on the quality of the information, its source, and subjective evaluation. However, LLMs relying solely on internal knowledge to assess information credibility are unstable and unreliable (Xie et al., 2023). Therefore, we aim to guide LLMs' acceptance of information by utilizing external indicators of credibility. We introduce **Credibility-aware Generation** (**CAG**), a universally applicable framework designed to address flawed information encountered during RAG. At its core, CAG seeks to equip models with the ability to discern and process information based on credibility. By assigning different credibility to information based on its relevance, timeliness, and the reliability of its source, and explicitly distinguishing them in the input, CAG significantly mitigates the issues arising from flawed information.

Unfortunately, we have discovered that existing LLMs are not inherently sensitive to directly provided credibility in the prompt. This deficiency restricts their capacity to optimally employ credibility for discerning and processing information. To endow models with the capability of CAG, we propose a novel data transformation framework. This framework transforms existing Question Answering (QA) datasets into data that integrates credibility, which can be employed to guide the model for credibility-based generation. Specifically, our process comprises two core steps: 1) Multi-granularity credibility annotation, which assigns credibility to text units at both document and sentence levels by dividing retrieved documents into varying granularities. 2) Credibility-guided explanation generation, which prompts LLMs to generate credibility-guided explanations given questions, retrieved documents with credibility annotation and *golden answers*. Finally, we employ instruction fine-tuning to train the model, enabling it to generate responses based on credibility.

To rigorously assess the ability of the model's credibility-aware generation in managing flawed information, we construct a comprehensive benchmark encompassing various real-world scenarios, including open-domain QA, time-sensitive QA, and misinformation polluted QA. In this benchmark, retrieval relevance, timeliness, and source authority are regarded as established measures of credibility. Experimental results on multiple datasets across multiple scenarios demonstrate the efficacy of our approach in utilizing credibility. Our model significantly outperforms various prevalent RAG approaches applied to both open and closed-source LLMs of diverse scales. Additionally, it exhibits robust resilience against noisy documents, maintaining high performance even as alternative strategies suffer sharp declines. All these results verify the effectiveness of the proposed CAG framework and corresponding training algorithm.

The main contributions of this study are summarized as follows [2]:

- We present Credibility-aware Generation, a universal framework to handle the flawed information challenge in RAG.
- We propose a novel data transformation framework that transforms existing datasets into data annotated with credibility and guides models to generate responses based on credibility, thereby equipping the model with Credibility-aware Generation capability.
- We construct a comprehensive benchmark and evaluate model performance in credibility-aware generation, encompassing real-world scenarios of open-domain QA, time-sensitive QA, and misinformation polluted QA.
- Experimental evidences demonstrate that our model effectively understands and employs credibility to generate responses, significantly surpasses other RAG-based strategies, and

[2]We uploaded the code and datasets as supplemental materials, which will be openly released after accepting.

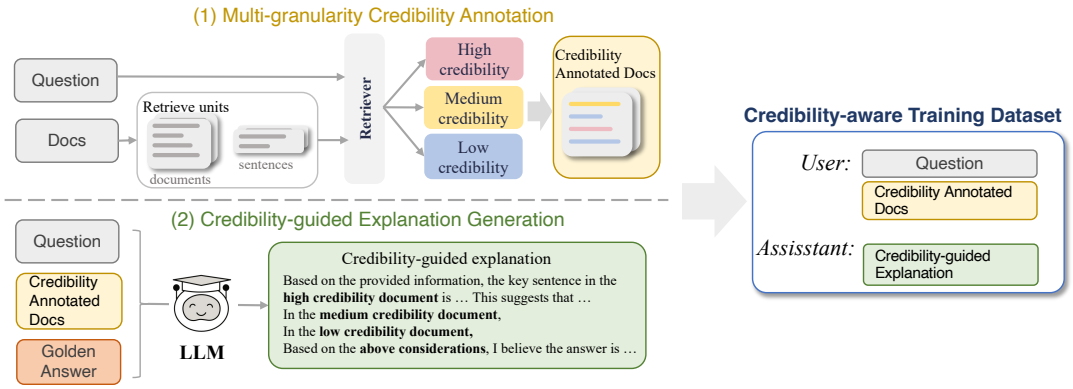


Figure 2: Overview of data transformation framework. The training data is constructed by assigning credibility to contexts via multi-granularity credibility annotation (§3.1) and prompting LLM to produce credibility-guided explanations (§3.2). The processed data is used to instruction fine-tuning (§3.3) to endow the model with the ability for Credibility-aware Generation.

maintains robustness despite the increasing noise in the context.

2 Credibility-aware Generation

Credibility-aware Generation is designed to enable models to discern and process information based on its credibility. Subsequently, we will provide formal definitions for both RAG and CAG, illustrating their divergence.

Definition In the Retrieval-Augmented Generation process, user input $\boldsymbol{x}$ initiates the retrieval of a set of related documents $D_{\boldsymbol{x}}$ from a large corpus C based on how closely these documents match the input. Then, it combines the input $\boldsymbol{x}$ with these documents $D_{\boldsymbol{x}}$ to generate responses $\boldsymbol{y}$, formalized as $\boldsymbol{y} = \text{LM}([\boldsymbol{x}, D_{\boldsymbol{x}}])$, where $[.,.]$ denotes the concatenation operation.

Compared to RAG, the Credibility-aware Generation offers additional credibility for each document. Initially, through credibility assessment based on various scenarios, each retrieved document has been assigned a level of credibility. Then, these documents $D_{\boldsymbol{x}}$ with their credibility C are synthesized with the user input $\boldsymbol{x}$ as augmented input. LM generates responses $\boldsymbol{y}$ based on this augmented input, formally represented as $\boldsymbol{y} = \text{LM}\left(\left[\boldsymbol{x}, \{[c_i, d_i]\}_{i=1}^{|D_{\boldsymbol{x}}|}\right]\right)$. This approach ensures that the generated responses not only incorporate the content of the documents but also consider the credibility of each document, thereby enhancing the reliability of responses.

3 Teaching Model to Credibility-aware Generation

In this section, we endow LLMs with the capability of CAG. A potential approach involves directly providing the credibility annotations of each document in the prompt. Unfortunately, as indicated in Table 2, our experiments reveal that even advanced LLMs, such as ChatGPT, exhibit limited sensitivity to credibility. To this end, we introduce a novel data transformation framework. Through multi-granularity credibility annotation and credibility-guided explanation generation, we transform existing QA datasets into data that includes credibility annotations which can guide the model to generate credibility-based responses. Then, through instruction fine-tuning, we train the model to generate responses grounded in credibility assessments.

3.1 Multi-granularity Credibility Annotation

To cater to the varied requirements for credibility across different scenarios and enhance the model's comprehension of credibility, we collect training data including open-domain QA, machine reading comprehension, and dialogue datasets and propose a multi-granularity credibility annotation method.

First, we divide the retrieved documents to create a multi-granularity corpus, encompassing sentence and document levels. Then, the retriever assesses the match between each retrieval unit and the query, assigning a relevance score, and classifies documents into three levels: high, medium, and low, using either equal count or equal interval methods. This approach of using levels instead of scores aims to simplify representation, thereby improving the model's understanding and providing a certain de-

gree of fault tolerance. Ultimately, we collect about 15k training data samples, all of which include documents with credibility annotations. The detailed composition of the training data is shown in the Appendix A.1.

3.2 Credibility-guided Explanation Generation

To facilitate the model's comprehension and effective utilization of credibility, we employ LLMs to generate credibility-guided explanations for the answers.

Given the limitations of current LLMs in comprehending credibility effectively, we design chain-of-thought prompts to guide LLMs to generate credibility-guided explanations given questions, retrieved documents with credibility and **golden answers**. In this case, LLMs only need to generate coherent explanations based on the document containing the answers, without distinguishing between documents with different credibility to generate answers. The credibility-guided explanation obtained includes an analysis that integrates both the credibility and the content of the documents, rather than merely focusing on deriving the answer.

Considering the accessibility and advanced capabilities, we employ GPT-3.5 for the generation of explanations. In this way, we obtain high-quality answer explanations. Then, we replace the original answers in the training data with credibility-guided explanations to form a novel QA dataset. In this dataset, the inputs include questions and external documents annotated with credibility, while the outputs are credibility-guided explanations.

3.3 Instruction Fine-tuning

Through the two steps above, the training dataset obtained contains credibility, which can be used to facilitate arbitrary language models in gaining the capacity for CAG. We fine-tune the language model on this dataset to empower the model to discern and process information according to its credibility. As defined by Iyer et al. (2023), the loss function is as follows:

$$\mathcal{L}(D_{\boldsymbol{x}};\theta) = -\sum_{i=1}^{N} \log p_\theta \left(\boldsymbol{y}_i \mid \left[\boldsymbol{x}, \{[c_i, d_i]\}_{i=1}^{|D_{\boldsymbol{x}}|} \right], \boldsymbol{y}_{<i} \right)$$

4 Credibility-aware Generation Benchmark

To rigorously evaluate the ability of credibility-aware model generation to handle flawed information, we construct the Credibility-aware Generation Benchmark (CAGB). This benchmark encompasses the following three specific scenarios where the integration of credibility is essential:

• **Open-domain QA** aims to accurately answer questions on a wide variety of topics without being limited to any particular area. It encompasses a broad spectrum of real-world applications that urgently require the integration of external knowledge to enhance the LLMs' ability to address queries. This scenario necessitates the ability to effectively identify and process noise information.

• **Time-sensitive QA** aims to give accurate and current answers. It poses a challenge for LLMs due to the dynamic internet information. The inevitable inclusion of outdated documents when incorporating external sources further complicates matters. Even with timestamps provided for documents, LLMs may erroneously prioritize outdated documents. This situation underscores the critical need for credibility in time-sensitive QA.

• **Misinformation polluted QA** aims to tackle the issue of ensuring accurate answers in an environment polluted with misinformation. It presents a substantial challenge to LLMs, attributed to the misuse of LLMs and the consequent proliferation of fake news and misinformation (Zhuo et al., 2023; Pan et al., 2023). Consequently, it is crucial to take into account the quality and credibility of any introduced external information. In the following, we will provide a detailed description of data construction for each scenario, and the statistics of CAGB are shown in the Table 1.

4.1 Credibility Assessment

We aim to establish a flexible credibility assessment mechanism that can be conveniently extended to consider additional factors and a broader range of application fields. In this benchmark, the credibility of the documents is evaluated by considering retrieval relevance, timeliness, and source reliability. Specifically, we establish a foundation based on retrieval relevance, then make adjustments according to timeliness, and finally integrate the reliability of the source to determine credibility. First, the retriever assigns relevance scores to documents based on query similarity. These relevance scores, which are distributed at equal intervals, enable to classify documents into three levels: high, medium, and low, collectively denoted as R. Subsequently, the temporal difference T between the query time and

Dataset	#Samples	#Documents	Noise Ratio
Open-domain QA			
HotpotQA	500	5000	0.8
2WikiMHQA	500	5000	0.6-0.8
MuSiQue	500	10000	0.9
ASQA	948	4740	-
RGB	300	11641	0.2-0.8
Time-sensitive QA			
EvolvTempQA	321	2247	0.4-0.8
Misinformation polluted QA			
NewsPollutedQA	480	2400	0.5-0.75

Table 1: Statistics of CAGB, which includes 7 dataset derived from 3 scenarios.

document publication is calculated, downgrading R if T surpasses a threshold. The formula integrating relevance and timeliness is as follows:

$$rt\_score(R, T) = \max(R - \text{floor}(T/\text{threshold}), 1)$$

Following this, the source reliability, denoted S, is customized to specific scenarios, similarly divided into three levels. Finally, we combine these factors, adopting the lower level as the credibility and the formula is expressed as follows:

$$Cred = \min(rt\_score(R, T), S)$$

In this way, the document of high credibility are concurrently characterized by high relevance, timeliness and source reliability. More details about the assessment can be seen in the Appendix A.9.

4.2 Open-domain QA

Our research utilizes data from several challenging QA datasets with noisy documents. HotpotQA (Yang et al., 2018) and 2WikiMHQA (Ho et al., 2020) both require reasoning across multiple documents, and feature a high proportion of distracting documents. Importantly, the data we utilize from HotpotQA is extracted from the dev subset, whereas our training dataset is derived from the train subset. Musique (Trivedi et al., 2021) questions are of higher complexity, with up to 90% of distracting passages. ASQA (Stelmakh et al., 2022) is a long format QA dataset focused on ambiguous questions. RGB (Chen et al., 2023) is a specialized benchmark used for evaluating the capabilities of models in the RAG scenario, with noise robustness being one of its aspects. We assign credibility to the documents provided in the dataset in terms of retrieval relevance.

4.3 Time-sensitive QA

In order to construct a diverse, high-quality, and up-to-date news dataset, we annotate 321 time-sensitive questions along with their corresponding dates. These questions originate from real-world scenarios, including news QA data from RealTime QA (Kasai et al., 2022), TAQA (Zhao et al., 2024), and questions adapted from news reports. To simulate the simultaneous occurrence of varied information on the Internet, we use Google search API to retrieve each query, selecting 3 relevant documents and 4 distracting documents. The distracting documents are either irrelevant to the query or outdated. This approach to document selection is crafted to emulate the intricate and heterogeneous nature of real-world information landscapes. Each news includes its publication date, thereby aiding in the evaluation of its timeliness. For document credibility annotation, we assess credibility based on relevance and time gap between the document's publication and the posed question. We ensure the accuracy of the answers by manually annotating.

The obtained time-sensitive dataset with outdated document settings and credibility annotation is named EvolvingTempQA.

4.4 Misinformation Polluted QA

We create a up-to-date multiple-choice quiz dataset, comprising both real and fake news for each question. The dataset construction bases on RealTime QA, utilizing weekly news quizzes from CNN and other news platforms. To maintain the dataset's real-time relevance, we select news from July 1, 2023, onwards, comprising 480 questions with four options and one supporting news item each.

To simulate the generation of fake news, we first generated a claim using LLMs, based on a question and a randomly selected incorrect option. This process transforms the question and incorrect option into a deceptive statement. Subsequently, we choose GPT-3.5 and Qwen (Bai et al., 2023) as the generators for fake news, guiding them to generate texts of varying styles based on the claim, including news style and Twitter style. The prompts used and examples are detailed in the Appendix A.15. The fictitious news articles produced by LLMs, due to their authenticity being deliberately compromised, are classified as having low source reliability. Conversely, news articles from reputable news websites are considered to possess high source reliability. We set the ratio of fake news at 0.5, 0.67, and 0.75 to evaluate the robustness of model against misinformation under various levels of pollution.

By simulating fake news generation, we create a misinformation polluted QA dataset in the news

Model	Open-domain QA					Time-sensitive QA	Misinfo polluted QA
	HotpotQA	2WikiMHQA	MuSiQue	ASQA	RGB	EvolvingTempQA	NewsPollutedQA
retrieval-based							
ChatGPT	0.334	0.368	0.194	0.404	0.773	0.579	0.231
LLaMA-2-7B	0.280	0.312	0.160	0.268	0.753	0.433	0.179
Vicuna-7B	0.278	0.296	0.116	0.358	0.677	0.567	0.229
Mistral-7B-Instruct	0.288	0.270	0.106	0.300	0.713	0.598	0.204
LLaMA-2-13B	0.366	0.370	0.164	0.321	0.820	0.495	0.204
LLaMA-2-70B	0.418	0.390	0.256	0.316	0.823	0.526	0.430
vanilla IFT	0.324	0.245	0.270	0.157	0.650	0.592	0.329
retrieval and reranking							
ChatGPT	0.396	0.394	0.216	0.388	0.790	0.632	0.427
LLaMA-2-7B	0.302	0.376	0.200	0.375	0.730	0.526	0.265
Vicuna-7B	0.355	0.306	0.164	0.494	0.757	0.620	0.275
Mistral-7B-Instruct	0.338	0.334	0.166	0.414	0.790	0.741	0.373
LLaMA-2-13B	0.370	0.372	0.180	0.390	0.823	0.561	0.308
LLaMA-2-70B	0.422	0.504	0.320	0.388	0.833	0.570	0.306
vanilla IFT	0.348	0.448	0.276	0.304	0.663	0.720	0.344
retrieval and credibility							
ChatGPT	0.422	0.402	0.182	0.440	0.807	0.673	0.408
LLaMA-2-7B	0.376	0.176	0.140	0.394	0.713	0.486	0.213
Vicuna-7B	0.349	0.266	0.091	0.490	0.740	0.642	0.279
Mistral-7B-Instruct	0.274	0.268	0.102	0.463	0.797	0.679	0.315
LLaMA-2-13B	0.360	0.384	0.164	0.385	0.803	0.520	0.227
LLaMA-2-70B	0.398	0.402	0.262	0.492	0.817	0.536	0.279
vanilla IFT	0.372	0.334	0.204	0.305	0.663	0.589	0.383
CAG-7B (*ours*)	0.509	0.578	0.340	0.496	0.897	0.826	0.442
CAG-13B (*ours*)	**0.514**	**0.604**	**0.408**	**0.525**	**0.917**	0.829	0.483
CAG-mistral-7B (*ours*)	0.502	0.540	0.384	0.505	0.900	**0.835**	**0.613**

Table 2: Model performance in our CAGB benchmark. The best/second best scores in each dataset are **bolded**/underlined. Our models substantially outperform previous strategies across all 3 scenarios in CAGB. The results shown for EvolvingTempQA and RGB are at noise_ratio setting of 0.8, while NewsPollutedQA is at noise_ratio setting of 0.75. The results of other metrics on the ASQA dataset are shown in the Appendix A.6.

domain, named NewsPollutedQA.

5 Experiments

To demonstrate the effectiveness of our framework in handling flawed information in real-world QA scenarios, we conduct comprehensive experiments within the CAGB. All these results verify the effectiveness of the CAG framework and the corresponding training algorithm. Additionally, our models maintain robustness even with the increasing noise in the context. In the following sections, we will discuss our experiments and conclusions in detail.

5.1 Setup

Baselines We compare our method with the following three strategies incorporated with 7 LLMs across various scales:

- **Retrieval-based** concatenates documents from the dataset with questions as input.
- **Retrieval and reranking** employs an advanced reranking mechanism to reorder retrieved documents, giving priority to those with greater relevance (Xie et al., 2023).
- **Retrieval and credibility** incorporates credibility as a prefix to the retrieved documents in the prompt, aiming to assess the model's ability to understand and utilize credibility.

We evaluate advanced models, including ChatGPT `(gpt-3.5-turbo-0613)`, LLaMA-2-7B, 13B, 70B, Vicuna-7B-v1.5 and Mistral-7B-Instruct (Jiang et al., 2023). Additionally, we create a dataset mirroring the model training data but without credibility annotations and with initial answers, on which we fine-tune the LLaMA-2-7B model, and named the trained model vanilla IFT.

Experimental settings We use LLaMA-2-7B, 13B and Mistral-7B as our base models. We train the LLaMA-2 model and the Mistral 7B model for 3 epochs with a learning rate of 1e-5 on A100-80G GPUs.To provide relevance scores, we use SPLADE (Formal et al., 2021) as our retriever. For all language models, we include 3-shot QA exam-

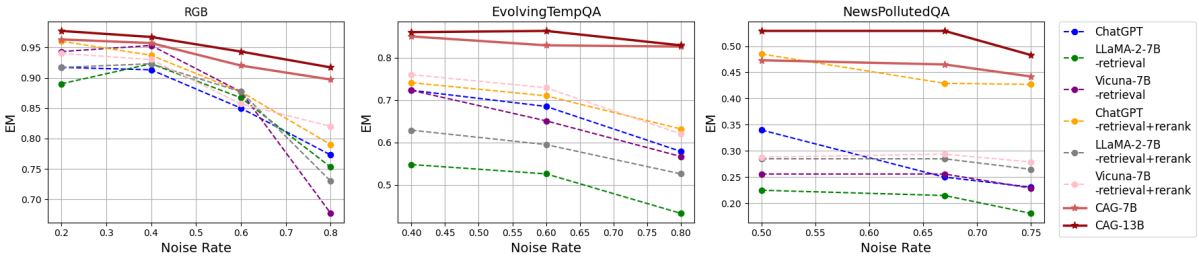


Figure 3: The performance of LLMs under varying noise ratios, which denote the proportions of retrieved noise documents. As the noise ratio increases, the performance of other methods markedly declines; in contrast, our model maintains stable performance in high noise ratio, attributed to its enhanced ability to prioritize accurate information.

ples within the prompt. We utilize Exact Match (EM) (Stelmakh et al., 2022) as the primary evaluation metric for all datasets. It is calculated by checking whether the short answers provided are exact substrings of the generation. We set the temperature to 0.01 during the inference. Additional experimental settings and the prompts used for evaluation and are provided in the Appendix A.11. These models can generate answers with reasoning processes based on the given prompt. Moreover, Appendix A.7 shows the results of the experiments using CoT prompt consistent with that used to generate training data.

5.2 Overall Results

The main results of the three scenarios are shown in the Table 2, we can clearly see that our model efficiently utilizes credibility to provide more accurate and credible responses. In the following, we analyze the experimental results in detail:

1) Previous approaches based on RAG severely suffer from the flawed information introduced during retrieval. In scenarios including open-domain QA, time-sensitive QA, and misinformation pollutedQA, existing LLMs, including ChatGPT and LLaMA-2-70B, face challenges due to interference from flawed information. In the retrieval-based open-domain QA, the average EM score for ChatGPT is only 41.5%, while 44.1% for LLaMA-2-70B. All models exhibit low performance on the Musique, NewsPollutedQA, which are characterized by high ratios of flawed information. Reranking with external relevance scores can assist the model to a certain extent, as the model is sensitive to the order of documents (Xie et al., 2023).

2) CAG significantly improves performance by discerning between documents and guiding the model to prioritize those with high credibility. Our models significantly surpass all baseline models across the 7 datasets under 3 scenarios, including ChatGPT and LLaMA-2-70B enhanced with retrieval and reranking. For instance, on the 2WikiMHQA dataset, our CAG-7B improves 26.6% of EM score over the LLaMA-2-7B model and 28.2% of EM score over the Vicuna-7B model under retrieval-based.

3) Our approach generalizes to scenarios previously unseen which require credibility and demonstrates compatibility with diverse base models. The models, developed through training on LLaMA 7B, 13B, and Mistral 7B with CAG, not only exhibit improved reliability in its outputs but also excel in new, challenging situations, including time-sensitive QA and misinformation polluted QA. This performance, achieved within an open-domain QA framework lacking temporal or source integration, effectively manages diverse flawed information and affirms the universality of CAG.

5.3 Analysis Study

In the following, we will present analysis against the robustness and limitation of current CAG model. Due to the space limit, experimental results on the effect of credibility annotation accuracy are shown in the Appendix A.2.

5.3.1 Noise Robustness Analysis

Previous research has demonstrated that an increase in the proportion of noise within the context significantly degrades model performance (Xie et al., 2023; Chen et al., 2023). To assess the robustness of diverse methods against flawed information, we vary the ratio of noisy documents across three distinct datasets: RGB, EvolvingTempQA and NewsPollutedQA, and observe the consistency in performance changes across different models.

We present the results in Figure 3 and can see that: **Credibility-aware Generation makes the model robust to flawed information, which en-**

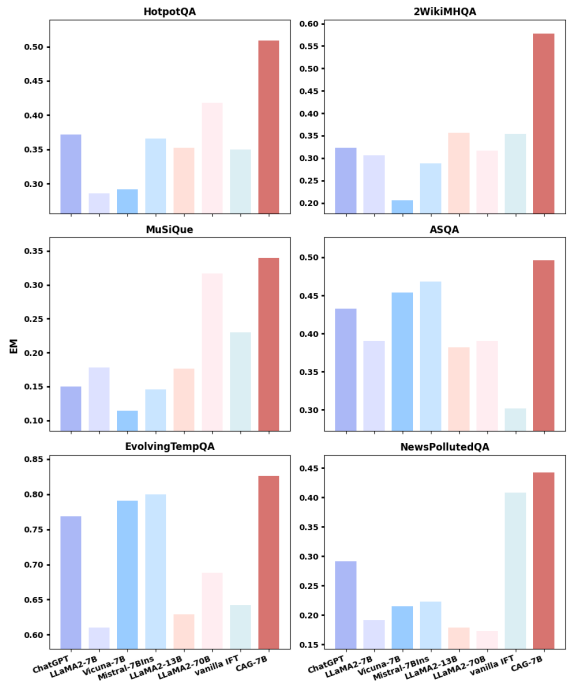


Figure 4: The comparison of performance of LLMs under discarding low credibility document setting and CAG-7B across six datasets.

hances its ability to discern and prioritize accurate information. As the proportions of noise in the context increases, most of the models exhibit performance degradation aligning with the observations made by Chen et al. (2023). However, our models show greater robustness compared to others, notably the improved performance of CAG-13B on EvolvingTempQA. The results of the noise robustness analysis for all LLMs are shown in the Appendix A.13.

5.3.2 Analysis of Discarding Low Credibility Documents

Upon assigning credibility to the documents in context, an alternative intuitive strategy is to simply discard low credibility documents. However, considering that credibility assessments are not precise, this strategy may inadvertently filter out helpful information, thereby impairing the accuracy of the model's responses. To demonstrate this, we compare the performance of LLMs in this setting with that of CAG-7B in our CAGB. The results are shown in Figure 4, and full results are in Appendix A.8. We can clearly see that: by preserving more document information and differentiating them based on explicit credibility in the prompt, our CAG framework mitigates the risk of losing valuable information. As a result, the accuracy and comprehensiveness of the responses are improved.

6 Related Work

Retrieval-Augmented Generation (Lewis et al., 2020) integrates a retriever with a generator to improve text generation quality by utilizing external knowledge (Izacard and Grave, 2021; Borgeaud et al., 2022; Shi et al., 2023b). However, the accuracy of RAG is compromised by flawed information, as the inclusion of noisy (Chen et al., 2023), outdated (Kasai et al., 2022; Wang et al., 2023a), or false information (Chen and Shu, 2023; Pan et al., 2023) during the retrieval negatively impacts the generator's outputs.

Previous studies have primarily focused on filtering, ranking, or manually evaluating retrieved documents to mitigate the impact of flawed information. For instance, Peng et al. (2023); Wang et al. (2023b) deploy various filtering algorithms to remove irrelevant text. Zhang and Choi (2023) utilizes timestamps to identify and discard outdated information. However, these approaches are limited by the accuracy of filtering algorithms, thereby discarding helpful information and impairing the effectiveness of RAG. Meanwhile, misinformation is primarily addressed by identifying falsehoods through fact-checking (Vijjali et al., 2020). However, this approach necessitates either human verification or further training of the discriminator (Baek et al., 2023), both of which can be resource-intensive and introduce bias (Draws et al., 2022; Su et al., 2023). In comparison, our work mitigates the impact of flawed information without discarding documents by introducing multi-feature dimensions of external information to assess the credibility level of each document.

Researchers fine-tune language models to better leverage the context provided in the input. For instance, Li et al. (2023) train the model using counterfactuals and irrelevant context to prioritize context. Yoran et al. (2023) include irrelevant context in the training samples, making the model robust to irrelevant documents. Asai et al. (2023) train the model on contexts with reflective tokens, enabling it to evaluate the relevance of passages during generation. However, these approaches focus mainly on irrelevant documents. Meanwhile, the model predominantly learns implicit rules, resulting in opaqueness of the generation.

7 Conclusions

This paper proposes Credibility-aware Generation to address the challenge of flawed information. To

equip the model with CAG capabilities, we introduce a data transformation framework aimed at generating credibility-based dataset, upon which we fine-tune the model. To effectively verify the ability of model Credibility-aware Generation to handle flawed information, we construct a benchmark from different real-world scenarios. Experimental results show that our model can effectively utilize credibility, exhibiting robustness in the face of flawed information and significantly outperforming other models with retrieval augmentation.

Moreover, through customizing the credibility, our approach can be applied to the real-world scenario including personalized response generation, for which we provide a detailed case study in the Appendix A.5.

Limitations

There are several limitations of our current CAG framework, which we plan to address in the future. Firstly, we have established a flexible credibility assessment mechanism, focusing more on endowing the model with the ability to generate based on credibility. However, credibility assessment is also a crucial part, and the current performance gap exists due to the retrieval strategy and influencing factors. In future research, we will delve further into credibility assessment to enhance the performance of our model. Secondly, despite our method demonstrating strong generalization capabilities, it still relies on additional training data annotation and training. In the future, we will explore how to enable existing models to perform confidence-aware generation without the need for further training. Thirdly, our methodology, effectively applied to RAG, acknowledges the broader research domain encompassing external resources like knowledge graphs and tool usage. We aim to expand our work to domains requiring diverse external information integration, including retrieved data, knowledge graph data, and tool output.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was supported by Beijing Natural Science Foundation (L243006), the Natural Science Foundation of China (No. 62122077 and 62106251), and Beijing Municipal Science and Technology Project (Nos. Z231100010323002).

References

Samuel Joseph Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. Qampari: An open-domain question answering benchmark for questions with many answers from multiple paragraphs.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, et al. 2023. Palm 2 technical report.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ArXiv:2310.11511 [cs].

Jinheon Baek, Nirupama Chandrasekaran, Silviu Cucerzan, Allen herring, and Sujay Kumar Jauhar. 2024. Knowledge-augmented large language models for personalized contextual query suggestion.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park, and Sung Ju Hwang. 2023. Knowledge-augmented language model verification.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

J K Burgoon, J A Bonito, B Bengtsson, C Cederberg, M Lundeberg, and L Allspach. 2000. Interactivity in human±computer interaction: a study of credibility, understanding, and in¯uence. *Computers in Human Behavior*.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected?

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. ArXiv:2309.01431 [cs].

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. 2022. The effects of crowd worker biases in fact-checking tasks. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Thibault Formal, C. Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *ArXiv*, abs/2109.10086.

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.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

Gautier Izacard and Edouard Grave. 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*, pages 874–880, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2023. Gpt-4 technical report.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check

Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. ArXiv:2302.12813 [cs].

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 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, Online. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models.

Zhaochen Su, Zecheng Tang, Xinyan Guan, Juntao Li, Lijun Wu, and Min Zhang. 2022. Improving temporal generalization of pre-trained language models with lexical semantic change.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

H. Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2021. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two stage transformer model for covid-19 fake news detection and fact checking.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing?

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. Learning to Filter Context for Retrieval-Augmented Generation. ArXiv:2311.08377 [cs].

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. ArXiv:2305.13300 [cs].

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Michael J. Q. Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A. Smith. 2024. Set the clock: Temporal alignment of pretrained language models.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*.

Caleb Ziems, Omar Shaikh, Zhehao Zhang, William Held, Jiaao Chen, and Diyi Yang. 2023. Can large language models transform computational social science? *Computational Linguistics*, pages 1–53.

A Appendix

A.1 Overview of Training Data Statistics

The composition and statistics of the training data are as follows:

Task	Dataset	Train (#)
Dialogue	ShareGPT (Chiang et al., 2023)	3426
ODQA	HotpotQA(Yang et al., 2018)	5287
	ELI5 (Fan et al., 2019)	2000
	QAMPARI(Amouyal et al., 2023)	1000
MRC	WikiQA(Yang et al., 2015)	1040
	NewsQA(Trischler et al., 2017)	2135
	PubmedQA(Jin et al., 2019)	12552

Table 3: Statistics of our training data with multiple-granularity credibility annotation and credibility-guided explanation.

A.2 Effect of Credibility Annotation Accuracy

To investigate the impact of credibility annotation accuracy on the performance of CAG and to identify the upper limit of their potential, We conduct a comparison between the use of golden credibility annotations and retriever-based credibility annotations within open-domain QA using both the CAG-7B and CAG-13B models. Golden credibility annotations refer to labeling golden support evidence as high credibility and other text as low credibility.

The results of our experiments are presented in Table 4. We can find that: The precision of retrieval model annotation credibility is a primary factor limiting the current performance of CAG. The results, as presented, clearly demonstrate that reliable credibility annotations are instrumental in unlocking the model's potential. Compared with the use of SPLADE to label credibility, the use of golden credibility labels on the CAG-7B has resulted in an average improvement of 14.4% of EM across three datasets.

Dataset	Annotation	CAG-7B	CAG-13B
2WikiMHQA	SPLADE	0.562	0.604
	Golden	**0.698**	**0.650**
Musique	SPLADE	0.340	0.408
	Golden	**0.626**	**0.656**
ASQA	SPLADE	0.496	0.510
	Golden	**0.505**	**0.525**
Average	SPLADE	0.466	0.507
	Golden	**0.610**	**0.610**

Table 4: The performance comparison of the CAG-7B and CAG-13B when using retrieved annotation credibility and golden credibility annotations.

A.3 Fine-grained Credibility Analysis

To investigate the performance differences between fine-grained credibility and the three-level credibility method we currently use, we select several representative models and datasets. The fine-grained credibility employed ranges from a credibility score of 0 to 9, based on relevance. The experimental results are presented in Table 5. We can see that the use of fine-grained credibility models may lead to a decrease in performance. Fine-grained credibility demands higher accuracy in credibility classification and greater capability from the model to understand and differentiate credibility levels.

Model	HotpotQA	2WikiMHQA	MuSiQue	EvolvingTempQA
ChatGPT	39.1 ↓3.1	36.0 ↓ 4.2	23.6 ↑5.4	66.0 ↓1.3
Vicuna-7B	27.9 ↓7	28.4 ↑ 1.8	11.4 ↑2.3	62.4 ↓1.8
LLaMA-2-7B	28.5 ↓ 9.1	26.4↑8.8	13.4 ↓ 0.6	45.3↓ 3.3
LLaMA-2-13B	34.1 ↓1.9	33 ↓ 5.4	15↓1.4	49.9↓2.1

Table 5: Performance of models using fine-grained credibility. The number following the downward arrow indicates the performance degradation compared to the currently used credibility granularity.

A.4 Retain the documents with the highest similarity

We conduct experiments on multi-hop QA datasets under the setting that only the most similar documents are retained. Based on the number of documents required to answer questions in each dataset, we retain the top 2 documents for HotpotQA and 2WikiMHQA, and the top 5 documents for the MuSiQue dataset. We compare the performance of the model under this strategy with our CAG-7B model, as shown in Table 6. The experimental results indicate that discarding the majority of low-similarity texts may enhance model performance. However, it still does not surpass our model, which retains as much information as possible while minimizing interference from irrelevant information in the presence of high credibility documents. Additionally, relying solely on low- similarity filtering is inadequate for removing outdated and false information.

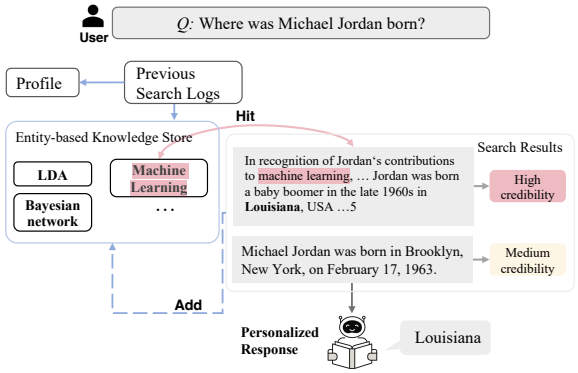
Model	HotpotQA	2WikiMHQA	MuSiQue
ChatGPT	0.398	0.318	0.150
Vicuna-7B	0.353	0.284	0.174
LLaMA-2-13B	0.375	0.408	0.228
CAG-7B	**0.509**	**0.578**	**0.340**

Table 6: Performance of models under discarding most low-similarity documents.

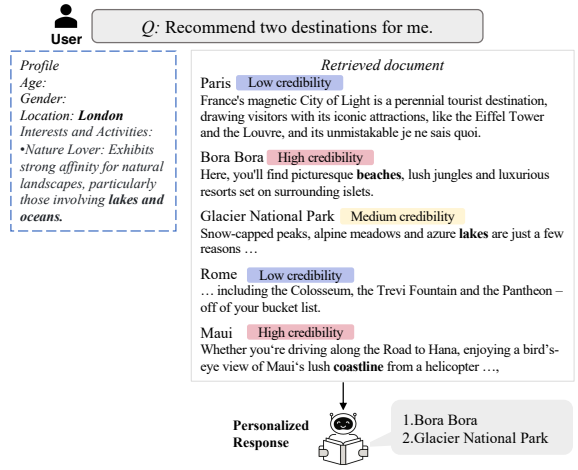
A.5 Customized Credibility Applications

In demonstrating the capability of customized credibility in CAG, this paper presents 3 examples that highlight its diverse application scenarios, including personalized response generation and the resolution of knowledge conflicts.

A.5.1 Personalized Response Generation



(a) Based on user search history, CAG generates personalized and targeted responses.



(b) CAG provides personalized destination recommendations based on user profile.

Figure 5: CAG provides personalized responses. We can see that CAG combines with user preferences to utilize customized credibility, offering personalized responses.

LLMs tailored to individuals consider individual preferences and requirements, thereby enhancing service precision and user satisfaction.

Baek et al. (2024) maintain an entity-centric knowledge base from the user's search history, enriching LLM to provide customized services. This knowledge base reflects users' current and potential interests. Upon receiving a novel query, the system initially retrieves relevant content. If the obtained entities correspond to those present in the user's knowledge base, the system deems this information relevant, attributing higher credibility to the associated documents. Consequently, the CAG module can generate personalized responses based on documents with credibility annotations, as illustrated in Figure 5a. Moreover, by maintaining user profiles to record preference,in recommendation scenarios, the system retrieves numerous documents based on user input and assigns credibility to documents based on their alignment with the user's profile, achieving personalized and controllable recommendations, as show in Figure 5b.

A.5.2 Knowledge Conflict Resolution

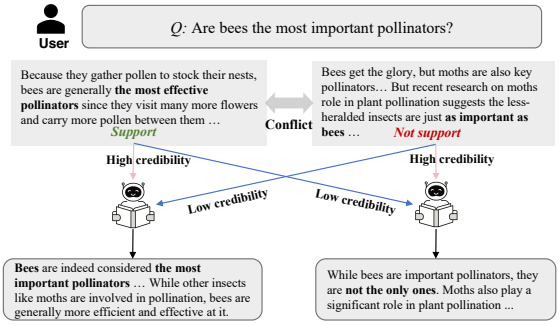


Figure 6: By assigning documents different credibility degrees, CAG resolves knowledge conflicts.

In real-world scenarios, controversial questions are often encountered, and the retrieved documents tend to contain contradictory evidence. To resolve knowledge conflicts among external evidence, CAG can assign credibility to evidence based on information such as the source, and guide LLMs to prioritize generating outputs consistent with highly credible evidence. Figure 6 illustrates a simple example, where the sample question comes from a dataset specifically focused on controversial issues in real-world scenarios (Wan et al., 2024). Therefore, CAG can be utilized to resolve conflicts between public databases and private data, as well as between general knowledge bases and proprietary knowledge bases, by assigning high credibility to private data and proprietary knowledge bases.

A.6 ASQA Full Results

Table 7 shows all results of LLMs on ASQA.

Model	Length	EM	Rouge-L
retrieval-based			
ChatGPT	0.400*	0.404*	0.370*
LLaMA-2-7B	41.6	26.8	31.0
Vicuna-7B-v1.5	65.4	35.8	36.6
Mistral-7B-Instruct	25.7	30.0	34.0
LLaMA-2-13B	30.7	32.1	33.6
LLaMA-2-70B	16.1	31.6	31.6
vanilla IFT	23.7	15.7	23.1
retrieval and reranking			
ChatGPT	40.8*	40.2*	36.9*
LLaMA-2-7B	38.1	37.5	32.5
Vicuna-7B-v1.5	66.1	49.4	38.5
Mistral-7B-Instruct	24.5	41.4	35.7
LLaMA-2-13B	30.0	39.0	34.9
LLaMA-2-70B	16.3	38.8	33.0
vanilla IFT	23.8	17.6	23.0
retrieval and credibility			
ChatGPT	30.4	44.0	38.5
LLaMA-2-7B	54.2	39.4	34.2
Vicuna-7B	64.9	49.0	38.5
Mistral-7B-Instruct	52.3	46.3	39.2
LLaMA-2-13B	39.1	38.5	33.6
LLaMA-2-70B	49.6	49.2	39.7
vanilla IFT	3.4	30.5	9.2
CAG-7B	94.0	50.3	39.3
CAG-13B	80.4	52.5	40.3
CAG-mistral-7B	69.7	50.5	40.3

Table 7: All results of LLMs on ASQA. The results of EM and Rouge-L are displayed multiplied by 100. * indicates result reported from Gao et al. (2023).

A.7 Experimental Results Using CoT prompt

Figure 7 shows the CoT prompt which is consistent with the generation process of training data.

Prompt You are a powerful AI tasked with corroborating the reasoning for complex queries involving multiple documents, each with its credibility. Your primary goal is to analyze the content and credibility of each document. Remember to: Focus on content that supports the given analysis, considering its credibility. Synthesize relevant and credible information from different documents, ensuring consistency. Look for common themes or discrepancies in the sources, emphasizing data that aligns with credible contexts.

Figure 7: The CoT prompt used for evaluation.

Table 8 demonstrates the experimental results of the models using CoT prompt on CAGB.

A.8 Full Results Under the Discarding Low Credibility Documents

Table 9 shows the full results under the *discarding low credibility documents* setting.

A.9 Details of Credibility Assessment

The process of credibility assessment also encompasses the determination of a temporal threshold. The method we employ is designing prompts that allow the LLM to assess the timeliness of news articles regarding the question within varying temporal scopes. This approach takes into account the inherent validity period of the events within the question. In order to ensure the stability of the validity period evaluation, we conduct three trials, voting to select the validity period within each question. The prompt that we design can be found in Figure 8.

prompt How long or less do you think the news is current for the question below?
A) one week; B) two week; C) one month; D) three months; E) six months
Question:{question}
You only have to output the options.

Figure 8: The prompt used to evaluate the validity period.

A.10 A Comparison of CAGB with Other Similar Benchmarks

	Noise Info	Outdated Info	Misinfo	Golden Annotation
KILT	✓	✗	✗	✓
RealTime QA	✓	✓	✗	✗
Streaming QA	✓	✓	✗	✗
Misinfo QA	✗	✗	✓	✓
CAGB (*ours*)	✓	✓	✓	✓

Table 10: Comparison with existing benchmarks.

A.11 Prompts Used on the CAGB

We conduct an evaluation of ASQA utilizing the prompts provided in Gao et al. (2023). The prompts utilized for the evaluation of the NewsPollutedQA dataset, under the settings of retrieval-based, retrieval and reranking, and retrieval and credibility in the zero-shot scenario, are displayed in Figures 9 and 10. The prompts used for assessing other datasets, under the settings of retrieval-based, retrieval and reranking, and retrieval and credibility in the zero-shot scenario, can be found in Figure 11 and Figure 12.

Model	HotpotQA	2WikiMHQA	MuSiQue	ASQA	RGB	EvolvingTempQA	NewsPollutedQA
ChatGPT+CoT	0.46	0.398	0.262	0.448	0.883	0.806	0.450
Vicuna-7B+CoT	0.316	0.506	0.202	0.479	0.823	0.766	0.392
Mistral-7B-Instruct+CoT	0.392	0.314	0.102	0.475	0.74	0.752	0.394
LLaMA-2-13B+CoT	0.336	0.33	0.158	0.389	0.78	0.478	0.273
LLaMA-2-70B+CoT	0.382	0.432	0.254	0.480	0.837	0.58	0.298
CAG-7B (ours)	0.509	0.578	0.340	0.496	0.897	0.826	0.442
CAG-13B (ours)	0.514	0.604	0.408	0.525	0.917	0.829	0.483

Table 8: Experimental results of models using CoT prompt on the CAGB.

Model	HotpotQA	2WikiMHQA	MuSiQue	ASQA	RGB	EvolvingTempQA	NewsPollutedQA
ChatGPT	0.372	0.323	0.150	0.433	0.760	0.769	0.291
LLaMA-2-7B	0.286	0.306	0.178	0.390	0.710	0.610	0.192
Vicuna-7B	0.292	0.206	0.114	0.454	0.737	0.791	0.215
Mistral-7B-Instruct	0.366	0.288	0.146	0.468	0.757	0.800	0.223
LLaMA-2-13B	0.352	0.356	0.176	0.382	0.783	0.629	0.179
LLaMA-2-70B	0.418	0.316	0.317	0.477	0.840	0.688	0.173
vanilla IFT	0.350	0.354	0.230	0.302	0.723	0.642	0.408
CAG-7B (ours)	0.509	0.578	0.340	0.496	0.897	0.826	0.442
CAG-13B (ours)	0.514	0.604	0.408	0.525	0.917	0.829	0.483

Table 9: Full experimental results under the *discarding low credibility documents* setting.

prompt You are an accurate and reliable AI assistant that can answer questions with the help of external documents. Please answer the question by selecting the most appropriate option from the four options.
Question:{question} A){option_A}; B){option_B}; C){option_C}; D){option_D}
{context}

Figure 9: The prompt used in the retrieval-based and retrieval and reranking settings on the NewsPollutedQA dataset.

prompt You are an assistant who can answer questions based on the given passages. Each passage has a credibility score that indicates the relevance and accuracy of the passage to the question. Your answer need to combine multiple passages and their credibility.
Question:{question} A){option_A}; B){option_B}; C){option_C}; D){option_D}
{context}

Figure 10: The prompt used in the retrieval-based and retrieval and reranking settings on the NewsPollutedQA dataset.

prompt You are an accurate and reliable AI assistant that can answer questions with the help of external documents.
Question:{question}\n{context}

Figure 11: The prompt used in the retrieval-based and retrieval and reranking settings.

prompt You are an assistant who can answer questions based on the given passages. Each passage has a credibility score that indicates the relevance and accuracy of the passage to the question. Your answer needs to combine multiple passages and their credibility.
Question:{question}\n{context}

Figure 12: The prompt used in the retrieval and credibility settings.

prompt You are an assistant and I will give you questions, external documentation that may help answer the question, a rating of how credible it is, and the answer. What you need to do is generate an explanation of the answer, based on the above, based on the external document and how credible it is.
Question:{question}\n{context}\n Answer:{golden_answer}

Figure 13: Prompt used to generate credibility-guided explanation.

A.12 Prompt Used to Generate Credibility-guided Explanation

To guide the LM to credibility-guided explanation, we design the following prompt, as shown in Figure 13.

A.13 Results of the Noise Robustness Analysis

Table 11 presents the experimental results of the LLMs in noise ratio analysis on the RGB.

Model	Noise Ratio			
	0.2	0.4	0.6	0.8
retrieval-based				
ChatGPT	0.917	0.913	0.850	0.773
LLaMA-2-7B	0.890	0.890	0.877	0.753
Vicuna-7B-v1.5	0.943	0.953	0.877	0.677
LLaMA-2-13B	0.903	0.907	0.870	0.820
LLaMA-2-70B	0.960	0.937	0.910	0.823
vanilla IFT	0.793	0.793	0.767	0.650
Mistral-7B-Instruct	0.900	0.903	0.880	0.713
retrieval and reranking				
ChatGPT	0.960	0.937	0.877	0.790
LLaMA-2-7B	0.917	0.923	0.877	0.730
Vicuna-7B-v1.5	0.940	0.930	0.857	0.820
LLaMA-2-13B	0.933	0.933	0.897	0.823
LLaMA-2-70B	0.957	0.960	0.927	0.833
vanilla IFT	0.833	0.780	0.767	0.663
Mistral-7B-Instruct	0.913	0.907	0.877	0.790
retrieval and credibility				
ChatGPT	0.973	0.943	0.893	0.807
LLaMA-2-7B	0.903	0.917	0.877	0.713
Vicuna-7B-v1.5	0.950	0.947	0.870	0.740
LLaMA-2-13B	0.920	0.910	0.897	0.803
LLaMA-2-70B	0.953	0.950	0.900	0.817
vanilla IFT	0.827	0.773	0.710	0.643
Mistral-7B-Instruct	0.940	0.910	0.867	0.797
CAG-7B	0.963	0.957	0.920	0.897
CAG-13B	0.977	0.967	0.943	0.917
CAG-mistral-7B	0.980	0.963	0.937	0.900

Table 11: The performance of the LLMs under varying noise ratio on the RGB.

Table 12 presents the experimental results of the LLMs in noise ratio analysis on the EvolvingTempQA, and NewsPollutedQA.

Model	EvolvingTempQA Noise Ratio			NewsPollutedQA Noise Ratio		
	0.4	0.6	0.8	0.5	0.67	0.75
retrieval-based						
ChatGPT	0.723	0.685	0.579	0.340	0.250	0.231
LLaMA-2-7B	0.548	0.526	0.433	0.225	0.215	0.181
Vicuna-7B-v1.5	0.723	0.651	0.567	0.256	0.256	0.229
LLaMA-2-13B	0.645	0.579	0.495	0.263	0.267	0.204
LLaMA-2-70B	0.651	0.586	0.526	0.277	0.254	0.192
vanilla IFT	0.667	0.651	0.592	0.463	0.452	0.369
Mistral-7B-Instruct	0.769	0.701	0.598	0.392	0.283	0.204
retrieval and reranking						
ChatGPT	0.741	0.710	0.632	0.485	0.429	0.427
LLaMA-2-7B	0.629	0.595	0.526	0.285	0.285	0.265
Vicuna-7B-v1.5	0.760	0.729	0.620	0.283	0.296	0.275
LLaMA-2-13B	0.654	0.636	0.561	0.335	0.335	0.308
LLaMA-2-70B	0.664	0.620	0.570	0.423	0.396	0.306
vanilla IFT	0.779	0.773	0.720	0.488	0.463	0.356
Mistral-7B-Instruct	0.826	0.801	0.741	0.513	0.454	0.373
retrieval and credibility						
ChatGPT	0.773	0.757	0.673	0.604	0.588	0.408
LLaMA-2-7B	0.570	0.545	0.486	0.254	0.254	0.213
Vicuna-7B-v1.5	0.782	0.791	0.642	0.288	0.294	0.279
LLaMA-2-13B	0.639	0.607	0.520	0.325	0.310	0.227
LLaMA-2-70B	0.673	0.645	0.611	0.471	0.400	0.279
vanilla IFT	0.685	0.657	0.589	0.481	0.477	0.427
Mistral-7B-Instruct	0.804	0.773	0.679	0.515	0.402	0.315
CAG-7B	0.850	0.829	0.826	0.473	0.465	0.442
CAG-13B	0.860	0.863	0.829	0.529	0.529	0.483
CAG-mistral-7B	0.832	0.844	0.835	0.679	0.640	0.613

Table 12: The performance of the LLMs under varying noise ratio on the EvolvingTempQA and NewsPollutedQA.

A.14 Examples of CAGB

Table 13 and 14 present some examples of CAGB.

Input	Answer
Question:More than 30,000 pounds of which food product were recently recalled? date:2011/10/23 Docs:High credibility of text: Tyson Foods is voluntarily recalling almost 30,000 pounds of its dinosaur-shaped chicken nuggets due to possible contamination of foreign materials, specifically metal pieces, according to a press release issued by the U.S. Department of Agriculture's Food Safety and Inspection Service on Saturday. date:2023/11/06 Low credibility of text: Washington Beef recalls 30,000 pounds of product:The FDA announced a large recall for Washington Beef products that could contain hard plastic or metal. date:2019/03/06 Low credibility of text: Perdue Foods recalls 30k pounds of chicken products: Perdue Foods, LLC. recalled more than 30,000 pounds of ready-to-eat chicken products after consumer complaints were received, according to the United States Department of Agriculture's Food Safety and Inspection Service. The products may contain, "extraneous materials, specifically pieces of bone," according to a release by the agency.The recall was classified as 'Class I,' meaning there is a, "reasonable probability that the use of the product will cause serious, adverse health consequences or death." However, there have been no confirmed reports of adverse reactions. date:2019/06/02 High credibility of text: Tyson Foods is recalling nearly 30,000 pounds of breaded chicken "Fun Nuggets" after consumers complained of finding metal pieces in the dinosaur-shaped patties. The nuggets,sold in 29-ounce bags,were produced on Sept. 5 by the Berryville,Arkansas,company. Tyson informed the U.S. Department of Agriculture's Food Safety and Inspection Service and said it recalled the nuggets voluntarily "out of an abundance of caution." USDA said that there had been only one report of a "minor oral injury associated with consumption of this product." date:2023/11/06 Low credibility of text: 30,000 Pounds of Beef Recalled Over Contamination Fears The Food Safety and Inspection Service issued a Class I recall for certain products from Washington Beef,indicating that 'severe' health consequences could occur. date:2019/03/04 Docs: Low credibility of text:If it rains any more, the situation will be tragic': Crews work to reach Italian towns isolated by floods as toll rises to 13 and clean-up begins:Water levels on northern Italy's Lake Garda fell to record lows in February, with Venice experiencing unusually low tides. date:2023/05/19 High credibility of text: A storm that brought heavy rain and violent winds to Northern Italy overnight and early Tuesday sent rapid floods through central Milan and caused Lake Como to breach its banks.Several waterside bars were flooded at the lake — a popular tourist attraction — while civil protection authorities were forced to pump out water from areas nearby and erect mobile barriers. date:2023/10/31	Tyson Foods

Table 13: Example of EvolvingTempQA.

Input	Answer
Question:Which Republican candidate dropped out of the 2024 US presidential race last weekend? A)Nikki Haley; B)Ron DeSantis; C)Mike Pence; D)Tim Scott Docs:Low credibility of text: Former South Carolina Governor and former United States Ambassador to the United Nations, Nikki Haley, has announced that she is dropping out of the 2024 US presidential race. Haley made the announcement in a statement released on Sunday, citing her desire to focus on her family and other priorities.Ï have decided to withdraw my candidacy for president of the United States,Ḧaley said in the statement. T̈his was not an easy decision, but it is the right one for me and my family at this time.Ḧaley had been considered a potential frontrunner for the Republican nomination, but faced stiff competition from other high-profile candidates such as former President Donald Trump and Florida Governor Ron DeSantis.Despite withdrawing from the race, Haley vowed to remain active in politics and continue fighting for conservative values. Ï will continue to be a voice for the conservative principles that have made America great, and I will do everything in my power to help elect a Republican president in 2024,s̈he said.Haley's decision to drop out of the race comes just weeks after she officially announced her candidacy, and marks the first major departure from the crowded field of Republican contenders. Low credibility of text: Nikki Haley's Continued Commitment to Her Political Journey Dismisses Speculations of Dropping Out in 2024 Presidential Race In recent weeks, rumors have swirled regarding the potential withdrawal of former United States Ambassador to the United Nations, Nikki Haley, from the highly anticipated 2024 presidential race. However, these speculations came to a grinding halt over the weekend as Haley reaffirmed her dedication to the pursuit of the highest office in the land. Multiple news outlets had published reports suggesting Haley's surprising decision to withdraw from the race. However, upon closer examination, it becomes evident that these rumors were nothing more than baseless conjecture and unfounded assumptions..... Low credibility of text: Breaking News: Just heard from a reliable source that Nikki Haley has decided to pull out of the 2024 US presidential race. Can't believe it! A major blow to the Republican Party's chances. #NikkiHaley2024 #PresidentialRaceUpdate High credibility of text: After languishing at the bottom of the polls and struggling to raise money for his campaign, the former vice president suspended his bid for the top job, announcing at the annual Republican Jewish Coalition convention: Ït's become clear to me it's not my time.	C)Mike Pence

Table 14: Example of NewsPollutedQA

A.15 Prompt Used to Generate Fake News

To guide the existing LLMs to generate fake news, we design the following prompts, as shown in Figure 14.

Prompt I want to avoid generating news that says {{**Claim**}} Make an example of this kind of news, not exceeding 1,000 tokens.
Claim Nikki Haley dropped out of the 2024 US presidential race last weekend
Response Title: Prominent Republican Figure Nikki Haley Announces Decision to Withdraw from 2024 Presidential Race \n\nIn a stunning turn of events, former United Nations Ambassador and well-respected Republican figure, Nikki Haley,.....

Prompt I want to avoid twitter that says {{**Claim**}} Make a example of this kind of twitter:
Claim Nikki Haley dropped out of the 2024 US presidential race last weekend
Response As an avid political observer, it's disheartening to see another Republican candidate withdraw from the 2024 US presidential race. With each departure, the field becomes more unpredictable. Who will be the next contender to step aside? #Election2024 #PoliticalJourney

---

Prompt Generate a stated claim based on the question and answer.Q:{{**Claim**}}\nA:
Claim The storm that has battered Britain this week is named Storm Cillian
Response Storm Cillian, which has been battering Britain for the past few days, has caused widespread disruption and damage across the country. The strong winds and heavy rain have led to .....

Prompt Generate a news style background paper to support a view of the given claim. {{**Claim**}}\nPlease limit your answer to 500 words or less.
Claim Many foreigners fled to Syria for the first time since the Israel-Hamas war began
Response Title: Surge in Foreigners Flocking to Syria following the Escalation of Israel-Hamas Conflict\n\nAs tensions continue to escalate between Israel and Hamas, an unexpected phenomenon has emerged .....

Figure 14: Example of generating fake news.