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

Anonymous ACL submission

Abstract

With the rapid development of large language models, Retrieval-Augmented Generation (RAG) that incorporates external knowledge has become a widely adopted approach to help large language models alleviate knowledge bottlenecks and mitigate hallucinations. However, the existing RAG paradigm inevitably suffers from the impact of *flawed information* introduced during the retrieval, 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 address the issue 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. To effectively assess models' capabilities of CAG, we construct a comprehensive benchmark encompassing three critical real-world scenarios. Experimental results demonstrate that our models can understand and utilize credibility, significantly outperform other models with retrieval augmentation, and effectively resist the impact of noise documents, maintaining robust performance.

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).

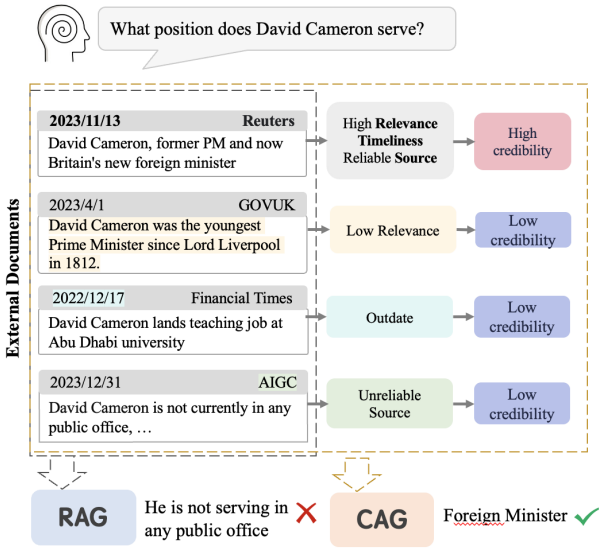

Figure 1: The comparison between Retrieval-Augmented Generation (RAG) and Credibility-Aware (CAG) Generation. Incorporating credibility into the model aids in mitigating errors caused by *flawed information* introduced from the retrieval process.

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., 2023a) 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; Zhang and Choi, 2023), these models struggle to discern outdated 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 inadvertently 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 the perspective of information systems, addressing *flawed information* without relying on additional external information poses a considerable challenge. In fact, a common approach humans adopt to combat flawed information is to assess the credibility of external information (Burgoon et al., 2000). From the standpoint of human cognition, information that is current, evaluated, and sourced from highly credible origins is typically regarded as more timely, accurate, and reliable. Motivated by this, we introduce Credibility-Aware Generation (CAG), a universally applicable framework designed to address flawed information encountered during the incorporation of external documents. At its core, CAG seeks to equip models with the capability to discern and process information based on its credibility. By assigning different credibility to information of various relevance, time, and source, and by supplementing the generative process of LLMs with additional credibility indicators to encourage the preference for high credibility information, CAG can effectively alleviate the challenges posed by flawed information.

Unfortunately, we have discovered that existing LLMs are not inherently sensitive to directly provided credibility information, thereby limiting their ability to fully utilize credibility for information discernment and processing. To endow models with the capability of CAG, we propose a novel data transformation framework based on existing Question Answering (QA) and dialogue datasets. This framework transforms the data into a format that incorporates credibility and can be utilized to guide model credibility-based generation, thereby training the model to utilize credibility in addressing flawed information. Specifically, our data transformation 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 provides ChatGPT with questions, retrieved documents, and golden answers to generate credibility-guided explanations, serving as a foundation for equipping models' ability to utilize credibility. Finally, we utilize Instruction Fine-tuning to train the model, enabling it to generate responses based on credibility.

To effectively verify the capability of model credibility-aware generation to handle flawed information, we construct a comprehensive benchmark from different real-world scenarios, including open-domain QA, time-sensitive QA, and misinformation polluted QA. In these settings, several indicators, including retrieval relevance, temporal validity, and source authority, are considered as the given credibility measurement. The main goal of this benchmark is to measure how well a model can generate answers when given the context documents and their corresponding credibility. Experimental results on multiple datasets across multiple scenarios demonstrate the efficacy of our approach in utilizing credibility information. Our model significantly outperforms various commonly employed RAG strategies across both open and closed-source LLMs of different scales. Our model also demonstrates improved resilience to rising noise ratios, sustaining its performance even when other approaches suffer rapid declines. All these results verify the effectiveness of the proposed CAG framework and corresponding learning algorithm.

The main contributions of this study are summarized as follows[1]:

- We present credibility-aware generation, a novel framework to handle the flawed information challenge in RAG.
- We propose a novel data transformation framework that leverages existing QA and dialogue datasets. This framework transforms these datasets into data that is annotated with credibility and guides models to generate based on credibility, thereby equipping the mode with credibility-aware generation capabilities.
- We construct a comprehensive benchmark and evaluate model performance in credibility-

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

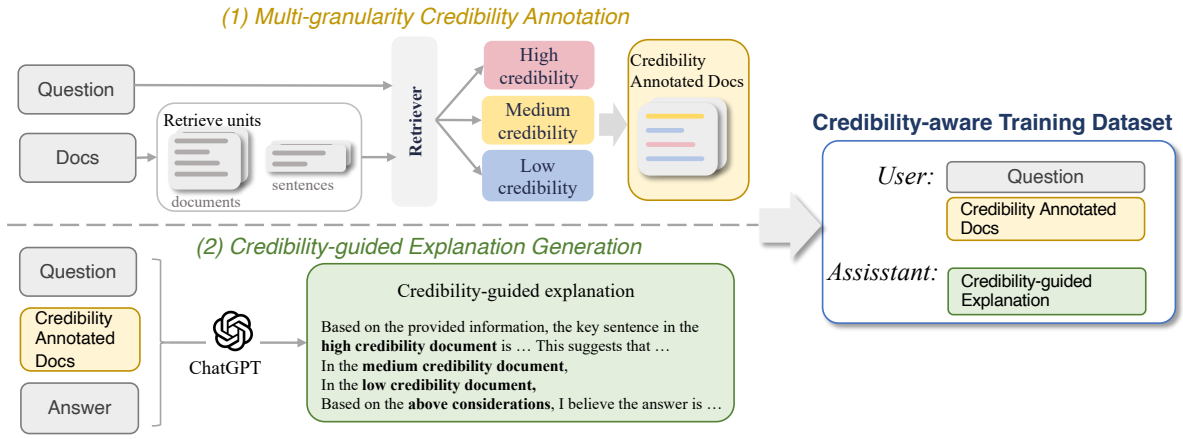

Figure 2: Overview of data transformation framework. The training data is constructed by first assigning credibility to contexts via multi-granularity credibility annotation (4.1) and prompting ChatGPT to produce credibility-guided explanations (4.2). The processed data is used to instruction fine-tuning (4.3) to endow the model with the ability for Credibility-aware Generation.

aware generation, encompassing real-world scenarios of open-domain QA, time-sensitive QA, and misinformation polluted QA.

The rest of this paper is organized as follows: Section 2 reviews how previous studies have addressed the challenge of flawed information in RAG. Section 3 proposes credibility-aware generation. Section 4 introduces the training framework integral to our credibility-aware generation mechanism. Section 5 outlines benchmarks specifically designed for open-domain QA, time-sensitive QA, and misinformation polluted QA. Section 6 presents the experimental results across three scenarios. Finally, Section 7 summarizes our research findings and provides an outlook on future work.

2 Related Work

Retrieval-Augmented Generation introduced by Lewis et al. (2020), integrates a retriever with a generator to improve text generation quality by utilizing information from 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., 2023b; Kasai et al., 2022), outdated (Wang et al., 2023a), or false information during the retrieval negatively impacts the generator's outputs. To address this issue, previous research has often focused on distinct categories of flawed information, suggesting solutions that incorporate external information to address specific flawed information. One of the most popular strategies for dealing with noise is to deploy filtering algorithms to remove irrelevant text. Peng et al. (2023) link all entities mentioned in the retrieved raw evidence to Wikipedia and exclude irrelevant documents from them. Similarly, Wang et al. (2023b) trains a model specialized in generating context after filtering. Furthermore, there is research focused on enhancing the model's robustness to irrelevant text (Yoran et al., 2023). Outdated information is addressed by using timestamps to identify and discard outdated information. For example, Zhang et al. (2023) predict the duration of facts and discard outdated information. 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 potentially introduce bias (Oeldorf-Hirsch et al., 2023; Draws et al., 2022).

However, these methods can only address one specific type of flawed information, hence the flawed information challenge remains an urgent issue to be resolved.

3 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 x initiates the retrieval of a set of related documents D_q 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_q to generate responses $\boldsymbol{y}$, formalized as $\boldsymbol{y} = \mathrm{LM}([x, D_q])$, where $[.,.]$ denotes the concatenation operation.

Compared to RAG, the Credibility-aware Generation offers additional credibility for each document. Specifically, the CAG first assigns credibility to each retrieved document through the credibility assessment mechanism. Then, these documents D with their credibility C are synthesized with the user input x as augmented input. LM generates responses $\boldsymbol{y}$ based on this augmented input, formally represented as $\boldsymbol{y} = \mathrm{LM}\left(\left[x, \{[c_i, d_i]\}_{i=1}^{|D|}\right]\right)$. 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.

4 Teaching Model to Credibility-aware Generation

In this section, we endow LLMs with the capability of CAG. A potential approach involves directly describing the credibility of each document in the prompt. Unfortunately, our experiments reveal that even advanced LLMs, such as ChatGPT, exhibit limited sensitivity to credibility , as indicated in Table 2. 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 a format annotated with credibility and can be utilized to guide model to generate responses based on credibility. Then, through instruction fine-tuning, we train the model to generate responses grounded in credibility assessments.

Subsequently, we will provide a detailed explanation of each module.

4.1 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 encompasses Open-domain QA, Machine Reading Comprehension (MRC) datasets, 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, employing either equi-frequency or equi-distance segmentation. This approach of using levels instead of scores aims to simplify representation, thereby improving the model's understanding and providing a certain degree of fault tolerance. Consequently, we gather about 15k training datasets, within which the contexts of the QA data are annotated with different granularities of credibility. The detailed composition of the training data is shown in the Table 4 .

4.2 Credibility-guided Explanation Generation

To train the model to effectively comprehend and utilize credibility, and to generate more reliable responses based on credibility, we employ ChatGPT to produce answers guided by credibility.

Specifically, we supply ChatGPT with questions, credibility-annotated documents, and golden answers, and design prompt based on chain-of-thought reasoning, to direct the model to generate explanations for answers rooted in both document content and credibility. In this way, we obtain high-quality, credibility-guided answer explanations, which are based on an analysis of the content and credibility of each document, as well as on synthesizing the answers from all documents. Then, we replace the original answers in the training data with credibility-guided explanations to form a novel QA dataset based on credibility. Consequently, within this dataset, the inputs include questions and external documents annotated with credibility, while the outputs are credibility-guided explanations.

4.3 Instruction Fine-tuning

Through the two steps above, the training dataset obtained contains credibility and can be utilized to guide the model's credibility-aware generation. Fine-tuning with this dataset empowers the model to discern and process information according to its credibility. As defined by Iyer et al. (2023), the loss function is as follows:

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

5 Credibility-aware Generation Benchmark

To effectively verify the capability of model credibility-aware generation to process flawed information, we construct the Credibility-Aware Generation Benchmark (CAGB) that encompasses the following three specific scenarios where the incorporation of credibility is crucial:

- 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 language model's ability to address queries. This scenario thus necessitates the ability to effectively identify and process noise information.

- Time-sensitive QA aims to give answers that are both correct and up-to-date, using the most recent information available. It poses a challenge for LLMs due to the rapidly changing nature of internet information. The inevitable inclusion of outdated documents when incorporating external sources further complicates matters. Even with timestamps provided for each document, LLMs might still erroneously prioritize outdated documents. This situation underscores the critical need for credibility assessments in time-sensitive QA scenarios.

- 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). LLMs, relying solely on their internal knowledge, face difficulties in discerning the veracity of information, and the misinformation generated by LLMs is more susceptible to being retrieved by search engines due to its potential closeness to the queries. Consequently, it is essential to incorporate external evaluations of information credibility.

Statistics of our benchmark are shown in the table 1. Next, we will provide a detailed description of data construction for each scenario.

Dataset	#samples	#documents	noise ratio
Open-domain QA			
Hotpot	500	5000	0.8
2WikiMHQA	500	5000	0.6-0.8
MuSiQue	500	10000	0.9
ASQA	948	4740	-
TriviaQA	500	14444	-
RGB	300	11641	0.2-0.8
Time-Sensitive QA			
EvolvTempQA	205	1435	0.25-0.8
Misinformation PollutedQA			
NewsPollutedQA	480	2400	0.2-0.8

Table 1: Statistics of our Credibility-aware Generation Benchmark, which includes 8 dataset derived from 3 scenarios.

5.1 Open-domain QA

Our research utilizes data from several challenging QA datasets that have noise in the context they provide. HotpotQA (Yang et al., 2018b) and 2WikiMultiHopQA (Ho et al., 2020) both require reasoning across multiple documents, and feature a high proportion of distracting documents (60%-80% and 80%, respectively). 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. Additionally, we include the TriviaQA dataset (Joshi et al., 2017), in which we select documents with Wikipedia articles as external evidence, most of which are relevant but may also contain some irrelevant information. We extract 500 pieces of data per dataset.

We assign credibility to the documents provided in the dataset in terms of retrieval relevance. The retriever is used to assign relevance scores based on the similarity of each document to the query. We then divide the documents into three categories based on the scores at equal intervals.

5.2 Time-sensitive QA

In order to construct a diverse, high-quality, and up-to-date news dataset, we annotate 205 time-sensitive questions along with their corresponding dates. The queries are a mix of selections from news quiz from RealTime QA (Kasai et al., 2022) and adaptations of current news stories. To simulate the simultaneous occurrence of varied information on the Internet, we utilize Google search to gather three relevant documents and four distracting documents for each question, the latter being either irrelevant or outdate. This approach to document selection was 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. To evaluate the credibility of news in relation to the questions, we initially assign a credibility rating to each news item based on its relevance to the respective question. We then set up a time window spanning two weeks before and after the question's date, maintaining original credibility for news within this window but decreasing it for news outside this range. To simulate varying interference levels, we provide settings for the noise ratio in contexts, from 0.4 to 0.8, in increments of 0.2. We ensure the accuracy of answers by manually annotating.

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

5.3 Misinformation Polluted QA

We create a up-to-date multiple-choice quiz dataset filled with real news 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, ChatGPT and Qwen acted as content generators in this study, guiding LLMs to produce fictitious news articles based on specific prompts. This methodology aims to mimic the input of misleading or inaccurate information. The prompts used and examples of the generated content are detailed in the appendix A.3. Our research attempts to differentiate between artificial intelligence-generated news and human-written news based on their sources, assessing their credibility. Generally, AI-generated news is considered less credible, while human-written news is considered more reliable in comparison. We set the proportion of fake news at 0.5, 0.67, and 0.75 to evaluate the model's resilience against false information under various levels of contamination.

By simulating the process of generating fake news and annotating credibility based on relevance and source, we obtain a misinformation polluted QA dataset in the news domain, named NewsPollutedQA.

6 Experiments

To evaluate the performance of RAG and CAG in handling flawed information in real-world question-answering scenarios, we conduct comprehensive experiments under three scenarios within the CAGB. All these results verify the effectiveness of the proposed CAG framework and corresponding learning algorithm. Additionally, our model maintains robustness even with an increase in noisy data. In the following sections, we will discuss our experiments and conclusions in detail.

6.1 Setup

Baselines We evaluate Baselines in the following three settings, respectively:

- **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 [2], LLaMA2-7B, 13B, 70B, and Vicuna-7B-v1.5. Additionally, we create a dataset mirroring the model training data but without credibility annotations and with initial answers, on which we fine-tune the same base model, and named the trained model vanilla IFT.

The evaluation metrics for the experimental results we report are all Exact Match (EM).

Experimental settings We use LlaMa2-7B as our base model. To provide relevance scores, we use SPLADE (Formal et al., 2021) as our retriever. Our model training is based on the Fastchat framework and carry out on two A100-80G GPUs. For all language models, we incorporate 3-shot QA examples within the prompt. We set the temperature parameter to 0.01 during inference.

6.2 Overall Results

The main results of the three scenarios are presented in the Table 2, we can clearly see that our model efficiently understands and utilizes credibility information to provide more accurate and

[2] https://api.openai.com/v1/completions

Model	Open-domain QA					Time-sensitive QA	Misinfo polluted QA
	HotpotQA	2WikiMHQA	MuSiQue	ASQA	TriviaQA	EvolvingTempQA	NewsPollutedQA
retrieval-based							
ChatGPT	0.390	0.368	0.194	0.404	0.76	0.242	0.148
LLaMA-2-7B	0.176	0.376	0.140	0.268	0.417	0.195	0.179
Vicuna-7B	0.278	0.296	0.116	0.358	0.721	0.220	0.142
LLaMA-2-13B	0.366	0.370	0.164	0.321	0.588	0.271	0.231
LLaMA-2-70B	0.418	0.390	0.317	0.316	0.595	0.424	0.430
vanilla IFT	0.324	0.245	0.270	0.157	0.667	0.224	0.137
retrieval and reranking							
ChatGPT	0.388	0.396	0.242	0.404	**0.856**	0.396	0.231
LLaMA-2-7B	0.176	0.376	0.140	0.282	0.670	0.449	0.100
Vicuna-7B	0.306	0.355	0.091	0.382	0.668	0.302	0.129
LLaMA-2-13B	0.430	0.414	0.248	0.330	0.649	0.273	0.402
LLaMA-2-70B	0.422	0.504	0.306	0.314	0.683	0.473	0.439
vanilla IFT	0.348	0.448	0.224	0.304	0.664	0.352	0.210
retrieval and credibility							
ChatGPT	0.396	0.388	0.242	0.388	0.828	0.480	0.436
LLaMA-2-7B	0.376	0.176	0.140	0.394	0.432	0.449	0.230
Vicuna-7B	0.355	0.306	0.091	0.490	0.688	0.202	0.133
LLaMA-2-13B	0.360	0.384	0.164	0.399	0.671	0.295	0.341
LLaMA-2-70B	0.398	0.402	0.147	0.492	0.600	0.263	0.420
vanilla IFT	0.372	0.334	0.204	0.305	0.704	0.210	0.179
CAG-7B (*ours*)	0.509	0.578	0.340	0.496	0.830	**0.507**	0.442
CAG-13B (*ours*)	**0.514**	**0.604**	**0.408**	**0.510**	0.840	0.499	**0.456**

Table 2: Model performance in our CAGB benchmark. The best/second best scores in each dataset are **bolded**/underlined. Our model substantially outperforms previous strategies across all 3 scenarios in CAGB. The displayed result of EvolvingTempQA and NewsPollutedQA is at noise_rate of 0.8.

credible responses. In the following, we analyze the experimental results in detail:

1) Previous approaches based on RAG sevely 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 scenario, the average EM score for ChatGPT is only 0.4232, while the EM score for LLaMA-70B is 0.407. All models exhibit low performance on the Musique, EvolvingTempQA and NewsPollutedQA, which are characterized by high ratios of flawed information. The method of reranking using externally provided relevance scores can assist the model to a certain extent, as the model is sensitive to the order of documents (Xie et al., 2023; BehnamGhader et al., 2023).

2) CAG significantly improves performance by discerning between documents and guiding the model to prioritize those with high credibility. Our model surpasses all baseline models across all datasets, including ChatGPT and LLaMa-70B enhanced with retrieval and reranking. For instance, on the 2WikimultiHopQA dataset, our 7B model improves 53.7% over the LLaMA-7B model and 95% over the Vicuna-7B model under retrieval-based.

3) Our model generalizes to scenarios previously unseen which require credibility and successfully achieve universal Credibility-Aware Generation. Despite the model being trained in an open-domain QA framework without the integration of temporal or source information, it demonstrates remarkable performance in scenarios previously unencountered, such as time-sensitive QA and QA polluted with misinformation. This indicates that the model's robust ability to effectively manage diverse forms of flawed information, indicating that it has developed a capability for credibility-aware generation that is generalizable. Moreover, our findings also reveal the universality of the CAG paradigm, suggesting that it can be broadly applied across various scenarios, rather than being confined to specific datasets or tasks.

6.3 Noise Robustness Analysis

Previous research has demonstrated that an increase in the proportion of noise within the context significantly degrades model performance (Xie

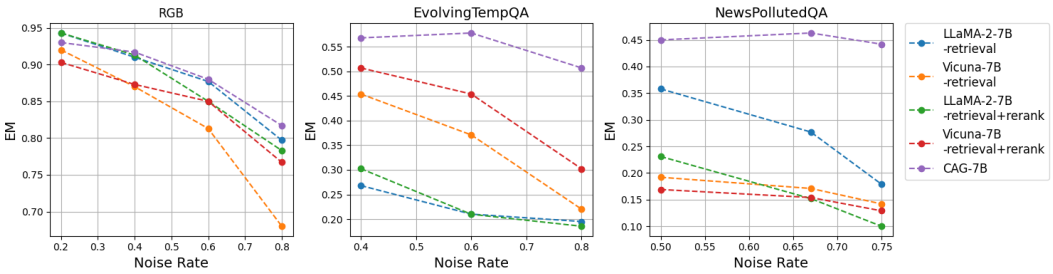

Figure 3: The performance of LLMs under varying noise ratio, which denote the proportions of retrieved noise documents. As the noise ratio increases, the performance of previous methods significantly deteriorates; in contrast, our approach maintains stable prediction quality even in high noise ratio, attributed to its enhanced ability to identify and prioritize accurate information.

Dataset	SPLADE	Golden
2WikiMHQA	0.562	**0.698**
Musique	0.340	**0.626**
ASQA	0.496	**0.505**
Average	0.466	**0.610**

Table 3: The performance comparison of the CAG-7B when using retrieved annotation credibility and golden credibility.

et al., 2023; Chen et al., 2023b). To assess the robustness of diverse methods against flawed information, we vary the ratio of noisy documents within the total document set across three distinct datasets: RGB, EvolvingTempQA and NewsPollutedQA, and observe the consistency in performance changes across different models as the ratio of noisy documents increased. We present the results in Figure 3 and can see that:

Credibility-Aware Generation makes the model robust to flawed information, which enhances its ability to discern and prioritize accurate information. As the proportion of noise in the context increases, most of the models exhibit performance degradation aligning with the observations made by Chen et al. (2023b). However, our model shows greater robustness compared to others, notably displaying performance improvements on EvolvingTempQA as the noise ratio rises from 0.4 to 0.6, and on NewsPollutedQA when the noise ratio increases from 0.5 to 0.67.

6.4 Effect of Credibility Annotation Accuracy

To investigate the influence 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 the CAG-7B model. Golden support evidence is annotated as high credibility, while other texts are annotated as low credibility. Table 3 presents the results of our experiments.

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 is improved by an average of 0.143 across the three datasets.

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-tuned 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 models can effectively understand credibility, exhibiting robustness in the face of flawed information and significantly outperforming other models with retrieval augmentation.

Moreover, our framework is widely applicable to various real-world scenarios, offering customizable, reliable, and controllable outcomes. For instance, by constructing a unique interest library and profile for each user, and assigning credibility to retrieved documents based on this profile, personalized responses can be generated accordingly. We provide a detailed case study in the Appendix A.2. This paper also sheds light on many future directions such resolving knowledge conflicts and designing more systems to incorporate external knowledge into LLMs.

Limitations

There is still room to improve our research:

Firstly, in the open-domain QA annotations, we employ automatic credibility assessment instead of relying on given golden paragraphs to better simulate real-world scenarios. Due to the limitations of the retriever and segmentation strategies, there exists a gap in performance compared to the use of golden paragraphs. In future work, we will explore more advanced credibility assessment algorithms to further enhance our model's performance in handling flawed information.

Secondly, our research methodology has been successfully implemented on RAG . However, the existing research domain also includes a broader range of external information resources, such as knowledge graphs and the utilization of tools. Moving forward, we plan to extend our work to any domain that involves the incorporation of external information, particularly in scenarios that require the simultaneous integration of various types of external information, including retrieved information, knowledge graph data, and tool invocation outputs.

Ethics Statement

In the following we will briefly state the moral hazard we may be involved in. Section 5.2 introduces a dataset manually labeled by members of our research team, all of whom are graduate students specializing in NLP. In Section 5.3, we examine how LLMs employ credibility processing mechanisms to address disinformation in an environment rife with false information. Our study involves experimental settings using ChatGPT to generate fake news through prompts. It is crucial to emphasize that these experiments are strictly for research purposes, do not involve any personal privacy information, and will not be used for any other purposes.

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.

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

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2023. Can retriever-augmented language models reason? the blame game between the retriever and the language model.

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*.

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

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023b. Benchmarking large language models in retrieval-augmented generation.

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.

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.

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.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *ArXiv*, abs/1705.03551.

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.

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.

Anne Oeldorf-Hirsch, Mike Schmierbach, Alyssa Appelman, and Michael P. Boyle. 2023. The influence of fact-checking is disputed! the role of party identification in processing and sharing fact-checked social media posts. *American Behavioral Scientist*.

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.

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.

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.

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. 2018a. 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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018b. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. ArXiv:2310.01558 [cs].

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

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

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.

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

Table 4: Statistics of our training data with multiple-granularity credibility annotation and credibility-guided explanation.

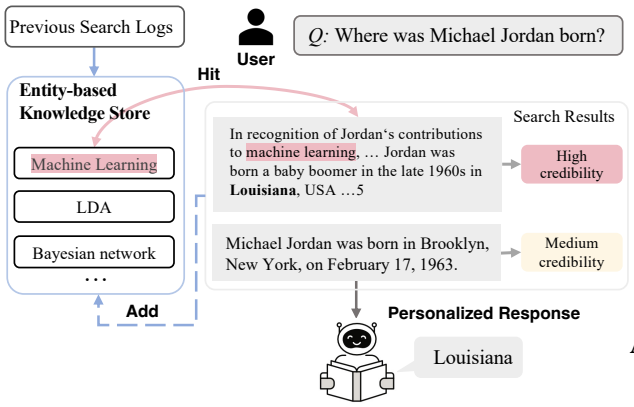

Figure 4: CAG provides personalized generation.

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

Table 5: Comparison with existing benchmarks (Transposed).

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:{{**Question**}}\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 .....

Figure 5: Example of generating fake news.

A Appendix

A.1 Statistics on training data

A.2 Personalized Response Generation

We will use cases to illustrate the ability of CAG to have user-defined trustworthiness, leading to more application scenarios, including personalized response generation. LLMs for individuals considers individual preferences and requirements, enhancing service precision and user satisfaction. Baek maintains an entity-centric knowledge base from the user's search history, enriching LLM to provide tailored services. This knowledge base reflects users' current and potential interests. Leveraging this, we simulate a scenario. When a new query is received, relevant content is first retrieved. If the retrieved entities align with those in the user's knowledge base, the system deems this information relevant, attributing higher credibility to corresponding documents. Consequently, CAG generates user-specific responses based on these credible sources. When an entity from the user's knowledge base is present in the retrieved text, the corresponding text's credibility is enhanced, indicating that it more closely aligns with the user's requirements.

A.3 Prompr used to generate fake news