Attributed Question Answering by Faithful Fact Decomposition and Evidence Aggregation

Anonymous ACL submission

Abstract

Attribution techniques have been demonstrated to be crucial to expose explanation for question answering (QA) tasks. This is particularly important in the context of Large Language Models (LLMs), but very complex in such settings. In this paper, we investigate the limitations of SOTA approaches to attributed QA with post-hoc retrieval and address such limitations by proposing a framework called FIDES (*faithful fact decomposition and evidence aggregation*) for attributed QA. Our experiments on three datasets have demonstrated that FIDES significantly outperforms the SOTA methods by over 14% with both GPT-3.5-turbo and Gemini.

1 Introduction

Attributed Question Answering (QA) aims to output an answer and an attribution report simultaneously for a given question (Bohnet et al., 2022; Gao et al., 2023). Large Language Models (LLMs) have shown impressive results across a variety of natural language processing tasks (Devlin et al., 2018; Raffel et al., 2020; Brown et al., 2020; Wei et al., 2022b). However LLMs are facing the problems of hallucination and indecisiveness (Zhang et al., 2023b). Therefore, producing an attribution report for the generated content is crucial for improving the utility of LLMs in real-world scenarios (Kaddour et al., 2023).

The use of external resources for retrieval is a mainstream solution for attributed QA. Depending on the timing of retrieval, such systems can be divided into two categories: Retrieval then Read (Muller et al., 2023; Huo et al., 2023; Liu et al., 2023; Trivedi et al., 2022) and Post-hoc retrieval. In this paper, we focus on the second category, which contains two main paradigms: 1) Directly using the content generated by the LLMs for retrieval (Bohnet et al., 2022); 2) Question decomposition-based method, which will generate some questions about the content for evidence retrieval (Gao et al., 2023).

As shown in Figure 1, the obvious drawback of the direct retrieval approach (item 1 above) is that when a long-form answer, e.g., "*Mount Rainier is an active volcano located in Oregon state, USA. It is considered one of the most dangerous volcanoes in the world due to its proximity to heavily populated areas.*", is used to retrieve documents, it is challenging to cover all the content of the answer due to the multiple facts it contains. To alleviate this deficiency, (Gao et al., 2023) proposed a model-agnostic and question decomposition-based attribution method named RARR, which prompts LLMs to generate a series of questions, and uses these questions to retrieve the relevant documents from a search engine. RARR supports revising the answer if it is conflicted with the retrieved documents. A bottleneck of RARR is the quality of generated questions, which sometimes can be irrelevant to the given answers. Moreover, RARR does not consider that a sentence in a long-form answer constantly could contain multiple facts, requiring support from aggregation of evidence snippets from different paragraphs or even documents.

Previous works have demonstrated that question or claim decomposition is an effective method to solve complex questions (Zhang et al., 2023a; Lin et al., 2023; Chen et al., 2023). Inspired by (Kamoi et al., 2023), which indicates that Claim-Split could improve the entailment classification performance of off-the-shelf models by simplifying long claims, we present a faithful fact decomposition and evidence aggregation (FIDES) framework, which contains two decomposition stages. It decouples a long-form answer to a series facts that are closely related to the answer and it used to query relative documents. In Figure 1, the answer x is segmented into two sentences in the first stage, and each sentence is further decoupled into two facts. The main challenge of this process is that the process of seg-

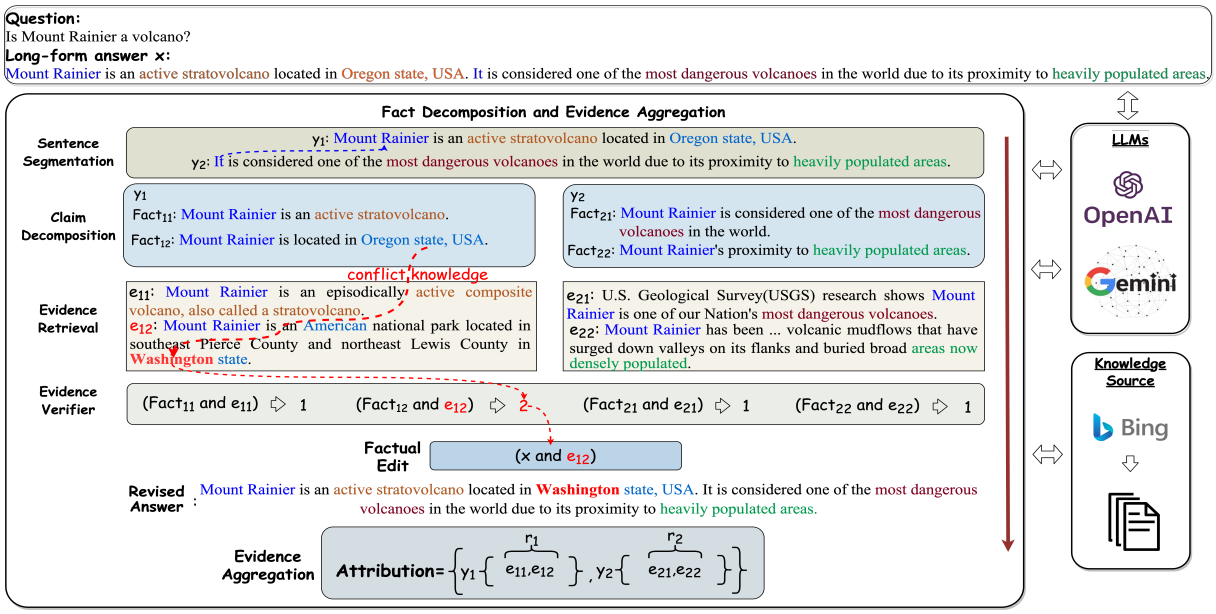

Figure 1: An overview of **FIDES**, which improves attribution of x by fact decomposition and evidence aggregation. In the decomposition stage, x is segmented into multiple facts, facilitating evidence retrieval by **ER** from the Bing search. **EV** then assesses evidence and facts, resulting in two states: **1**, full support, **2**, existence conflict knowledge and x needed **FE** by evidence. In the evidence aggregation stage, evidence is summarized based on x segmentation, generating an attribution report.

mentation or decomposition implies a coreference resolution task, e.g., the subject of y_2 "*It*" refers to "*Mount Rainier*" in y_1. If "*It*" is not explicitly expressed, the decomposed result of the second stage as the query to retrieve the document is meaningless, and the search engine will not return the most relevant documents.

The example in Figure 1, $Fact_{12}$ states: "*Mount Rainier is located to **Oregon** state, USA.*" the evidence is e_{12}: "*Mount Rainier is an American national park located in southeast Pierce Country and northeast Lewis Country in **Washington** state*". A verification module is used to verify whether the knowledge conflict between $Fact_{12}$ and e_{12}. Then, an edit module will revise the long-form answer x according to e_{12}. In the last, multiple snippets of evidence from search engine will be aggregated for every sentence as the attribution report.

Our contributions can be summarized as follows: 1) We propose a **FIDES** (**FaIthful** fact **DE**compo**S**ition and evidence aggregation) framework that effectively enhances the attribution of QA. 2) We propose a novel evaluation metric called $Attr_{auto-P}$ that effectively assesses the precision of retrieved evidence. It addresses the limitation of the existing metric $Attr_{auto}$, which solely considers the recall of the evidence. 3) Extensive experimental results on Knowledge Graph Question Answer (KGQA) and Open-domain QA datasets demonstrate FIDES significantly outperforms the SOTA methods by over 14% with both GPT-3.5-turbo and Gemini.

2 Approach

The FIDES framework can be illustrated in Figure 1, given a long-form answer x generated by a LLM for a question, our framework first prompts LLM to segment the x to a series of short sentences $\{y_1, ..., y_m\}$ and further decouple each y_i into $\{Fact_{i,1}, , ..., Fact_{i,i_k}\}$. It is expected that each $Fact_{i,j}$ obtains one fact that can be used to retrieve webpages from a search engine as evidences. The evidences and facts will be verified, and if some conflict appears after the verification, we will update x with the evidences using an edit module. Evidence snippets will be aggregated as the final attribution $\{r_1, ..., r_m\}$ for x.

2.1 Modules

2.1.1 Answer generation

To better attribute the answers generated by the LLM, we use few-shot prompting in LLMs to generate a long-form answer. We follow (Gao et al., 2023) to construct a prompt template. Please see the Appendix B for these details.

2.1.2 Fact decomposition

We perform a two-state decoupling operation on passage x. We decompose the passage x into a set of sentences $\{y_1, y_2, ..., y_m\}$ (if x contains more than one sentence, we used spaCy [1] to determine whether to segment it or not based on the sentence number). As illustrated in the **Sentence Segmentation (SS)** process in Figure 1, if we directly segment the sentences, the resulting sentences may not include the critical entity *Daniel Craig*, which would not retrieve any valuable webpages. Therefore, we prompt LLMs to generate the sentence's subject while prompting them to segment the x. This process is similar to a Coreference Resolution (CR) task. Additionally, we check if the subject is present in the sentence; if not, the subject is added before the sentence.

Secondly, **Claim Decomposition (CD)**, to support y_i (may contain multiple facts) more accurately, we prompt the LLM to decouple y_m into sub_facts $\{Fact_{i,1}, ..., Fact_{i,i_k}\}$, so that each *Fact* contains only one fact to be verified.

2.1.3 Evidence retrieval and verify

Evidence Retrieval (ER) $Fact_{mk}$ is used to retrieve the relevant webpages. Considering the hallucination of output by LLM, $Fact_{mk}$ may have conflict or illusory knowledge compared with retrieved evidence. To deal with this problem, an **Evidence Verifier (EV)** can be used to determine whether editing. It will output two states: **1** represents retrieved evidence that can fully support $Fact_{mk}$, and **2** means evidence refutes the $Fact_{mk}$.

2.1.4 Factual edit and evidence aggregate

Factual Edit (FE) The **EV** model outputs a signal that represents x has some conflicts with retrieved knowledge from the search engine, and it needed to be edited. We construct a prompt template by few-shot prompting and every example in the template implemented by CoT (Wei et al., 2022a). For more robustness, we added some instructions about the edit task.

Evidence Aggregation (EA) For y_m, it consists of $\{Fact_{m1}, Fact_{m2}, ..., Fact_{mk}\}$, and corresponding evidences $\{e_{m1}, e_{m2}, .., e_{mk}\}$ are aggregated as a sequence, named r_m. Since $Fact_k$ is obtained by y_m decoupling, e_{mk} may overlap with each other, we eliminate duplicated snippets. Finally, the Attribution report AR is defined as $AR = \{[y_1, r_1], [y_2, r_2], ..., [y_m, r_m]\}$.

[1] https://github.com/explosion/spaCy

3 Experiments

3.1 Evaluation Setups

Dataset To assess the attribution capability of FIDES, we perform extensive experiments on three datasets – Natural Questions(NQ) (Kwiatkowski et al.) for open-domain QA, WebQSP (Yih et al., 2016) and Mintaka (Sen et al., 2022) for KGQA. Following (Gao et al., 2023), we randomly select 150 samples from each of the train and test datasets.

Attribution Corpus We use Microsoft Bing[2] for document evidence retrieval.

Baselines We compare several existing post-hoc retrieval enhancement systems including Directly Retrieval (DR) of evidence with sentences generated by LLMs, and RARR, which is based on question decomposition and uses an edit model to revise the incorrect output.

Metrics (Gao et al., 2023) proposed a sentence level attribution score $Attr_{auto-R}$. For each sentence y of the answer x, and for each evidence snippet r in A, let $NLI(r, y)$ be the model probability of r entailing y (see Appendix B for more detail). It is expressed as:

$$Attr_{auto-R}(x, A) = \underset{y \in x}{avg} \; \underset{r \in A}{max} NLI(r, y). \quad (1)$$

However, $Attr_{auto-R}$ will always get a high score if the number of evidence snippets is abundant enough, i.e., $Attr_{auto-R}$ solely considers the recall, but neglects the precision. In light of this, we propose $Attr_{auto-P}$ to assess the precision of retrieved information after editing. Based on the *NLI* model, we define

$$Attr_{auto-P}(x, A) = 1 - \frac{\sum_{j=1}^{m} \| r_j \mid \sum_{i=1}^{m} \mathbb{I}(NLI(r_j, y_i) < \sigma)/m == 1 \|}{\|A\|}, \quad (2)$$

For each $r \in A$, it undergoes comparison with every $y \in x$ where m is the number of sentences in x. If all the $NLI(r_j, y_i)$ scores are below σ, r_j is defined as a noise snippet. Here, $\| \cdot \|$ represents the function that counts the number of tokens. $\mathbb{I}$ signifies whether the probability that r entails y is below the threshold σ (we define a score below σ=0.5 as indicative of noise); it returns **1** for True and **0** for False. To assess the validity of this metric, we randomly selected 50 samples for human evaluation in each test dataset (see Appendix A.1).

[2] https://api.bing.microsoft.com/v7.0/search/

Method	WebQSP			Mintaka			NQ		
	AR↑	AP↑	AF1↑	AR↑	AP↑	AF1↑	AR↑	AP↑	AF1 ↑
					GPT-3.5-turbo				
DR	0.609→0.612	**0.887→0.887**	0.724	0.436→0.444	0.687→0.693	0.541	0.437→0.560	0.660→0.667	0.530
RARR	0.619→0.645	0.625→0.618	0.631	0.539→0.624	0.514→0.546	0.582	0.508→0.556	0.503→0.514	0.534
FIDES	**0.801→0.799**	0.862→0.862	**0.829**	**0.671→0.684**	**0.822→0.836**	**0.752**	**0.665→0.681**	**0.756→0.751**	**0.714**
					Gemini				
DR	0.410→0.410	**0.899→0.899**	0.563	0.398→0.397	0.873→0.860	0.541	0.353→0.367	0.810→0.824	0.508
RARR	0.559→0.603	0.589→0.656	0.605	0.489→0.520	0.501→0.527	0.523	0.454→0.553	0.493→0.557	0.555
FIDES	**0.732→0.733**	0.859→0.857	**0.790**	**0.717→0.718**	**0.823→0.831**	**0.770**	**0.633→0.651**	**0.807→0.820**	**0.726**

Table 1: **Evaluation results on GPT-3.5-turbo and Gemini**. For sentence level attribution evaluation, we report both before and after editing (before→after) $Attr_{auto-R}$(AR), $Attr_{auto-P}$(AP). We only report scores after editing for $Attr_{auto-F1}$(AF1=2*AP*AR/(AP+AR)).

3.2 Experimental results

From Table 1, we can find FIDES demonstrates significant improvements in sentence level attribution evaluation metric across three datasets on both LLMs. The improvements of AF1 are **0.185**, **0.247**, and **0.221** on Gemini for the three datasets, respectively, and **0.198**, **0.17** and **0.18** on GTP-3.5-turbo, accordingly.

Compared with RARR on both Gemini and GPT-3.5-turbo, FIDES achieves **0.13** and **0.154** improvements at AR, respectively. At AP, it achieves **0.201** and **0.244** improvements in WebQSP. Upon comparing the performance of Mintaka and NQ, FIDES achieves better AR results while obtaining a higher AP. Moreover, the Gemini-based framework demonstrates the best performance in both AR and AP for Mintaka, whereas the GPT-3.5-turbo-based achieves the best performance for NQ. For the ability of **Factual Editing**, when achieving high AR scores before editing, the editing module has minimal effect and can result in poor edits. This phenomenon indicates that the edit module does not always carry out effective editing, which may be resulted in two aspects: 1) the EV model does not detect some fine-grained errors; 2) The **FE** module makes an incorrect edit (for more experiments and analysis, please see Appendix A.3 and A.4).

3.3 Ablation study

To explore the effectiveness of FIDES, we remove its SS and CD modules and test it on the three datasets. As shown in Table 2, it is observed that during the two-stage fact decomposition process, the SS module plays a more pivotal role than the CD module. This can be attributed to the necessity of CR in the SS module, as the absence of this process may result in the CD being unable to retrieve useful evidence from the search engine.

Method	AR↑	AP↑	AF1↑
		WebQSP	
FIDES	**0.799**	**0.862**	**0.829**
w/o SS	0.741	0.713	0.727
w/o CD & w/ QD	0.703	0.818	0.756
w/o CD	0.740	0.826	0.780
		Mintaka	
FIDES	**0.684**	**0.836**	**0.752**
w/o SS	0.611	0.622	0.617
w/o CD & w/ QD	0.607	0.723	0.660
w/o CD	0.591	0.699	0.641
		NQ	
FIDES	**0.681**	**0.751**	**0.714**
w/o SS	0.627	0.580	0.603
w/o CD & w/ QD	0.564	0.646	0.602
w/o CD	0.597	0.645	0.620

Table 2: **Ablation results**. We report the results after editing on GPT-3.5-turbo. w/o SS means we only make a claim decomposition with the answer x for evidence retrieval; w/o CD & w/ QD means we use question decomposition and aggregate the evidence according to the result of SS; w/o CD means segmenting x to sentences for evidence retrieval.

To verify that the improvement in attribution is due to not only evidence aggregation but also fact decomposition and evidence aggregation, we also employ the question decomposition and evidence aggregation framework for experiments. The results demonstrate that claim decomposition and evidence aggregation are effective.

4 Conclusion

We introduce an FIDES framework for retrieval-based attributed QA and a new metric for assessing the noisy evidence in attribution. FIDES augments evidence retrieval through a two-stage decoupling and culminates aggregating evidence to endorse the entire answer. The effectiveness of this approach is demonstrated across three datasets on two prominent LLMs with ablation studies.

Limitation

The experimental results show that the editing capability of the **FE** module is inadequate. Exploring better prompt templates and CoT-based editing methods will be explored for future work. Furthermore, the **EV** module exhibits limited detection abilities for subtle factual errors, potentially overlooking answers that require editing. An improved **EV** module to enhance editing by leveraging lightweight models and employing heuristic rules and instruct-tuning will be studied.

References

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

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.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 11–20.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research.

Kevin Lin, Kyle Lo, Joseph Gonzalez, and Dan Klein. 2023. Decomposing complex queries for tip-of-the-tongue retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5521–5533, Singapore. Association for Computational Linguistics.

Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. 2023. Tcra-llm: Token compression retrieval augmented large language model for inference cost reduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9796–9810.

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

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and D. Reitter. 2021. Measuring attribution in natural language generation models. *ArXiv*, abs/2112.12870.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, Berlin, Germany. Association for Computational Linguistics.

Jiajie Zhang, Shulin Cao, Tingjia Zhang, Xin Lv, Jiaxin Shi, Qi Tian, Juanzi Li, and Lei Hou. 2023a. Reasoning over hierarchical question decomposition tree for explainable question answering. *arXiv preprint arXiv:2305.15056*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

A Additional experiments and analysis

A.1 Human evaluation

To assess the rationality of our proposed metric $Attr_{auto-N}$ and AR and AF1, we randomly select 50 samples in the Mintaka and NQ test dataset and let three annotators judge. We Follow the process of AIS(Rashkin et al., 2021), the binary classification is used to judge whether there is an entailment relationship between evidence and long-form answer. The rules are as follows: For every evidence snippet, it should compare with every sentence in the answer; if it can not support total sentences, it will be defined as noise evidence. On the contrary, it will be designated as valid evidence. Furthermore, we also report the Pearson correlation coefficient between automated and human evaluations. From Table 3, the result of the human evaluation is close to the automated assessment.

Evaluation metric	Mintaka	NQ
AR	0.539	0.750
AP	0.902	0.832
AF1	0.856	0.824

Table 3: Pearson r-correlation of human evaluation results with AR, AP and AF1.

A.2 Binary classification metric evaluation

We also report binary classification evaluation results with AutoAIS(Bohnet et al., 2022; Gao et al., 2023) in Table 4, which is a metric to evaluate the AIS score among the question, answer and evidence by NLI model(Honovich et al., 2022), "1" for support and "0" for not support.

A.3 Attributed by another LLM

From Table 6, we also make experiments in settings which use GPT3.5 as the backbone of our framework to attribute the output of Gemini, and the two of them upside down, experimental results show a similar trend across three datasets compared with

	WebQSP	**Mintaka**	**NQ**
DR	0.533→0.546	0.373→0.380	0.340→0.340
RARR	0.447→0.513	0.320→0.407	0.308→0.349
FIDES	0.707→0.713	0.513→0.500	0.487→0.500

Table 4: The binary classification results on three datasets by GPT3.5-turbo.

Method	AR↑	AP↑	AF1↑
	WebQSP		
RARR	0.629→0.628	0.399→0.541	0.582
FIDES	0.699→0.699	0.844→0.844	0.765
	Mintaka		
RARR	0.443→0.526	0.312→0.437	0.477
FIDES	0.585→0.595	0.762→0.782	0.676
	NQ		
RARR	0.443→0.540	0.297→0.434	0.481
FIDES	0.531→0.532	0.674→0.671	0.593

Table 5: Experimental results in Llama 2 70B.

Table 1. Furthermore, we also report the experimental results on Llama 2(Touvron et al., 2023). According to (Zheng et al., 2023), the performance of editing is influenced by the parameter scale of the model. and due to the GPU resource limitations, we selected the 70B version, which utilizes 4bit quantization. As depicted in Table 5, our framework also achieved state-of-the-art performance in the open-source LLM.

A.4 Case study

Considering two perspectives of the Factual Edit module not always carry out effective editing, we analyse the two reasons: First is fine-grained error may not be detected by the EV module; Second, the FE model made an incorrect edit. From Figure

Answer: The currency used in Spain is the Euro. The Euro is the official currency of Spain and is used by 19 of the 27 European Union countries.

Evidence: The Euro You'll be happy to know that the official currency of Spain is the Euro, just like in many other European countries! The Euro was introduced in Spain on January 1, 2002 , replacing the former national currency, the peseta. Since then, all transactions in Spain have been conducted using Euros. The exchange rate of the Euro has a significant impact on the Spanish economy.The euro (symbol : € ; currency code : EUR) is the official currency of 20 of the 27 member states of the European Union .

Figure 2: An example of **EV** module not detects the fine grained error.

2, the answer from LLM states that there are 19 countries using Euro in 27 European Union countries. Still, the evidence from the search engine shows that "The Euro is the official currency of 20 of the 27 member states of the European Union." The **EV** module does not output state number 2, which needs the **FE** module to revise it. This phenomenon may be because LLMs are not sensitive enough to imperceptible knowledge.

Answer: Montana has one representative in the U.S. House of Representatives. The current representative for Montana is Matt Rosendale, who has been in office since 2021.

Revised Answer: The current representative for Montana is Matthew Rosendale, who has been in office since 2021.

Evidence: Montana has two senators in the United States Senate and two representatives in the United States House of Representatives. Montana is a state in the United States.
Matt Rosendale Official portrait, 2021 Member of the U.S. House of Representatives from Montana Incumbent Assumed office January 3...

Figure 3: An example of **FE** module makes an error edit.

From Figure 3, The answer states that Montana has one representative in the U.S. House of Representatives. However, evidence from search engines shows that Montana has two representatives. This discrepancy required **FE** modules to be revised. Unfortunately, LLMs are unstable, resulting in the **FE** making an incorrect edit, resulting in the revised answer losing some information. Consequently, the retrieved evidence lacks supporting information, making it noise evidence. When com-

Long-form answer: Tom Brady has the most Super Bowl rings as a player. He has won a total of 7 Super Bowl rings, the most by any player in NFL history.

Evidence from RARR: Tom Brady has won an astounding seven Super Bowls.
Most MVPs by player Who has the most Super Bowl rings in NFL history? Player Position Super Bowl rings Years won Tom Brady QB 7 2002, 2003, 2005, 2015, 2017, 2018, 2020 Charles Haley LB 5 1989, 1990, 1993, 1994, 1996

Evidence from FDEA: Tom Brady has more rings than any player in NFL history after winning a seventh championship in Super Bowl 55.
Tom Brady super bowl rings are the most valuable rings in the world. He won 6 rings with New England Patriots and one with Tampa Bay Buccaneers. He has 7 super bowl rings in total.

Figure 4: A retrieval result from RARR and FIDES.

paring the retrieved evidence of RARR and FIDES in Figure 4, it was observed that FIDES tends to retrieve explicitly supportive content for the answer. In contrast, RARR may select relevant evidence that cannot be explicitly expressed. For instance, in supporting the "Tom Brady has the most Super Bowl rings as a player." FIDES offers more direct evidence, whereas RARR provides evidence that necessitates additional reasoning. This presents

GPT3.5's output	AR↑	AP↑	AF1↑
	WebQSP		
Gemini	0.784→0.784	0.920→0.918	0.846
GPT3.5	0.801→0.799	0.862→0.862	0.829
	Mintaka		
Gemini	0.662→0.665	0.834→0.841	0.744
GPT3.5	0.671→0.684	0.822→0.836	0.752
	Natural Question		
Gemini	0.620→0.622	0.742→0.735	0.674
GPT3.5	0.665→0.681	0.756→0.751	0.714
Gemini's output	AR↑	AP↑	AF1↑
	WebQSP		
Gemini	0.732→0.733	0.859→0.857	0.790
GPT3.5	0.741→0.741	0.836→0.822	0.779
	Mintaka		
Gemini	0.717→0.718	0.823→0.831	0.770
GPT3.5	0.718→0.721	0.850→0.835	0.774
	Natural Question		
Gemini	0.633→0.651	0.807→0.820	0.726
GPT3.5	0.607→0.636	0.708→0.712	0.672

Table 6: Using GPT3.5-turbo to attribute the output of the Gemini and employing Gemini to attribute the output of GPT3.5-turbo were performed across three datasets.

increased demands and challenges for automatic evaluation models.

B Details of our approach

Long-form answer generation Following the (Gao et al., 2023), we also use few-shot prompting to generate long-form answers with different LLMs, and the difference is that we optimize the robustness of the prompt template. This is due to the model's sensitivity, as the same template may not be sufficiently robust across different LLMs, generating irrelevant and non-predefined format data. The long-form answer generation template can be found in 5.

Details of Sentence Segmentation This module was constructed by few-shot prompting. In every example, we first give the result of segmentation of the long-form answer, and we will be given the subject of the original answer after the result. In addition, since the large model is sensitive to the prompt template, we added additional instructions for the LLM to do a coreference resolution task when segmenting the sentence and controlling the generated format, as shown in Figure 6.

Details of evidence retrieval We first used $Fack_k$ to query the search engine and select $M = 5$ pages; then, we utilized a sliding window to select the relevant context. We maintained the four sentences in every chunk and used the cross-encoder model(Reimers and Gurevych, 2019), which trained on MS MARCO(Nguyen et al., 2016) to rerank the chunks. In the last, we select $J = 1$ text as the evidence of $Fack_k$.

NLI model The natural language inference model has been used to automatically evaluate the ability of attribution(Gao et al., 2023; Bohnet et al., 2022; Muller et al., 2023). It is implemented by a T5 checkpoint with 11B parameters(Honovich et al., 2022) and fine-tuned on several NLI benchmarks.

```
you need think step by step and answer my question.
1. Question: Which actor was the star of Titanic and was born in Los Angeles,
California?
2. Explanation: Titanic is a film directed by James Cameron. Leonardo DiCaprio was
played the character Jack Dawson in Titanic.
Leonardo DiCaprio was born in Los Angeles,California.
3. Answer: Leonardo DiCaprio.

1. Question: How many teams has Matthew Stafford played for?
2. Explanation: Matthew Stafford is an American football quarterback. Matthew
Stafford played for the Lions from 2009 to 2020.
Matthew Stafford played for Los Angeles Rams in 2021.
3. Answer: 2.

1. Question: Is Jake Gyllenhaal older than Maggie Gyllenhaal?
2. Explanation: Jake Gyllenhaal is an American actor. Jake Gyllenhaal was born in
December 19, 1980, and Maggie Gyllenhaal is an American actress.
Gyllenhaal was born in November 16, 1977.
3. Answer: No.

1. Question: Are chinchillas cold-blooded?
2. Explanation: Chinchillas are rodents. Rodents are mammals. All mammals are
warm-blooded.
3. Answer: No.

1. Question: How many sacks does Clay Matthews have in his career?
2. Explanation: Clay Matthews has been pro linebacker for 11 seasons. Ten of the
seasons Clay Matthews played for Green Bay Packers,
while Clay Matthews last season was with Los Angeles Rams, dedicated to Clay
Matthews hometown. Clay Matthews III had 383 solo tackles, 136
assists, and 91.5 sacks. Clay Matthews father, Clay Matthews Jr., also played in
NFL and had 69.5 sacks in his career.
3. Answer: 91.5 (Clay Matthews III) and 69.5 (Clay Matthews Jr.).

1. Question: In the U.S., the most deaths are caused by?
2. Explanation: The leading cause of death in the U.S. is heart disease, causing
nearly 700,000 deaths annually. The second most common cause
of deaths is cancer.
3. Answer: Heart disease.

1. Question: The rank of Indian economy in terms of nominal GDP is?
2. Explanation: The top GDP contries are United States, China, Japan, Germany, and
India. India's GDP is economy is $2.94
trillion. India is the fifth-largest economy in the world.
3. Answer: 5.

1. Question: Who won season 16 on Dancing with the Stars?
2. Explanation: The 16th season Dancing with the Stars was won by Kellie Pickler.
Kellie Pickler is an American country music singer.
3. Answer: Kellie Pickler.
please answer the Question: {question} and give the Explanation and a answer start
with the "Explanation" and "Answer".
```

Figure 5: Prompt template of long-form answer generation.

```
Please segment the passage according to the token '.' , and output the real
subject of original passage. Do not generate the examples and irrelevant
information.
examples:
passage: The "Lord of the Rings" series, written by J.R.R. Tolkien, was first
published between 1954 and 1955. On the other hand, the "Harry Potter" series,
written by J.K. Rowling, began with the publication of "Harry Potter and the
Philosopher's Stone" in 1997.
sub_sentence: The "Lord of the Rings" series, written by J.R.R. Tolkien, was first
published between 1954 and 1955. subject: Lord of the Rings
sub_sentence: On the other hand, the "Harry Potter" series, written by J.K.
Rowling, began with the publication of "Harry Potter and the Philosopher's Stone"
in 1997. subject: "Harry Potter" series

passage: The Mississippi River is significantly longer than the Colorado River.
The Mississippi River is the second-longest river in the United States, stretching
approximately 2,320 miles, while the Colorado River is shorter, with a length of
about 1,450 miles.
sub_sentence: The Mississippi River is significantly longer than the Colorado
River. subject: Mississippi River
sub_sentence: The Mississippi River is the second-longest river in the United
States, stretching approximately 2,320 miles, while the Colorado River is shorter,
with a length of about 1,450 miles. subject: Mississippi River

passage: The United States experiences the most tornadoes globally. The capital of
the United States is Washington, D.C.
sub_sentence: The United States experiences the most tornadoes globally. subject:
United States
sub_sentence: The capital of the United States is Washington, D.C. subject: United
States

passage: Elizabeth Warren served as the senator of Massachusetts. She was also
previously married to Jim Warren.
sub_sentence: Elizabeth Warren served as the senator of Massachusetts. subject:
Elizabeth
sub_sentence: Elizabeth Warren was also previously married to Jim Warren. subject:
Elizabeth

passage: Marcus Elmore Baldwin was an American right-handed professional baseball
pitcher. He played seven seasons in Major League Baseball (MLB).
sub_sentence: Marcus Elmore Baldwin was an American right-handed professional
baseball pitcher. subject: Marcus Elmore Baldwin
sub_sentence: Marcus Elmore Baldwin played seven seasons in Major League Baseball
(MLB). subject: Marcus Elmore Baldwin

passage: Patsy Cline was an American singer. She is considered one of the most
influential vocalists of the 20th century.
sub_sentence: Patsy Cline was an American singer. subject: Patsy Cline
sub_sentence: Patsy Cline is considered one of the most influential vocalists of
the 20th century. subject: Patsy Cline

passage: The Matrix is a series of science fiction films that starred Neo as the
titular character, Keanu Reeves. The series consists of 3 films, released between
1999 and 2003.
sub_sentence: The Matrix is a series of science fiction films that starred Neo as
the titular character, Keanu Reeves. subject: Matrix
sub_sentence: The Matrix series consists of three films, released between 1999 and
2003. subject: Matrix

passage: Chadwick Boseman played the role of T'Challa, the Black Panther, in the
Marvel Cinematic Universe film released in 2018.
sub_sentence: Chadwick Boseman played the role of T'Challa, the Black Panther, in
the Marvel Cinematic Universe film released in 2018. subject: Chadwick Boseman

Please follow the example's pattern(Do not generate the examples and irrelevant
information. sub_sentence and real subject of original passage need in same line)
segment a new passage and generate subject (every sub_sentence use the original
entity name, the 'subject' need follow the 'sub_sentence' in same line don't
generate "\n", don't break new lines) start with "sub_sentence:" and
"subject:",respectively. Do not generate irrelevant information.
passage: {claim}
```

Figure 6: Prompt template of sentence segmentation module.