Towards Improved Multi-Source Attribution for Long-Form Answer Generation

Nilay Patel[*]
UC Santa Cruz
nilay@ucsc.edu

Shivashankar Subramanian
Amazon
ssangu@amazon.com

Siddhant Garg[†]
Meta AI
sidgarg@meta.com

Pratyay Banerjee
Amazon
pratyay@amazon.com

Amita Misra
Amazon
misrami@amazon.com

Abstract

Teaching large language models (LLMs) to generate text with attribution to evidence sources can reduce hallucinations, improve verifiability in question answering systems (QA), and increase reliability of retrieval augmented LLMs. Despite gaining increasing popularity for usage in QA systems and search engines, current LLMs struggle with attribution for long-form responses which require reasoning over multiple evidence sources. To address this, in this paper we aim to improve the attribution capability of LLMs for long-form answer generation to multiple sources, with multiple citations per sentence. However, data for training multi-source attributable QA systems is difficult and expensive to annotate, and therefore scarce. To overcome this challenge, we transform existing QA datasets for this task (MULTIATTR), and empirically demonstrate, on a wide range of attribution benchmark datasets, that fine-tuning on MULTIATTR provides significant improvements over training only on the target QA domain. Lastly, to fill a gap in existing benchmarks, we present a multi-source attribution dataset containing multi-paragraph answers, POLITICITE, based on PolitiFact articles that discuss events closely related to implementation statuses of election promises.

1 Introduction

Large language models (LLMs) are one of the most popular tools for a variety of NLP tasks, including question answering (QA). However, a number of issues have arisen as research in LLMs continues, notably their tendency to hallucinate facts despite otherwise fluent generations (Ji et al., 2023). Several strategies have been proposed to eliminate this shortcoming, and in this work, we focus on retrieval-augmented generation (RAG), and particularly an important adjacent area of attributed QA (Bohnet et al., 2023), i.e., to cite the sources used for writing the LLM's responses. The ability to cite evidence sources accurately can help to improve model development and debugging, explainability of generations, and more importantly the end-user trust on the LLM outputs (Menick et al., 2022; Ras et al., 2021; Nakano et al., 2021).

While the attribution problem in general is non-trivial (Gao et al., 2023b; Yue et al., 2023), the *multi-source attribution* task is particularly challenging as multiple retrieval sources are necessary to support statements in a long response (see Figure 1 for an example). This is exemplified by the citation quality evaluations done by Liu et al. (2023), where they show that even the popular generative search engines (e.g., Bing and perplexity.ai) which support multi-source reasoning, have just 52% of statements to be fully supported by their citations, and only 75% of citations supporting the statement.

So far, existing works (Li et al., 2023a) for this task have either utilized a multi-task learning setup by jointly predicting the response and in-line citations (Menick et al., 2022; Cohen et al., 2022; Glaese et al., 2022) or perform source attribution post generating the responses (Gao et al., 2023b; Yue et al., 2023). One shortcoming of the former is that they predominantly deal with commercial search engines, and are primarily for single-source attribution per statement. The shortcoming of the latter is that while they perform well on single-source attribution, they fail to perform well on attribution prediction in human-annotated *multi-source* attribution datasets. A major challenge for *multi-*

[*] Work done during internship at Amazon
[†] Work completed at Amazon

source attribution is lack of availability of RAG training data with citations, as data-efficient methods like few-shot prompting underperform (Gao et al., 2023b). In this paper, we address these challenges by instead using few-shot prompting to automatically transform existing QA datasets to generate training data for *multi-source attribution*.

Finally, all the existing human-annotated long-form attribution datasets (Kamalloo et al., 2023; Bajaj et al., 2016; Liu et al., 2023) focus primarily on single source attribution, or at most, contain answers that are one paragraph in length. To address this, we present POLITICITE, a multi-paragraph and multi-source attribution QA benchmark dataset sourced from PolitiFact,[1] with 428 long, expert-written articles that analyze to what extent a politician has kept a promise. Politics is not just a domain with rich text-data, but also where text forms a crucial data-source to improve accountability in a society (Subramanian, 2021), and the dataset lies at the core of this purpose. We ask human annotators to retrieve evidence snippets from the cited sources and indicate the degree to which the evidence supports a statement. As POLITICITE provides gold evidences, it allows the study of attribution independent of the retrieval quality. Improving attribution quality is important in the political domain, which is rampant with misinformation, and can directly improve the public trust on LLM-generated text.

We summarize our paper's contributions as follows:

- We highlight the challenges of the *multi-source attribution* prediction task, which has received less community attention for both the multi-task in-line citation generation and post response generation approaches.
- To address the lack of training data for multi-source attribution prediction, we transform existing QA datasets using a simple few-shot prompting approach, and empirically show that models pre-trained on this data can significantly improve performance when they are fine-tuned on the target domain datasets.
- To address the lack of availability of a multi-paragraph multi-source human-annotated evaluation benchmarks for this task, we collect and release the POLITICITE dataset[2].

[1] https://www.politifact.com/

[2] Data can be found at https://github.com/offendo/politicite

2 Related Work

LLMs often hallucinate when answering questions, generating otherwise fluent responses while still containing misinformation. There is significant prior work studying hallucination and improving reliability of generations (Zhang et al., 2023) through methods such as improving quality of data used during pre-training (Li et al., 2023b), adding domain-targeted data during supervised fine-tuning (Elaraby et al., 2023), designing suitable reward models and alignment to human preferences using reinforcement learning (Ouyang et al., 2022) and retrieval-augmented generation (Izacard et al., 2022; Borgeaud et al., 2022).

Retrieval-Augmented Generation In retrieval-augmented generation (RAG) (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022), for a question, relevant text is retrieved (say from web) and used as input to the LLM model. This allows the model provide supported responses, which can reduce hallucinations and improve the response quality. Other work (Gao et al., 2023a) allow retrieval after generation, updating the response based on the new information.

RAG with In-line Attribution In a RAG system, there is no guarantee that generations are completely grounded in the sources. Requiring models to attribute generations to sources can mitigate both concerns, while also improving explainability. To this end, Bohnet et al. (2023) introduce attributed QA as a task, though only include single-source attribution. Currently, joint answer-citation generation (*in-line citation*) models have poor attribution performance (Liu et al., 2023; Gao et al., 2023b; Bohnet et al., 2023), and improving quality has been of interest (Li et al., 2023a).

However, *multi-source attribution* has not seen as much focus. While some datasets were created specifically to require reasoning over multiple sources (Yang et al., 2018; Ho et al., 2020; Qi et al., 2021), their primary aim is in improving answer quality rather than attribution performance.

Attribution Evaluations Liu et al. (2023), through human annotation, find just 52% of statements are fully supported by their citations, indicating a need for improvement. To measure improvement, Rashkin et al. (2022) introduce AIS, a system for annotating attribution in responses. Since AIS is not automatic, using NLI models to approximate

human judgment is a common approach (Honovich et al., 2022; Kamoi et al., 2023; Gao et al., 2023b), and recently fine-tuned attribution models are employed for evaluation (Yue et al., 2023). Gao et al. (2023b) used multi-source datasets in their end-to-end evaluation system ALCE, but do not have gold labels for retrieved sources, making evaluation reliant on NLI models which we find to have gaps on multi-source long-form datasets.

Benchmark datasets There are useful datasets for attributed question answering (Liu et al., 2023; Kamalloo et al., 2023; Yue et al., 2023) and some which could be adapted with additional annotation or automatic heuristics (such as Stelmakh et al. (2023); Fan et al. (2019); Amouyal et al. (2023) as in Gao et al. (2023b)). Of note, the dataset from Liu et al. (2023) (VJ) provides a difficult attribution benchmark which isn't easily solvable by off-the-shelf single-source NLI models. Multi-hop datasets such as HotpotQA (Yang et al., 2018), while useful for training, make poor benchmarks for attributed QA at source-level granularity: simple fine-tuning of Llama-2 (7b) with LoRA (Hu et al., 2021) achieves over 97% citation accuracy on a held out set. Kamalloo et al. (2023) introduced HAGRID, a dataset of LLM generations with annotated sources, which could prove useful for fine-tuning models for attributed QA. We leave incorporating it into our training data for future work.

Both VJ and HAGRID have LLM generated answers, and the datasets used in ALCE have human written responses but no annotated sources. In contrast, our benchmark (POLITICITE) has both high-quality (expert-written) answers and human retrieved evidences, mitigating propagated LLM errors. Furthermore, current benchmarks measure attribution only on single-sentence or few-sentence responses. POLITICITE is the first multi-source attributed QA benchmark with multi-paragraph answers, filling an important gap in this area.

Leveraging LLMs for Data Additionally, LLM-human collaboration has proven useful in generating datasets for a variety of tasks (Yue et al., 2023; Wiegreffe et al., 2022; Liu et al., 2022; Wang et al., 2023; Honovich et al., 2023; Mekala et al., 2022; Bonifacio et al., 2022). We follow Yue et al. (2023) in automatically repurposing datasets, but instead focus on *multi-source attribution* with fine-tuned attribution prediction models. These could be used for both attribution of existing RAG-LLMs and for evaluation as in previous work.

3 Multi-Source Attribution in QA

First, we introduce some terminology we use for the remainder of the paper. A **source** (s_i) is a text segment (boundaries dependent on the chunking strategies employed), obtained using any suitable retrieval approach, which we assume to be true. In the retrieval augmented generation (RAG) setting (Lewis et al., 2020), given a query Q, a generative QA model generates a **response** R which is made up of one or more verifiable statements (r_i, $R = \bigcup_{i \in m} r_i$), grounded on the top-k evidence sources. For *multi-source* attributed QA, we want the individual statements (r_i) to be attributed to one or more sources, i.e., we define a function $\phi(S, r_i) \rightarrow \{0, 1\}$ which takes a set of sources S ($S = \bigcup_{i \in k} s_i$) and a statement r_i and generates 1 for a subset of sources, S_g, which entail r_i, and 0 otherwise. We denote the generated response with in-line attributions as R_c.

Existing Approaches There are two popular approaches of solving the multi-source attributed QA task: (a) *post-gen. citation*: perform citation generation synthesizing the retrieved evidences using the (already generated) response by a RAG model (i.e., (Q, S, R) $\rightarrow R_c$), and (b) *in-line citation generation*: a multi-task learning setup by jointly generating the response and in-line citations, (Q, S) $\rightarrow R_c$. In this work, we assume the existence of a retrieval model to fetch top-k relevant evidence text, and the goal of our system is to get *multi-source* attributed responses to these already retrieved evidences.

The advantage of the post-gen. citation approach is that it can be applied on top of any existing LLM that is used for response generation, and allows to improve the post-gen. model for citation prediction performance independent of the other response quality indicators (e.g., fluency). At the same time, the drawback is that the response generation model cannot use the attribution signals to improve the quality of output, as shown to be feasible by (Yang et al., 2018) for short answers with multi-source reasoning. On the other hand, a joint approach makes it necessary to retrain the model each time we change the citation style or granularity of attributions (e.g., sentence, sub-sentence or a paragraph). In general, a joint model is harder to train than a post-gen. model, as the response is already generated in the latter approach, and the

Sources:
[1] Jonathan Stark (born April 3, 1971) is a former professional tennis player from the United States. During his career he won two Grand Slam doubles titles (the 1994 French Open Men's Doubles and the 1995 Wimbledon Championships Mixed Doubles). Stark reached the World No. 1 doubles ranking in 1994.
[2] Henri Leconte (born 4 July 1963) is a former French professional tennis player. He reached the men's singles final at the French Open in 1988, won the French Open men's doubles title in 1984, and helped France win the Davis Cup in 1991. Leconte's career-high singles ranking was world No. 5.

Question: Which tennis player won more Grand Slam titles, Henri Leconte or Jonathan Stark?
Given (Short) Answer: Jonathon Stark

Paragraph Response: Jonathon Stark won two Grand Slam doubles titles (the 1995 French Open Men's Doubles and the 1995 Wimbledon Championships Mixed Doubles)[1], while Leconte won only one Grand Slam doubles title (the 1984 French Open Men's Doubles) [2]. Stark also reached the World No. 1 doubles ranking in 1994 [1], while Leconte's career-high singles ranking was world No. 5 [2].

Figure 1: An example of HotpotQA → MULTIATTR by expanding short answers to paragraphs with in-line citations.

Original Source: "Whenever Holly (her daughter) said or did anything brattish, I'd just go, 'Is there a problem, "Veruca"? Can I help you?' That "always" snapped her out of it. But I didn't call Barnaby (her son) anything because there was no need; he never gave me any such trouble." Julie Dawn Cole Julie Dawn Cole (born 26 October 1957) is an English actress and politician who has been active for more than 40 years. She began as a child performer in what remains her best-remembered film, 1971's "Willy Wonka & the Chocolate Factory", playing the spoiled Veruca Salt.

Question: Who played veruca in the original willy wonka?
Answer Julie Dawn Cole played Veruca in the original Willy Wonka.

Generated Positive Source: Veruca Salt, the bratty, sassy, and demanding child in the 1971 film "Willy Wonka & the Chocolate Factory," was played by none other than Julie Dawn Cole. Interestingly, Cole's portrayal of Veruca was so memorable that it has become her most recognizable role to date
Generated Negative Source: The 1971 film Willy Wonka & the Chocolate Factory was a musical fantasy film directed by Mel Stuart and released by Paramount Pictures. It was based on the 1964 novel Charlie and the Chocolate Factory by Roald Dahl. The film starred Gene Wilder as Willy Wonka and introduced Peter Ostrum as Charlie Bucket. The film's storyline concerns five children who win a golden ticket to visit Wonka's mysterious and magical chocolate factory.

Figure 2: An example of NQ → MULTIATTR by augmenting additional positive and negative references. The answer "Julie Dawn Cole" appears in the generated positive source while also including other information. The negative source discusses the movie, but importantly doesn't mention information required to answer the question.

former approach may not always result in overall improvements (Menick et al., 2022).

Evaluation For post-gen. citation, we use gold answers for both training and evaluations, which allows us to compare the predicted attributions directly against the gold citations to compute sentence-level *citation accuracy* (exact-match) and *citation F1* following by comparing against gold labels (Liu et al., 2023).

For the joint approach (in-line citation generation), where we predict both the response and in-line citations together, there is no one-to-one alignment of sentences between the prediction and gold annotations. We employ two different evaluation strategies in this case. First, the unsupervised evaluation framework ALCE (Gao et al., 2023b), which assumes access to ground-truth responses but not the citations. Here, generated answer correctness is estimated based on entailment of individual claims from the gold response, and citation quality assessments are NLI-based (checking entailment of generated response given the evidence sources). Second, when we have both ground truth answers and citations, we break the gold response into individual claims and align them to predicted answer sentences, similar to the way Gao et al. (2023b) does for measuring answer correctness. Finally, we compute citation F1 by comparing the gold citations against the aligned sentences' predicted citations (see Figure 3).

4 MULTIATTR Data

In this section, we describe our method for creating MULTIATTR by transforming existing QA datasets with attributions, targeting multi-source attribution in long-form answers. There can be two possible scenarios in multi-source attribution: an answer may require multiple sources to support it or multiple sources can support a claim independently. We consider both scenarios in this work.

Our proposed strategies are: (a) given a question, multiple retrieved sources and a short response as in the case of HotpotQA (Yang et al., 2018), we use few-shot prompting similar to Yue et al. (2023) to expand the short responses into long answers, and (b) given a attributed QA dataset with single-source citations, as found in the NaturalQuestions (Kwiatkowski et al., 2019), we propose a few-shot prompting strategy to augment additional positive and negative evidence sources. In this paper, we demonstrate this approach focusing on two datasets, but this can easily be extended to more QA datasets using a combination of these prompts, to generate large-scale repurposed multi-source attribution

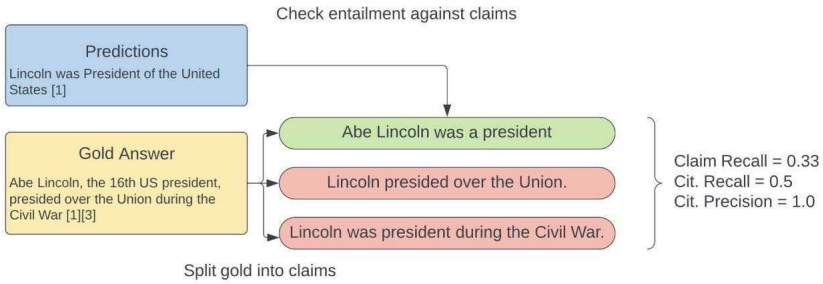


Figure 3: Example for Citation F1 and Claim recall for a single predicted and gold response. In a multi-sentence setting, citation F1 is computed over all the predicted sentences, and a claim recall is 100% if every claim was entailed by at least one sentence in the predicted text.

datasets, which we leave for future work.

Expanding Short to Long Answers Each example in HotpotQA consists of a question, a short answer, and ten source paragraphs, of which two contain information necessary to correctly answer the question. We use a few-shot prompting approach (similar to Yue et al. (2023)) with Llama-2 Chat (13b), to generate longer responses with attributions. See §A.1 for the exact prompt used for generation and example outputs. In addition, the responses often incorporate additional information from the sources to create a more informative answer as shown in Figure 1.

Generating Positive and Negative Sources We use the Natural Questions dataset from (Yue et al., 2023) by selecting answers labeled "attributable" (i.e., fully-grounded on the evidence source). As each example has a single gold source, we augment three additional sources which can be positive or negative references. First, we paraphrase the source paragraph conditioned on the question and gold answer in order to generate another positive source to augment. We instruct the model to use a different style, word choice, and length, while keeping the relevant information. Second, to generate distracting sources, we prompt the model to generate two paragraphs on the same topic as the source, but with no information overlap with the provided statement. This has the effect of generating an on-topic but irrelevant paragraph, which acts as a stronger distractor than randomly selecting an irrelevant reference (see §A.2 for prompts used). Refer to Figure 2 for an example output.

Finally, we use an NLI model (Honovich et al., 2023) to validate and filter positive and negative sources against the gold answer for entailment and non-entailment respectively.

5 Experiments

In this section, we describe our experimental settings, followed by a discussion of the results.

5.1 Datasets

First, we describe the evaluation datasets that we use, along with any details of pre-processing done.
Verifiability Judgments (VJ): We use the human-annotated dataset from Liu et al. (2023). The examples are annotated generations from four generative search engines with attribution (Bing Chat, NeevaAI, perplexity.ai, and YouChat). We use examples which are completely supported by the union of the cited sources. Given evidence snippets are single sentences, to bring uniformity as in a typical RAG setting, we expand evidences to a 400-token snippet by searching the full source document. In the end, we have 1907 training and 494 testing examples (with disjoint query sets).
MSMARCO: MSMARCO (Bajaj et al., 2016) is a large multi-task QA dataset using queries from Bing search, evidence sources retrieved from the web and human generated answers. We evaluate on MSMARCO to ensure our approach still maintains competency in the single-source attribution case. We select examples for which there is a long answer, and this results in having ~ $15k$ samples for training and ~ $12.5k$ for testing.
ALCE: ALCE (Gao et al., 2023b) is an end-to-end question-answering evaluation system, designed to measure fluency, correctness, and citation quality. ALCE combines 1000 examples from each of 3 datasets: ELI5 (Fan et al., 2019), ASQA (Stelmakh et al., 2023), and Qampari (Amouyal et al., 2023). We report results on ELI5 and ASQA (as QAMPARI is for short-form QA). Answers are human-written gold answers, while evidence sources are obtained via retrieval and labeled via exact-match

Dataset	Length	Human response?	Human-annotated citations?
VJ (Liu et al., 2023)	Paragraph	✗	✓
HAGRID (Kamalloo et al., 2023)	Paragraph	✗	✓
HotpotQA (Yang et al., 2018)	Phrase/sentence	✓	✓
ALCE (Gao et al., 2023b)	Paragraph	✗	✗
POLITICITE	Document	✓	✓

Table 1: A comparison of features of several multi-source attribution datasets. POLITICITE is the first long-form multi-source QA dataset with both human expert written answers and human annotated citations, and is the first in a particularly difficult domain (political fact checking).

recall (ASQA) or claim NLI (ELI5).

POLITICITE: Our new benchmark is a multi-paragraph, multi-source attributed QA evaluation benchmark with 428 articles regarding implementation statuses of political promises (sourced from PolitFact). We have gold evidence retrievals from cited sources, i.e., all the evidence sources are relevant to the question. Unlike previous datasets, this has both gold answers and human-annotated multi-source attributions, and requires longer responses with more citations than previous benchmarks, making it challenging for this task. The differences are summarized in Table 1.

5.2 Models

All experimentals are using Llama-2 (7b) (from Hugging Face). Models were trained for 3 epochs with early stopping (based on validation performance). Models have a max sequence length of 2048, total batch size of 64 (8 per GPU × 8 GPUs), and a linear decaying learning rate with 6% warmup ratio, up to a maximum of 2×10^{-4}. We use deterministic greedy decoding for inference. We use LoRA (Hu et al., 2021) with $rank$=8, α=32, $dropout$=0.05.

For post-generation citation models, we add a classifier head with k output labels[3] and train them as multi-label classifiers. In-line citation models are trained as text-to-text models with completion-only fine-tuning. We compare models trained with and without MULTIATTR (based on HotpotQA + NQ), and fine-tune each of them on the target evaluation datasets.

5.3 Experimental Results

Our experiments are organized as follows: we first highlight the challenges of the multi-source attribution setting by showing that some strong baselines, which correlate well with human judgment in single-source benchmarks, don't perform well in the multi-source setting. Following this, we present our results, highlighting the performance gains from MULTIATTR, and an ablation study to demonstrate the impact of varying the training data size of MULTIATTR and the target in-domain datasets. Finally, we discuss model performance on our POLITICITE and remaining challenges.

5.3.1 Single-source Models for Multi-source Attributed QA

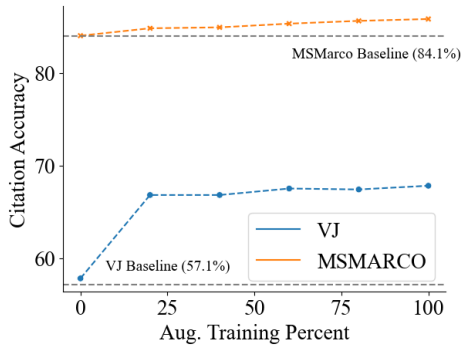
We evaluate two recently popular single-source attribution models:[4] (1) a T5-based NLI model (Honovich et al., 2023) and (2) Alpaca 7B (Taori et al., 2023) trained on single-source attribution evaluation data (Yue et al., 2023) — on VJ and MSMARCO (see results in Table 2). We present results of zero-shot evaluation of these models on the target datasets, and post fine-tuning on the VJ dataset. Despite good correlation with human judgments on single-source datasets, and high zero-shot performance on MSMARCO (also single-source), both models fail to generalize to the multi-source VJ dataset (as highlighted by low absolute performance in Table 2). Even post fine-tuning, the best fine-tuned NLI model achieves only 33.2% F1 score, which is 24.5% points behind Llama-2 (7b) post-gen. attribution prediction model fine-tuned on VJ. This highlights that existing strong models for the single-source attributed QA task do not generalize well to multi-source attributed QA.

5.3.2 Pre-training on MULTIATTR

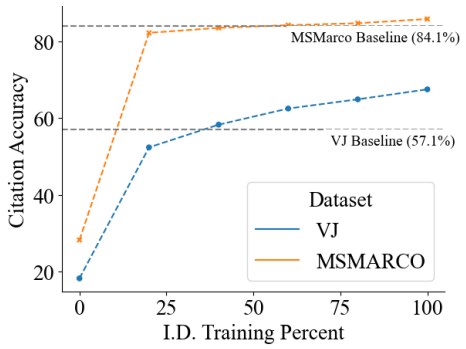
Next, we train models with and without our MULTIATTR data; and compare performance in both

[3] k corresponds to the maximum number of retrieved evidences, and is 5 unless specified otherwise

[4] We provide a single evidence source and statement as input in the intended format. While models may work in a multi-source setting by concatenating evidence sources to be treated as one, this exponentially increases the possible combinations of sources, making it infeasible to evaluate.



(a) Varying the size of MULTIATTR data used for pre-training when using the entire in-domain training data for fine-tuning.



(b) Varying the size of *in-domain* (I.D.) training data post pre-training on the entirety of MULTIATTR .

Figure 4: Performance of post-generation attribution models by varying the size of the pre-training MULTIATTR and in-domain training datasets. MSMarco and VJ Baselines are trained only with in-domain data, without any MULTIATTR for pre-training.

NLI Model	VJ	MSMARCO
T5 xxl 2023	20.6/29.5	65.4/79.7
+ Fine-tuned	10.7/33.2	-
Alpaca 7b 2023	17.0/28.9	59.2/75.7
+ Fine-tuned	17.8/29.0	-
Post-gen. Llama-2 7b	**57.8/65.3**	**84.0/88.4**

Table 2: Performance (citation accuracy/F1) of NLI models (T5 and Alpaca) trained on single-source data as attribution evaluators.

Training data	VJ	MSMARCO
Random	12.0/20.9	8.0/20.5
Fine-tuned	57.8/65.3	84.0/88.4
HotpotQA (short)	8.1/35.3	26.0/68.1
+ Fine-tuned	58.5/62.6	85.0/88.5
HotpotQA (long)	19.2/51.5	28.4/71.0
+ Fine-tuned	**66.8/72.5**	**85.8/88.9**

Table 3: Accuracy/F1 of post-generation citation models trained on HotpotQA with (long) and without (short) our transformation. HotpotQA (short) is the original dataset (Yang et al., 2018).

post-generation and in-line generation settings.

Post-generation Citation For post-gen. models, we vary the size of MULTIATTR, and in-domain training data and observe the change in performance. We find that pre-training on just 20% of MULTIATTR starkly increases accuracy (57.8% → 66.8%), and adding the remaining 80% contributes an additional 1% (66.8% → 67.8%). On MSMARCO, we see smaller improvements (a simpler task), but note that even 20% of the pre-training data makes a 0.8% improvement (Figure 4a).

After pre-training on the entire MULTIATTR, fine-tuning on just 20% of the in-domain data improves accuracy by 34.1% (18.3% → 52.4%). Additional in-domain data continues to improve accuracy, seeing considerable benefits after including all 100% on both datasets, reaching 67.8% and 85.8% on VJ and MSMARCO respectively (Figure 4b). Notably, it only requires 40% and 60% of the in-domain data to surpass the baseline (trained only with the in-domain data, without MULTIATTR). These results highlight the benefits of transforming diverse existing QA datasets to tackle the lack of training data for multi-source attributed QA.

Ablation: Length of answers in MULTIATTR To determine whether transforming HotpotQA by extending answers from short to long-form helps with the downstream performance on the target datasets for post-gen. models, we perform an ablation study (results in Table 3). We see minimal improvement on both VJ and MSMARCO after training on short answers, and a decrease in F1 by 2.7% for VJ. In contrast, training on the extended long-form answers improves performance on VJ by 9% accuracy over the baseline, and by 8.3% accuracy over the model trained on short answers, suggesting that a similar answer shapes is one of the important aspects for positive knowledge transfer between the pre-training and fine-tuning datasets.

In-line Answer+Citation Generation For in-line models, we see a similar trend as in post-generation models (Table 4). That is, while the models pre-trained only on the augmented data alone perform poorly on the evaluation datasets, fine-tuning on the in-domain data post pre-training on MULTIATTR significantly improves performance over the

Model	MAUVE	Cor.	C. Rec	C. Prec
VJ				
VJ	**9.2**	**40.2**	33.5	48.8
MULTIATTR	2.1	29.8	16.5	25.2
MULTIATTR+VJ	8.4	38.2	**35.7**	**49.8**
MSMARCO				
MS	53.3	52.4	**82.6**	83.0
MULTIATTR	45.3	52.6	37.5	34.5
MULTIATTR+MS	**56.0**	**53.3**	82.5	**83.5**
ASQA				
VJ	19.3	25.4	50.4	56.9
MULTIATTR	22.9	26.9	27.7	32.8
MULTIATTR+VJ	**50.9**	**27.5**	**62.1**	**66.2**
ELI5				
VJ	46.1	16.0	46.1	46.8
MULTIATTR	35.7	**18.1**	20.5	27.8
MULTIATTR+VJ	**53.9**	16.9	**49.4**	**49.4**
ELI5 (Few-shot; reported)				
ALCE-L-2 13b	50.0	3.9	3.1	5.3
Vicuna 13b	**58.2**	10.0	15.6	19.6
Chat 13b	34.7	**13.4**	17.3	15.8
Chat 70b	38.6	12.8	**38.3**	**37.9**

Table 4: Fluency (MAUVE (Pillutla et al., 2021)), correctness (claims NLI/exact-match), and citation precision/recall for in-line citation models, measured using ALCE on various datasets. We compare results on models trained only on VJ, only on MULTIATTR, and both. We also report few-shot results of ALCE-Llama-2 models (Gao et al., 2023b).

baseline. On VJ and MSMARCO, we see significant improvements in recall and precision, by 5% and 8% on VJ, and 2% and 1.5% on MSMARCO. On ASQA, the MULTIATTR pre-trained models have higher correctness (+2.1% exact-match recall), higher citation recall and precision (+11.7% and 9.3%), and greatly improved fluency. Similarly, on ELI5, we achieve an improvement of 3.3% and 2.6% recall and precision, while also significantly improving fluency and answer correctness over the baseline. Notably, our fine-tuned models show sizable improvements over few-shot results of very large models such as 70B Llama-2 Chat. We believe that combining pre-training on MULTIATTR with scaling the LLM size should lead to even further performance improvements.

6 POLITICITE

Annotations We collected 1412 truth-o-meter promise articles from PolitiFact, and filtered out articles having fewer than five citations, which yields 428 articles. We split articles by paragraphs, using those with at least one cited source, and select each sentence as a claim (4,839 sentences out of 10,685 total). We employed crowdworkers from Amazon Mechanical Turk to visit the cited sources and retrieve an evidence snippet which supports the claim. Many cited sources are not necessary for support, but rather to provide context or background information on a topic. To address this, we additionally ask annotators to indicate whether the retrieved evidence (1) completely supports, (2) partially supports, or (3) is relevant to but doesn't support the claim. Annotators were allowed to add up to five evidences snippets per claim. Each example was annotated by three annotators with conflicts resolved by a majority vote. Annotations had a BLEU score of 37.2 between annotators, indicating a high agreement in the choice of evidence snippets. Overall statistics of the dataset is given in Table 5. The dataset would be open-sourced[5].

# articles	428
Avg. length of articles	553 words
# claims annotated	10, 685
# claims annotated	4,839
Claims with multi-source citations	1,997
Avg. citations per claim	1.5

Table 5: Statistics about POLITICITE. Average citations are computed over claims with at least one citation.

Why is this dataset challenging? While each claim is typically short (avg. of 24 words), articles are very long, with an average of 553 words (~851 tokens), and up to 1900 words for the longest. Additionally, articles have an average of 7 cited sources, but 11 unique retrieved evidences (some sources have multiple). Both the long length and high citations make for several challenges, including a difficult domain for maintaining fluency, mixing information from sources, and dealing with a large context in models with limited context windows.

Empirical Benchmarking on POLITICITE As there are often many more than 5 sources, we retrieve the top-5 evidences for post-generation method consistent with other datasets we use for training and top-10 for in-line generation models as they struggle with limited context window, using BM25. This reduces the recall upper-bound to 67%, and we leave the full-fledged modeling of large text for attributed QA to future work.

We benchmark results on POLITICITE in Table 6. We observe that both single-source NLI and our post-generation multi-source attribution

[5] Politifact Dataset.

Sources:
[1] https://www.consumerfinance.gov/about-us
The Consumer Financial Protection Bureau is a 21st century agency that implements and enforces Federal financial law and ensures that markets for consumer [...]
[8] https://www.cato.org/people/dan-quan
Dan Quan is an adjunct scholar at the Cato Institute's Center for Monetary and Financial Alternatives, [...]

Question (Promise): What is the status of Biden's promise to *create a public credit reporting agency*?

Answer (Article):
President Joe Biden's campaign proposal to create a new agency within the Consumer Financial Protection Bureau to provide credit scores has not been successful. The Consumer Financial Protection Bureau [1], which implements and enforces federal consumer financial law, told PolitiFact it has "taken significant actions" to help consumers combat coercion through increased guidance to the consumer reporting industry [...]

Figure 5: POLITICITE article discussing Biden's promise to "create a public reporting agency".

Training Data	Post-gen	In-line
T5-xxl NLI (2023)	9.3	-
VJ	9.8	5.5/28.1/4.4
MULTIATTR	9.5	8.1/13.5/2.4
MULTIATTR+ VJ	11.3	6.1/28.7/4.4
ChatGPT	-	15.3/27.9/12.3

Table 6: Results of post-generation (citation F1) and in-line (claim recall/sentence-level citation F1/article-level citation F1) models on POLITICITE. We omit accuracy due to inflated scores across the board, since many sentences have no citations.

prediction models perform poorly on this dataset, with the former achieving 9.3% F1 and the latter at 11.3% F1 at best. Similarly, we find that our best-performing in-line generation models are very poor on this dataset, in large part due to their preference for short answers, at an average of just 92 words (compared to 553 for gold answers). This is apparent in the claim recall and article-level F1, which both require longer answers to improve, and are currently (at best) just 6.1% and 4.4% respectively. Even zero-shot inference using ChatGPT only has 15% claim recall and 12% article-level F1, and achieves comparable sentence-level F1 to our fine-tuned models.

The low performance of best-performing existing models on this dataset highlights the gap area in attribution performance, and positions POLITICITE as a challenging benchmark for this task. Moreover, this emphasizes the need for target-domain training data, and training the attribution models suitably, which we leave for future work. We believe POLITICITE can prove to be a useful resource for benchmarking newer improved models and research in the field of multi-source attribution of long-form answers in the QA community.

7 Conclusion

In this paper we have studied the increasingly important problem of attribution in long-form QA, specifically for multiple sources in a RAG setting. We show that single-source NLI models, despite otherwise being strong baselines, under perform on multi-source attribution QA datasets. We investigate fine-tuning multi-source attribution prediction models, and address the lack of training data by automatically transforming existing QA datasets for this task via few-shot prompting. We show pre-training on the MULTIATTR significant performance improvements on the target domain, in both post-generation attribution prediction and in-line answer-citation generation settings.

Finally, we observe a limitation in existing attributed QA benchmarks, in that they deal with at most one paragraph, and have a limited number of citations. To fill this gap, we present POLITICITE, the first multi-paragraph, multi-source attributed QA dataset, in which expert-written articles from PolitiFact analyze the extent to which a politician has kept a promise. Our best models perform very poorly on POLITICITE, indicating that the length and high-citation counts pose a significant but important challenge. Improving performance on POLITICITE will require both target domain training data and improvements in long-form attributed QA including models' ability to understand paragraph structure and use dependencies between cited sources.

Limitations

Our dataset collection and LLM supervised fine-tuning require access to GPU resources. Our dataset and experiments for multi-source attributed QA is developed only for English language and should extend to other languages.

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.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2387–2392, New York, NY, USA. Association for Computing Machinery.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung ching Chang, Claire Cui, Cosmo Du, Daniel De Freitas Adiwardana, Dehao Chen, Dmitry (Dima) Lepikhin, Ed H. Chi, Erin Hoffman-John, Heng-Tze Cheng, Hongrae Lee, Igor Krivokon, James Qin, Jamie Hall, Joe Fenton, Johnny Soraker, Kathy Meier-Hellstern, Kristen Olson, Lora Mois Aroyo, Maarten Paul Bosma, Marc Joseph Pickett, Marcelo Amorim Menegali, Marian Croak, Mark Díaz, Matthew Lamm, Maxim Krikun, Meredith Ringel Morris, Noam Shazeer, Quoc V. Le, Rachel Bernstein, Ravi Rajakumar, Ray Kurzweil, Romal Thoppilan, Steven Zheng, Taylor Bos, Toju Duke, Tulsee Doshi, Vincent Y. Zhao, Vinodkumar Prabhakaran, Will Rusch, YaGuang Li, Yanping Huang, Yanqi Zhou, Yuanzhong Xu, and Zhifeng Chen. 2022. Lamda: Language models for dialog applications. In *arXiv*.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models.

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.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. Rarr: Researching and revising what language models say, using language models.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023a. A survey of large language models attribution. https://github.com/HITsz-TMG/awesome-llm-attributions.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks are all you need ii: phi-1.5 technical report.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. ArXiv:2304.09848.

Dheeraj Mekala, Tu Vu, Timo Schick, and Jingbo Shang. 2022. Leveraging QA datasets to improve generative data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9737–9750, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes.

Reiichiro Nakano, Jacob Hilton, S. Arun Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers.

Peng Qi, Haejun Lee, Oghenetegiri "TG" Sido, and Christopher D. Manning. 2021. Answering open-domain questions of varying reasoning steps from text. In *Empirical Methods for Natural Language Processing (EMNLP)*.

Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. 2021. Explainable deep learning: A field guide for the uninitiated.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2022. Measuring attribution in natural language generation models.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2023. Asqa: Factoid questions meet long-form answers.

Shivashankar Subramanian. 2021. *Natural Language Processing for Improving Transparency in Representative Democracy*. Ph.D. thesis, UNIVERSITY OF MELBOURNE.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

A Appendix

A.1 HotpotQA prompt

You are given two sources with related information, a question, and a correct answer. Your job is to write a 1-3 sentence response which answers the question and adds relevant information from the sources, while also citing your sources. Do NOT just copy words from the source in your response. Instead, use the information and creatively paraphrase the text.

Sources:
{sources}

Question/Answer:
{question} {answer}

Response:
{response}"""

Figure 6: Prompt template used for augmenting HotpotQA to long responses. 3 examples with this template (including instructions) are used for few-shot prompting. To generate output, {response} is left blank.

A.2 NaturalQuestions prompt

Prompt to augment a positive source is given in Figure 7, and the prompt to augment negative sources is given in Figure 8.

Instructions:
You are given a reference paragraph and a statement which contains information from the reference. Your job is to rewrite the reference using different style, word choice, and length, without removing the information found in the statement. For example:

Reference
Figure skating at the 2018 Winter Olympics – Men's singles The men's single figure skating competition of the 2018 Winter Olympics was held on 16 and 17 February 2018 at the Gangneung Ice Arena in Gangneung, South Korea. The short program was held on 16 February and the free skating was held on 17 February. This medal event was the 1000th medal event in the history of the Winter Olympic Games. With his victory at the 2018 Winter Olympics, Yuzuru Hanyu became the first male figure skater to win two consecutive gold medals after Dick Button, who did so in

Statement
Yuzuru Hanyu won men's figure skating at the Olympics.

Output
Olympic figure skating had some surprises for us in 2018. For the first time since Dick Button in 1952, the reigning men's figure skating gold medalist, Yuzuru Hanyu of Japan, defended his title to take home a second consecutive gold medal.

Reference
{reference}

Statement
{statement}
Output:

Figure 7: Prompt given to NaturalQuestions to paraphrase the gold source into a positive source. We only use the one example, and find it is sufficient.

Instructions
You are given a reference paragraph and a statement. Your job is to generate two paragraphs on the same topic as the reference and statement without repeating any information from either. Be absolutely certain that statement _cannot_ be inferred by the generated paragraphs. For example:

Reference
Figure skating at the 2018 Winter Olympics – Men's singles The men's single figure skating competition of the 2018 Winter Olympics was held on 16 and 17 February 2018 at the Gangneung Ice Arena in Gangneung, South Korea. The short program was held on 16 February and the free skating was held on 17 February. This medal event was the 1000th medal event in the history of the Winter Olympic Games. With his victory at the 2018 Winter Olympics, Yuzuru Hanyu became the first male figure skater to win two consecutive gold medals after Dick Button, who did so i

Statement
Yuzuru Hanyu won men's figure skating at the Olympics.

Output 1
Figure skating was first contested as an Olympic sport at the 1908 Summer Olympics, in London, United Kingdom. As this traditional winter sport could be conducted indoors, the International Olympic Committee (IOC) approved its inclusion in the Summer Olympics program. It was featured a second time at the Antwerp Games, after which it was permanently transferred to the program of the Winter Olympic Games, first held in 1924 in Chamonix, France.

Output 2
Ice dance competitions formerly consisted of three phases: one or more compulsory dances; an original dance to a ballroom rhythm that was designated annually; and a free dance to music of the skaters' own choice. Beginning in the 2010–11 season, the compulsory and original dances were merged into the short dance, which itself was renamed the rhythm dance in June 2018, before the 2018–19 season.

Reference
{reference}

Statement
{statement}

Output 1

Figure 8: Prompt given to NaturalQuestions to generate additional negative sources. We only use the one example, and find it is sufficient.