Post-Hoc Answer Attribution for Grounded and Trustworthy Long Document Comprehension: Task, Insights, and Challenges

Abhilasha Sancheti[†][*], **Koustava Goswami**[‡], **Balaji Vasan Srinivasan**[‡]
[†]University of Maryland, College Park [‡]Adobe Research
sancheti@umd.edu, {koustavag,balsirini}@adobe.com

Abstract

Attributing answer text to its source document for information-seeking questions is crucial for building trustworthy, reliable, and accountable systems. We formulate a new task of post-hoc answer attribution for long document comprehension (LDC). Owing to the lack of long-form abstractive and information-seeking LDC datasets, we refactor existing datasets to assess the strengths and weaknesses of existing retrieval-based and proposed answer decomposition and textual entailment-based optimal selection attribution systems for this task. We throw light on the limitations of existing datasets and the need for datasets to assess the actual performance of systems on this task.

1 Introduction

Users now benefit from the help of automatic question-answering (QA) systems on a day-to-day basis when faced with an information need. Such systems are integrated into search engines (*e.g.*, BingAI[1]) and digital assistants (*e.g.*, ChatGPT). However, such systems are prone to generating answers lacking sufficient grounding to knowledge sources (Dziri et al., 2022; Ji et al., 2023), leading to the risks of misinformation and hallucination (Metzler et al., 2021; Shah and Bender, 2022; Huo et al., 2023). Therefore, attributing the generated answers to the respective sources is crucial for building trustworthy, reliable, verifiable, and accountable systems (Bohnet et al., 2022; Huang and Chang, 2023; Rashkin et al., 2023; Yue et al., 2023); by allowing users to verify outputs.

Input

Question: When does the next assasins creed come out?
Document: [1] Ubisoft has announced that its next Assassin's Creed game will be revealed in September 2022.
[2] Ubisoft shared the first trailer for the game on Saturday.
[3] Assassin's Creed Mirage, the next entry in Ubisoft's long-running action-adventure series, will arrive in 2023.
[4] The publisher announced the release date today during its Ubisoft Forward event. . . .
Answer: The next Assassin's Creed game, Assassin's Creed Mirage, will arrive in 2023 according to Ubisoft's announcement during its Ubisoft Forward event. It will be released for Xbox . . . The game will be revealed in September 2022.

Output

Attributed answer: The next Assassin's Creed game, Assassin's Creed Mirage, . . . Ubisoft's announcement during its Ubisoft Forward event [3,4] . . . The game will be revealed in September 2022 [1].

Table 1: An example taken from reformulated verifiability dataset (Liu et al., 2023) that includes a question, a document,[2] and an answer as inputs, and the document-grounded attributions for each sentence (some may not have any attribution) in the answer as output.

Existing works mainly consider generating attributed text in open-ended settings. These attributions are generated along with the answers either one per answer paragraph (Bohnet et al., 2022; Hu et al., 2024) or per answer sentence (Gao et al., 2023a,b; Malaviya et al., 2023). Evidence retrieval is used to select an answer in reading comprehension setting (Wang et al., 2019; Yadav et al., 2020; Cui et al., 2022) for short and extractive answers. Attribution becomes challenging when answers are abstractive such that each sentence could be composed of multiple sentences in the source document, requiring more sophisticated approaches. To address this gap, we aim to identify fine-grained attributions (*i.e.*, sentences grounded in a provided long document) for each sentence (unlike paragraph or article) of a long-form abstractive answer to an information-seeking question asked over a user-provided document (closed-domain). Such fine-grained attributions can lead to more trustworthy, reliable, and accountable systems. Specifically, we propose a new task (Table 1) of **post-hoc answer attribution for long document comprehension** wherein the input to a system is a **(question, answer, document)** triplet, and output is an **attributed answer** consisting of pointers to sentences in the document that provide supporting evi-

[*] This work was done when the author was at Adobe.

[1] https://www.microsoft.com/en-gb/bing?form=MW00X7

[2] A subset of sentences is shown due to space constraints.

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 49 - 57
June 20-21, 2024 ©2024 Association for Computational Linguistics

dence for each sentence in the answer.

Building systems for this task is challenging due to the unavailability of appropriate datasets as answers in existing information-seeking reading comprehension datasets (*e.g.*, Dasigi et al., 2021) are short and extractive. Moreover, obtaining attribution annotations is cognitively demanding, labor-intensive, and expensive as it requires expertise (Kamalloo et al., 2023). Thus, we (a) propose to reformulate existing datasets curated for evaluating citation verifiability in generative search engines (Liu et al., 2023), and generating attributed explanations in generative information-seeking systems (Kamalloo et al., 2023), and (b) assess the feasibility of using existing textual entailment models by proposing `ADiOSAA`– consisting of an answer decomposer and a textual entailment-based attributor that uses an optimal selection strategy to find attributions for each sentence of an answer.

This work **contributes** the following: (1) introduces the task of **post-hoc answer attribution for LDC** for building trustworthy, verifiable, reliable, and accountable QA systems (§2); (2) reformulates existing datasets for this task, owing to the lack of availability of long-form abstractive reading comprehension datasets (§2), and (3) assesses the strengths and weaknesses of existing retrieval-based systems, and proposed answer decomposition and textual entailment-based optimal selection system, `ADiOSAA` (§3), by adopting information retrieval measures (§4).

2 Adapting existing datasets for our task

Task Definition We formalize the task of post-hoc answer attribution for long document comprehension as: given a query Q, a set of sentences $S = s_1, \ldots, s_n$ from document D (namely, source sentences), and an answer (either generated from a system or ground-truth) to query Q, the goal is to identify supporting sentences (namely, attributions) $s_i \in S$ for each answer sentence $a_i \in A = a_1, \ldots, a_m$ (may be attributed to multiple source sentences or none). Since there are no datasets that match the needs of our task, we propose to reformulate the Citation Verifiability dataset (Liu et al., 2023) and Hagrid dataset (Kamalloo et al., 2023) for the proposed task.

Split	**Size**	**Avg. No. of source sentences**	**Avg. No. of attributions per sentence**	**Avg. No. of sentences per answer**	**Avg. No. of answers per question**
		Verifiability/Hagrid			
Train	1138/1922	128.58/2.82	1.45/1.26	2.11/1.63	2.63/1.67
Dev	146/716	141.68/2.83	1.49/1.40	2.18/1.71	2.56/1.84
Test	136/–	130.03/–	1.60/–	2.13/–	2.75/–

Table 2: Dataset statistics. No test set in Hagrid.

Reformulation Citation Verifiability Dataset Citation verifiability dataset (Liu et al., 2023) consists of questions from NaturalQuestions (Kwiatkowski et al., 2019) and ELI5 (Fan et al., 2019) and answers are generated from different generative search engines; Bing Chat, NeevaAI, perplexity.ai, and YouChat. These answers are embedded with inline citations pointing to the web pages. Human annotators were shown a question and a verification-worthy sentence from the generated answer with its corresponding generated citations and were asked to judge if the citations *fully, partially, or do not support* the sentence. For sentences that are *fully* supported, annotators also provide sentences on the webpage that support the answer sentence. In this open-domian setup, the citations in an answer may belong to multiple web pages. To obtain a pseudo document for a question, we focus on questions anchored to a given document by combining fully supported web page contents cited for sentences. Hence, we have a corpus with questions, answers, a document to which questions are grounded, and ground truth attributions for sentences in an answer.

Reformulating Hagrid Dataset Kamalloo et al. (2023) introduced Hagrid which is constructed based on human and LLM collaboration by first automatically collecting attributed answers (for information-seeking questions in MIRACL (Zhang et al., 2022) dataset) that follow an inline citation style using GPT-3.5. Then, asking human annotators to evaluate the LLM answers based on informativeness and attributability. We establish benchmarks for this dataset by considering the LLM-generated answers to be the gold-answers required as input (as opposed to the task formulation of Hagrid, wherein output is an attributed answer), attributability annotations as attributions for sentences in an answer, and labeled relevant passages as the document. We provide dataset statistics in Table 2.

3 Answer Decomposition and Optimal Selection for Answer Attribution

We propose an **A**nswer **D**ecompos**i**tion and **O**ptimal **S**election **A**nswer **A**ttribution system for

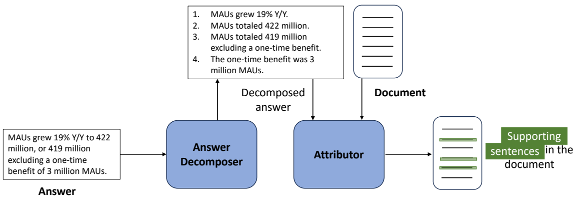
1. MAUs grew 19% Y/Y.
2. MAUs totaled 422 million.
3. MAUs totaled 419 million excluding a one-time benefit.
4. The one-time benefit was 3 million MAUs.
Decomposed answer
Document
MAUs grew 19% Y/Y to 422 million, or 419 million excluding a one-time benefit of 3 million MAUs.
Answer
Answer Decomposer
Attributor
Supporting sentences in the document

Figure 1: Overview of proposed answer attribution system, `ADiOSAA`. The **answer decomposer** breaks the given answer into *information units*, and the **attributor** finds the supporting sentences as attributions for each *information unit* in the answer.

the introduced task. `ADiOSAA` consists of two components (Figure 1): (1) An **answer decomposer** to break each sentence of an answer into one or more *information units* (Nenkova and Passonneau, 2004; Stanovsky et al., 2018; Ernst et al., 2021) as we believe that an answer sentence is composed of information from multiple sentences in the input document. (2) An **attributor** to find supporting sentences in the document for a given *information unit* in the answer sentence.

Answer Decomposer We prompt ("Please breakdown the following sentence into independent facts: ..") ChatGPT (OpenAI, 2023) to decompose the given answer into its information units, following Min et al. (2023) who found such decompositions to be effective and close to human. This decomposition resembles past frameworks derived from OpenIE (Stanovsky et al., 2018; Ernst et al., 2021) or Pyramid (Nenkova and Passonneau, 2004; Shapira et al., 2019), but avoids relying on annotated data and achieves greater flexibility by using ChatGPT. Such decomposition to information units has been successfully used for claim-verification (Kamoi et al., 2023) and propositional semantic representations (Chen et al., 2023).

Attributor Once the answer is decomposed into its information units, each unit needs to be mapped to sentences in the input document to provide the desired attributions. We pose this task of finding supporting sentences in the document for a given information unit as a textual entailment task. Textual entailment is the task of identifying if a given premise ($\mathcal{P}$) entails or does not entail the given hypothesis ($\mathcal{H}$). For our purpose, we consider sentence(s) in the document as the premise and an information unit as the hypothesis. We use

Algorithm 1 Optimal Selection Algorithm

```
Inputs: Information unit ($iu$), $D = d_1, d_2 \ldots d_n$, $Attr(\mathcal{P}, \mathcal{H})$, $\delta$
Outputs: $L$ = A list of supporting sentences in $D$ which together attribute $iu$
$L \leftarrow []$, RS $\leftarrow D$, prev_score $\leftarrow -1$    // RS: remaining sentences; Initialization
while RS is not empty do
    curr_score $\leftarrow \max_{d_i \in \text{RS}}$ Attr($L + d_i, iu$)
    $d_{max} \leftarrow \arg\max_{d_i \in \text{RS}}$ Attr($L + d_i, iu$)
    if curr_score > prev_score + $\delta$ then
        $L$ += $d_{max}$
        RS $-=$ $d_{max}$
        prev_score = curr_score
    else
        break
    end if
end while
```

RoBERTa-L (Liu et al., 2019) pretrained[3] on DocNLI (Yin et al., 2021) dataset (contains paragraph-level (premise, hypothesis) pairs, see §B for more details) as the entailment model (attributor) to predict if the given information unit can be inferred from the given sentence(s) from the document.

Optimal Selection An answer sentence could be attributed to multiple sentences in the provided document when: (a) the same information is available in the document at multiple places, and (b) pieces of information in the answer sentence is available in different parts of the document. (a) can be solved by considering the top k (premise hypothesis) pairs where the premise is the sentence from the document and the hypothesis is the sentence or information unit of the answer. To solve (b), it is required to check if a sentence or information unit of an answer can be entailed from a combination of sentences in the document as a premise. However, this becomes computationally expensive; for a document consisting of N sentences, there will be 2^N combinations. To address this issue, we propose an optimal selection approach that greedily selects sentences from the document that has the maximum probability of entailment as described in Algorithm 1. Attr($\mathcal{P}$,$\mathcal{H}$) refers to DocNLI-based attributor which takes sentences from the input document and the information unit (or sentence in an answer) and outputs the probability of entailment of $\mathcal{H}$ from $\mathcal{P}$. For each information unit in a sentence, Algorithm 1 iteratively selects a sentence from the set of remaining source sentences that maximizes the probability of entailment until the entailment score keeps increas-

[3] We use the official code and trained model available at `https://github.com/salesforce/DocNLI`.

Model	Verifiability			Hagrid		
	(P/R/F1)@1	(P/R/F1)@2	(P/R/F1)@4	(P/R/F1)@1	(P/R/F1)@2	(P/R/F1)@4
BM25	0.669/0.529/0.567	0.443/0.648/0.499	0.270/0.722/0.369	0.815/0.686/0.722	0.740/0.919/0.788	0.678/0.990/0.760
GTR	0.656/0.511/0.550	0.432/0.623/0.483	0.270/0.723/0.371	0.899/0.768/0.804	0.744/0.918/0.790	0.677/0.987/0.759
MonoT5	**0.698**/**0.552**/**0.593**	0.466/**0.675**/**0.522**	0.284/**0.757**/0.389	**0.962**/**0.827**/**0.864**	0.763/**0.946**/**0.811**	0.680/**0.993**/0.762
ADiOSAA	0.545/0.428/0.459	**0.484**/0.546/0.487	**0.476**/0.604/**0.499**	0.856/0.734/0.768	0.848/0.810/0.799	0.848/0.817/**0.801**
ADiOSAA - D	0.473/0.388/0.412	0.445/0.418/0.412	0.442/0.418/0.411	0.869/0.749/0.782	**0.861**/0.758/0.783	**0.861**/0.758/0.783
ADiOSAA - OS	0.375/0.295/0.317	0.280/0.333/0.284	0.256/0.360/0.276	0.793/0.679/0.710	0.745/0.783/0.736	0.743/0.830/0.752
ADiOSAA - D - OS	0.269/0.234/0.243	0.269/0.234/0.243	0.269/0.234/0.243	0.567/0.466/0.494	0.567/0.466/0.494	0.567/0.466/0.494

Table 3: Evaluation results: ADiOSAA systems use top 150 source sentences (see Table 6 in Appendix for results with GTR, MonoT5, and all the source sentences) retrieved using BM25 for the Verifiability dataset. D denotes Answer Decomposer, and OS refers to Optimal Selection.

ing above a threshold δ as compared to that in the previous iteration.

We reorder the attributions for each information unit based on their score and deduplicate (as different information units may be attributed to the same source sentence) them to obtain the predicted attributions for each sentence of an answer.

4 Evaluation

As answer sentence attribution to sentences in the source document could also be considered as an information retrieval task, we benchmark the performance of a range of retrieval-based systems: **(1) BM25 (sparse)**, **(2) GTR (dense)**, and **(3) MonoT5**, considering an answer sentence as the query, and the sentences/passages from the input document as the document (refer to §A). Because our task assumes the answer as an input, inline attribution-based systems like vanilla LLM prompting (Tay et al., 2022; Weller et al., 2023) and retrieve-and-read-based systems (Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022) do not fit here. For the Verifiability dataset, ADiOSAA system and its variants use top 150 retrieved sentences as the source sentences. As Hagrid has only 2.83 passages per question in total, we consider all the passages as the source sentences. Additionally, we perform ablation experiments to demonstrate the importance of decomposition and optimal selection in ADiOSAA in the following ways.

ADiOSAA - D considers an answer sentence as the information unit instead of decomposing it. This system establishes the importance of the answer decomposer in ADiOSAA.

ADiOSAA - OS decomposes each answer sentence into its information units, and then ranks source sentences based on their entailment probabilities from the Attr($\mathcal{P}$,$\mathcal{H}$) for each information unit. To obtain attributions for each sentence of the answer, it deduplicates and reorders the attributions for all the information units of the sentence based on the entailment probabilities.

ADiOSAA - D - OS neither uses the answer decomposer or the optimal selection algorithm rather for each sentence in the answer, it ranks source sentences based on their entailment probabilities from the Attr($\mathcal{P}$,$\mathcal{H}$). This system demonstrates the effectiveness of both the components in ADiOSSA.

Evaluation Measures We report precision (P), recall (R), and F1@k $\in \{1, 2, 4\}$ predicted attributions per sentence of an answer[4] for the test set of Verifiability dataset and development set of the Hagrid dataset (as no test set is available). We tune the threshold for attributor's entailment probability (=0.5) and δ (=0.3) in Algorithm 1 based on the Verifiability development set.

5 Results and Discussion

While MonoT5-based retrieval system outperforms (Table 3) others for the top-1 prediction, ADiOSAA variants attain the highest precision when top 2 or 4 predictions are considered. Having a high precision for top 2 or 4 predictions is important as the mean number of attributions per sentence > 1 (see Table 2) and with the increase in the number of predictions, recall may increase or remain the same however, precision may increase, decrease, or stay the same. ADiOSAA variants retain higher precision (as compared to retrieval-based systems) even with the increase in the number of predictions, indicating that retrieval-based systems are good at retrieving one attribution correctly but fail for the second (or more) one compared to our systems. This shows that our systems capture abstractive and compositional attributions more correctly. Optimal selection results in a significant improvement. Higher gains due to optimal selection under no decomposition (difference between ADiOSAA-D

[4] We filter out the instances where answer sentences were extracted directly from the documents.

and ADiOSAA-D-OS) than under decomposition (difference between ADiOSAA and ADiOSAA-OS) shows that the answer sentence is composed of multiple document sentences which are better captured with optimal selection. However, under decomposition, it is more likely that now the decomposed units could be attributed to a single sentence in the document. Decomposition also helps in better predictions (compare ADiOSAA-OS with ADiOSAA-D-OS) showing that compositional answers have multiple attributions to different sentences in the input document. Further, due to a small number of source sentences (avg. 2.83) in Hagrid, the precision and recall values are higher as compared to that in the Verifiability dataset.

Good performance of retrieval-based systems indicate that the existing datasets are less abstractive for long-form comprehension, suggesting the need for research in creating more challenging datasets to foster the development of trustworthy, reliable, and accountable systems that can be used in real-world information-seeking scenarios.

Quality of Decompositions Prior works have used ChatGPT for decomposing facts (Min et al., 2023) or claims (Kamoi et al., 2023) and have shown it to perform reasonably well. We manually examine a subset of decompositions and find that the decomposer might sometimes over-decompose a simple sentence, or generate hallucinated information units (see Table 4 in the appendix for examples). We leave a careful analysis of error categories, and ways to mitigate hallucinations and over-decompositions for future work.

6 Conclusion

We introduce a task of post-hoc answer attribution for long document comprehension, reformulate existing datasets, and asses the feasibility of existing textual entailment and retrieval-based systems in performing this task. Evaluation shows that retrieval-based systems are good at top one prediction however, our proposed answer decomposition and textual entailment-based optimal selection system, ADiOSAA, performs better when more than one predictions are considered. This further indicates the need for highly abstractive long-form reading comprehension datasets that can foster the development and evaluation of more sophisticated attribution systems.

7 Limitations

We note the following limitations of our work. (1) The decompositions are obtained without taking into consideration the source document which might result in unnecessary answer decompositions. This issue can be resolved if the information units are explicitly constrained in the input document, and (2) ADiOSAA is a post-hoc inference time attribution system which uses off-the-shelf trained model, DocNLI. However, future work may consider developing supervised systems for performing the task on the verifiability dataset, and building end-to-end systems where decomposition and optimal selection may happen in an interactive manner. (3) We acknowledge the performance dependence of ADiOSAA on the Attributor. Further investigation into the impact of NLI model's performance on the final results is an avenue for future work.

References

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

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

Sihao Chen, Hongming Zhang, Tong Chen, Ben Zhou, Wenhao Yu, Dian Yu, Baolin Peng, Hongwei Wang, Dan Roth, and Dong Yu. 2023. Sub-sentence encoder: Contrastive learning of propositional semantic representations. *arXiv preprint arXiv:2311.04335*.

Yiming Cui, Ting Liu, Wanxiang Che, Zhigang Chen, and Shijin Wang. 2022. Expmrc: explainability evaluation for machine reading comprehension. *Heliyon*, 8(4).

Curation. 2020. Curation. 2020. curation corpus base.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.

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 Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. 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, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Sheng Bi, Tongtong Wu, and Jeff Z Pan. 2024. Benchmarking large language models in complex question answering attribution using knowledge graphs. *arXiv preprint arXiv:2401.14640*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*.

Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. 2023. Retrieving supporting evidence for llms generated answers. *arXiv preprint arXiv:2306.13781*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

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 Computing Surveys*, 55(12):1–38.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. *arXiv preprint arXiv:2303.01432*.

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.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. In *Acm sigir forum*, volume 55, pages 1–27. ACM New York, NY, USA.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

OpenAI. 2023. OpenAI. (2023). ChatGPT.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, pages 1–66.

Chirag Shah and Emily M Bender. 2022. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 221–232.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019. Evidence sentence extraction for machine reading comprehension. *arXiv preprint arXiv:1902.08852*.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. " according to..." prompting language models improves quoting from pre-training data. *arXiv preprint arXiv:2305.13252*.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4514–4525, Online. Association for Computational Linguistics.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages. *arXiv preprint arXiv:2210.09984*.

Appendix

A Baseline Models

- **BM25 (sparse)** is a classical bag-of-words based sparse retrieval method that relies on lexical overlap, term frequency heuristics, inverse document frequency and document length for retrieval relevant passages given a query.

- **GTR (dense)** is a dense retrieval method that embeds both documents and queries into low-dimensional representations using T5-based (Raffel et al., 2020) dual encoders, with one of the encoders tailored to the queries and the other to the documents.

- **MonoT5** is a T5-based model that takes a query and a document, and outputs the probability of relevance of document with respect to the query. The documents are ranked based on this probability.

B Entailment model DocNLI

We have used RoBERTa-L model trained on DocNLI dataset as our go-to entailment model. DocNLI contains an array of reformulated versions of existing datasets (adversarial NLI (ANLI) (Nie et al., 2019), the question answering benchmark SQuAD (Rajpurkar et al., 2016) and three summarization benchmarks (CNN/DailyMail (Nallapati et al., 2016), DUC2001[5], and Curation (Curation, 2020))) by transforming various summarization and question answering datasets into natural language inference form to ensure that the premise and hypothesis are paragraph-level long and that the dataset does not contain any artifacts such as hypothesis length bias, direct overlap between premise and hypothesis. Table 5 reports the accuracy results of DocNLI on various NLI datasets.

C Additional Results

We present results on verifiability dataset when top 150 sentences retrieved using GTR and MonoT5 are used by `ADiOSAA` and when all the sentences in a document are used in Table 6. In all the cases, `ADiOSAA` outperforms all the ablations - removing decompose, optimal selection or both. This shows that each of the components of the proposed approach is important for the attribution task.

C.1 Implementation Details

The RoBERTa-L model contains 355 million parameters. We use off-the-shelf model so no training is required. We only perform inference on one NVIDIA T4 16GB GPU machine. All the reported scores are from one run of the model inference. Hyperparameter tuning is done for δ and entailment probability threshold using the development set of Verifiability dataset.

[5] `https://www-nlpir.nist.gov/projects/duc/guidelines/2001.html`

S.No.	Input	Output Decompositions
	Correct Decompositions	
1.	Based on the given contexts, the University Library of the Library of Congress in Washington D.C. is the biggest library in the world, with more than 167 million items including books, printed materials, recordings, photographs, maps, sheet music, and manuscripts.	The University Library of the Library of Congress is in Washington D.C. The University Library of the Library of Congress is the biggest library in the world. The University Library of the Library of Congress has more than 167 million items. The items include books, printed materials, recordings, photographs, maps, sheet music, and manuscripts.
2.	The Islamic Golden Age is traditionally dated from the 8th century to the 14th century.	The Islamic Golden Age is traditionally dated. The Islamic Golden Age is traditionally dated from the 8th century. The Islamic Golden Age is traditionally dated to the 14th century.
	Erroneous Decompositions	
1.	Ronnie Van Zant was born in Jacksonville, Florida.	Ronnie Van Zant was born. Ronnie Van Zant was born in Jacksonville, Florida
2.	The Beer Store, also known as Brewers Retail Inc, was founded in 1927.	The Beer Store is also known as Brewers Retail Inc. The Beer Store was founded in 1927. The Beer Store was founded in Ontario, Canada

Table 4: Sample outputs from the Answer Decomposer. 1. shows over-decomposition, and 2. shows hallucination error under Erroneous Decompositions.

FEVER	MCTest (v160)	MCTest (v500)	SciTail	MNLI
88.84	90.00	85.83	78.17	91.13

Table 5: Accuracy of DocNLI (used as the Attributor in our work) model on various NLI datasets. We report the numbers as-is from Yin et al. (2021).

Model	Top 1			Top 2			Top 4		
	P	R	F1	P	R	F1	P	R	F1
All + ADIOSAA	**0.537**	**0.422**	**0.452**	**0.479**	**0.540**	**0.482**	**0.471**	**0.598**	**0.494**
All + ADIOSAA - Decomposer	0.462	0.381	0.404	0.435	0.408	0.402	0.433	0.408	0.401
All + ADIOSAA - Optimal Selection	0.368	0.289	0.311	0.272	0.327	0.279	0.250	0.353	0.270
All + ADIOSAA - Decomposer - Optimal Selection	0.262	0.226	0.236	0.262	0.226	0.236	0.262	0.226	0.236
GTR + ADIOSAA	**0.538**	**0.423**	**0.453**	**0.479**	**0.541**	**0.483**	**0.471**	**0.598**	**0.494**
GTR + ADIOSAA - Decomposer	0.463	0.382	0.405	0.435	0.409	0.403	0.433	0.409	0.402
GTR + ADIOSAA - Optimal Selection	0.372	0.294	0.315	0.275	0.332	0.282	0.252	0.358	0.273
GTR + ADIOSAA - Decomposer - Optimal Selection	0.265	0.229	0.238	0.265	0.229	0.238	0.265	0.229	0.238
MonoT5 + ADIOSAA	**0.537**	**0.422**	**0.452**	**0.479**	**0.540**	**0.482**	**0.471**	**0.598**	**0.494**
MonoT5 + ADIOSAA - Decomposer	0.467	0.385	0.408	0.439	0.412	0.407	0.437	0.413	0.406
MonoT5 + ADIOSAA - Optimal Selection	0.371	0.292	0.314	0.274	0.330	0.281	0.251	0.356	0.272
MonoT5 + ADIOSAA - Decomposer - Optimal Selection	0.265	0.229	0.238	0.265	0.229	0.238	0.265	0.229	0.238

Table 6: Evaluation results with GTR, MonoT5 and all sentences for Verifiability dataset.