Document Attribution: Examining Citation Relationships using Large Language Models

Vipula Rawte[1]*, Ryan Rossi[2], Franck Dernoncourt[2], Nedim Lipka[2]

[1]Adobe Inc. [2]Adobe Research

vrawte@adobe.com

Abstract

As Large Language Models (LLMs) are increasingly applied to **document-based tasks** - such as document summarization, question answering, and information extraction - where user requirements focus on retrieving information from provided documents rather than relying on the model's parametric knowledge, ensuring the trustworthiness and interpretability of these systems has become a critical concern. A central approach to addressing this challenge is *attribution*, which involves tracing the generated outputs back to their source documents. However, since LLMs can produce inaccurate or imprecise responses, it is crucial to assess the reliability of these citations.

To tackle this, our work proposes two techniques. (1) A **zero-shot** approach that frames attribution as a straightforward textual entailment task. Our method using `flan-ul2` demonstrates an improvement of 0.27% and 2.4% over the best baseline of ID and OOD sets of AttributionBench Li et al. (2024), respectively. (2) We also explore the role of the **attention mechanism** in enhancing the attribution process. Using a smaller LLM, `flan-t5-small`, the F1 scores outperform the baseline across almost all layers except layer 4 and layers 8 through 11.

1 Introduction

Attribution in Large Language Models refers to tracing the origins of information embedded in the model's outputs. This involves identifying the specific datasets, documents, or text segments contributing to the generated response. Attribution is essential for verifying the information's provenance, ensuring the generated content's accuracy and reliability, and addressing concerns regarding plagiarism, accountability, and transparency in AI systems. Attribution methods typically involve mapping responses to the relevant data sources that influenced the model's generation.

In LLMs, attribution systematically links the model's outputs to their source materials, facilitating the identification of the exact documents, datasets, or references that informed the generated response. The primary goal is to uphold transparency, validate factual correctness, and give proper credit to sources. This process is critical for maintaining the credibility and accountability of generative AI systems.

Attribution methods are fundamental for enhancing the interpretability and dependability of LLMs. They support the model's output by providing citations or references, improving accuracy, and reducing the risk of misinformation. This ensures that each response is substantiated by relevant evidence, forming a basis for assessing the sufficiency and relevance of the underlying data.

Research on LLM attribution methodologies encompasses citation generation, claim verification, and hallucination detection techniques. These strategies are aimed at improving the quality and reliability of LLM-generated content. However, challenges remain in implementing adequate attribution, including the need for robust validation mechanisms, managing cases where sources influence the model's reasoning indirectly, handling structured data or non-textual sources (e.g., tables, figures, or images), and addressing the complexities of multi-lingual or cross-lingual data. Overcoming these challenges is essential for successfully integrating attribution methods within LLMs.

As AI and machine learning systems become increasingly prevalent, the demand for accountability, transparency, and reliability intensifies. Attribution techniques are pivotal in achieving these objectives, positioning them as a key area of research and development to advance AI technologies and ensure their responsible deployment.

The main **contributions** of this work are:

*Work done while the first author was an intern at Adobe Research

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 132–136

- A simple zero-shot prompting technique following the idea of textual entailment.
- An attention-based binary classification technique exploring whether attention could help achieve the attribution better.

2 Related Work

Attribution in LLMs has become a vital research area focused on tracing content origins and ensuring accuracy and accountability. Key studies have introduced various techniques and addressed challenges in this field.

Pasunuru et al. (2023) propose a minimal-supervision method for eliciting attributions, improving scalability, and reducing the need for extensive human input. An interactive visual tool for attribution is introduced Lee et al. (2024), aiming to enhance transparency by making attributions more accessible to non-technical users. Zhou et al. (2024) explore attribution in low-resource settings, emphasizing its potential to explain model behavior when data and resources are limited.

The Captum interpretability library is used in Miglani et al. (2023) for generative LLMs, offering insights into the factors influencing model predictions. Khalifa et al. (2024) argue that source-aware training enhances attribution by linking knowledge to specific sources, improving content reliability. The issue of false attribution, stressing the need for more accurate methodologies, is highlighted in Adewumi et al. (2024).

Bohnet et al. (2023) focus on attribution in question-answering systems, proposing methods for evaluating and modeling attributions in QA contexts. A survey of LLM attribution research, summarizing key techniques, challenges, and developments, is provided in Li et al. (2023). Lastly, Yue et al. (2023) explores the automated evaluation of attribution, aiming to streamline validation processes in practical applications.

3 Method and Experimental Setup

The attribution task defined in AttributionBench Li et al. (2024) is framed as a binary classification problem, where the objective is to determine whether a given claim is attributable to its associated references. The work in AttributionBench explores this problem using both zero-shot inference and fine-tuning of LLMs. Similarly, our formulation adopts the same approach to the problem. However, we restrict our methodology to zero-shot experiments due to computational limitations. Additionally, we also investigate if attention layers could help improve the attribution.

3.1 Zero-shot Textual Entailment

We frame this attribution task as a textual entailment problem to ensure simplicity and efficiency.

Textual entailment refers to the relationship between two text fragments, typically a premise and a hypothesis, where the goal is to determine whether the premise entails the hypothesis. Formally, given two sentences S_1 (premise) and S_2 (hypothesis), textual entailment can be defined as a binary relation $\text{Entail}(S_1, S_2)$, where:

$$\text{Entail}(S_1, S_2) = \begin{cases} 1, & \text{if } S_1 \text{ entails } S_2 \\ 0, & \text{otherwise} \end{cases}$$

Here, S_1 entails S_2 if the meaning of S_1 logically supports or guarantees the truth of S_2. The task is to model this relation using techniques, such as deep learning models, to predict this entailment relationship based on large corpora of annotated text pairs.

Why zero-shot Textual Entailment? The core challenge in zero-shot textual entailment is to build models that can generalize well to unseen tasks and relationships, relying purely on contextual understanding rather than task-specific fine-tuning. This is typically achieved through techniques like transfer learning, where models use their broad language understanding to handle specific inference tasks on the fly. For example, a model may be able to infer whether the statement "It is raining outside" entails "The ground is wet" without having been specifically trained on this exact inference.

QUESTION: how much of the world's diamonds does de beers own?

RESPONSE: De Beers owns 40% of the world's diamonds.

CLAIM: De Beers owns 40% of the world's diamonds.

REFERENCE: Title: Diamond Section: Industry, Gem-grade diamonds. The De Beers company, as the world's largest diamond mining company, holds a dominant position in the industry, and has done so since soon after its founding in 1888 by the British businessman Cecil Rhodes.
.....
.....
De Beers sold off the vast majority of its diamond stockpile in the late 1990s – early 2000s and the remainder largely represents working stock (diamonds that are being sorted before sale). This was well documented in the press but remains little known to the general public.

Setting	Model (Size)	ExpertQA (#=612)			Stanford-GenSearch (#=600)			AttributedQA (#=230)			LFQA (#=168)			ID-Avg.
		F1 ↑	FP ↓	FN ↓	F1 ↑	FP ↓	FN ↓	F1 ↑	FP ↓	FN ↓	F1 ↑	FP ↓	FN ↓	F1 ↑
Zero-shot Li et al. (2024)	**FLAN-T5 (770M)**	38.2	1.3	47.4	73.5	15	11.5	80.4	12.2	7.4	37.2	0	48.2	57.3
	FLAN-T5 (3B)	55.6	15.8	27.9	74	17.2	8.7	79.8	15.2	4.8	75.3	6.5	17.9	71.2
	AttrScore-FLAN-T5 (3B)	55.7	32.4	9.6	64.6	27.3	6.5	80.5	16.5	2.6	71.4	21.4	6.5	68.1
	FLAN-T5 (11B)	52	36.4	7.5	59.2	32.7	5	78.6	18.3	2.6	79.8	10.1	10.1	67.4
	T5-XXL-TRUE (11B)	54.5	17.8	27.3	68.5	16.2	15.3	**85.2**	7.8	7	80.4	1.2	17.9	72.2
	FLAN-UL2 (20B)	59.4	22.5	18	72.5	19.2	8	82.5	13	4.3	80.1	4.2	15.5	73.6
	AttrScore-Alpaca (7B)	47.4	11.1	37.7	68.6	21.2	9.8	79	14.8	6.1	68.7	10.1	20.8	65.9
	GPT-3.5 (w/o CoT)	55.3	30.4	12.1	62	30.5	3.8	74.7	20.9	3.5	72.6	22	4.2	66.2
	GPT-3.5 (w/ CoT)	**60.4**	23	16.2	66.1	25.5	7.2	78.9	14.3	6.5	73.4	19.6	6.5	69.7
	GPT-4 (w/o CoT)	56.5	32.8	8	59.8	33.2	3.5	81	15.7	3	71.6	23.2	4.2	67.2
	GPT-4 (w/ CoT)	59.2	26.3	13.9	71.7	19.5	8.5	82.2	10	7.8	80.2	14.9	4.8	73.3
Our Zero-shot	**gpt4-o (05-13-2024)**	52	13.1	33	64.7	**14**	21.2	71.5	10	18.3	81.14	23.8	15.47	64.1
	flan-ul2 (20B)	55	32.7	10	**75.2**	16	8.7	84.16	20.86	12.17	**85.38**	16.6	13.09	**73.8**

Table 1: We evaluate our zero-shot approach against the AttributionBench. Results highlighted in **bold** denote the highest performance. Our method performs better than existing approaches on the Stanford-GenSearch and LFQA sub-datasets. The average ID achieved using our method with `flan-ul2` is **73.8**, representing the highest value.

```
Answer the question with ONLY a 'YES'
or 'NO.' Does the REFERENCE entail the
CLAIM?
```

Figure 1: For our zero-shot experiments, we used this prompt template to query the LLM for determining whether the REFERENCE entails the CLAIM.

In our problem formulation, we task the LLM with a textual entailment problem by utilizing the prompt outlined in Fig. 1. This process involves evaluating the relationship between the given claim and its associated references, as defined in AttributionBench.

3.2 Attention-based attribution

Given the computational limitations, we designed experiments using a single LLM, specifically the `flan-t5-small` model, to analyze attention layers in addressing the attribution task.

Experimental Setup: We utilized the attention weights from each layer as input to a fully connected layer for binary attribution classification. We did this for all 12 layers.

4 Results and Analysis

In the initial phase of our evaluation of the attribution task, we conduct zero-shot experiments. The framework presented in AttributionBench is divided into two key components: in-distribution (ID) and out-of-distribution (OOD) sampling of the dataset. In their experimental setup, AttributionBench employs F1 score, False Positive (FP), and False Negative (FN) rates as evaluation metrics. Consistent with their methodology, we adopt the same metrics - **F1**, **FP**, and **FN** - for the evaluation in this study.

4.1 Evaluation Metrics

F1: The F1 score is a metric used to evaluate the performance of a classification model, specifically its balance between precision and recall.
FP: The False Positive Rate (FP) is a measure used to evaluate the performance of a classification model, specifically in binary classification tasks. It quantifies the proportion of negative instances that are incorrectly classified as positive.
FN: The False Negative (FN) is a metric used to evaluate the performance of classification models. It represents the proportion of actual positive instances incorrectly classified as negative.

4.2 Zero-shot

In this zero-shot setup, we formulate the attribution binary classification task as a simple *textual entailment* problem. To do so, we prompt the LLM using the template shown in Fig. 1. We compare our zero-shot method with the baseline zero-shot approach given in Li et al. (2024). With this simple question, we outperform the baselines in both ID and OOD sets.

We present our zero-shot experimental results in Table 1 for ID data distribution. We mainly used two LLMs: `gpt4-o` Achiam et al. (2023) and `flan-ul2` Raffel et al. (2020). We observe that `flan-ul2` performs better with F1 accuracy metrics in Stanford-GenSearch and the LFQA sub-dataset. The best ID-average (`flan-ul2`) = **73.8**.

Similar to the results observed for in-distribution (ID) data, the highest-performing model for out-of-distribution (OOD) tasks, as presented in Table 2, is `flan-ul2`, specifically for the AttrScore-GenSearch and HAGRID sub-datasets. When evaluating the OOD performance, our approach, leveraging the `flan-ul2` model, achieves the highest average score, reaching an impressive value of

Setting	Model (Size)	BEGIN (# = 436)			AttrScore-GenSearch (# 162)			HAGRID (# = 1013)			OOD-Avg.
		F1 ↑	FP ↓	FN ↓	F1 ↑	FP ↓	FN ↓	F1 ↑	FP ↓	FN ↓	F1 ↑
Zero-shot Li et al. (2024)	**FLAN-T5 (770M)**	79.6	9.2	11.2	80.8	6.2	13	75.9	13.1	10.9	78.8
	FLAN-T5 (3B)	80.2	13.3	6.4	82	6.2	11.7	79	16.9	3.8	80.4
	AttrScore-FLAN-T5 (3B)	78.9	17.7	3	76.3	16.7	6.8	68.6	26.9	2.6	74.6
	FLAN-T5 (11B)	72.3	25	1.1	78.1	16.7	4.9	64.5	30.6	2	71.6
	T5-XXL-TRUE (11B)	**86.4**	4.8	8.7	76.4	2.5	20.4	78.6	14.4	6.8	80.5
	Flan-UL2 (20B)	82.2	13.1	4.6	87.7	5.6	6.8	73.9	21.4	3.9	81.3
	AttrScore-Alpaca (7B)	75.9	20.4	3	82.1	6.8	11.1	73.9	19.9	5.6	77.3
	GPT-3.5 (w/o CoT)	79.4	15.8	4.4	76.7	18.5	4.3	70.1	25.2	2.8	75.4
	GPT-3.5 (w/ CoT)	77.6	14.9	7.3	82.1	11.1	6.8	74	19.7	5.1	77.9
	GPT-4 (w/o CoT)	77.5	19.7	2.1	84.3	14.2	1.2	72.1	23.9	2.8	78
	GPT-4 (w/ CoT)	77.5	18.3	3.7	83.3	8	8.6	75.9	18.5	5.2	78.9
Our Zero-shot	**gpt4-o (05-13-2024)**	79.69	42.66	5.5	**88.24**	17.28	7.4	76.54	42.37	14.41	81.48
	flan-ul2 (20B)	81.55	32.56	8.71	88.05	9.87	13.5	**80.71**	42.79	6.36	**83.43**

Table 2: We evaluate our zero-shot approach against the AttributionBench. Results highlighted in **bold** denote the highest performance. Our method performs better than existing approaches on the AttrScore-GenSearch and HAGRID sub-datasets. The out-of-distribution (OOD) average achieved with our approach utilizing the `flan-ul2` model is the highest, reaching a value of **83.43**.

83.43. This demonstrates the robustness and superior generalization capability of the `flan-ul2` model across both ID and OOD settings.

4.3 Using Attention layers

Preliminary results comparing zero-shot and varying attention layers on the LFQA attribution subset are presented in Table 3. We present layer-wise performance results for all three evaluation metrics. Although the results are mixed, the F1 scores generally outperform the baseline across nearly all layers, except for layers 4 and 8 to 11. Additionally, lower values of false positives (FP) and false negatives (FN) compared to the zero-shot baseline suggest improved performance.

LFQA (#=168)			
	F1 ↑	FP ↓	FN ↓
Our Zero-shot	20	17.85	86.9
using attention			
layer 1	66.67	100	0
layer 2	66.93	98.8	0
layer 3	66.67	100	0
layer 4	0	0	100
layer 5	66.13	100	1.19
layer 6	66.13	100	1.19
layer 7	65.6	100	2.38
layer 8	10.31	9.52	94.04
layer 9	2.35	0	98.8
layer 10	0	0	100
layer 11	66.67	100	0
layer 12	66.93	98.8	0

Table 3: With balanced classes (84 each Class 0/1) using `flan-t5-small`, F1 scores exceed the baseline across most layers, except 4 and 8–11, indicating improved performance, further supported by reduced false positives and false negatives.

5 Conclusion and Future Work

In this paper, we conducted zero-shot experiments on AttributionBench to assess the performance of textual entailment-based approaches for attribution tasks. Our findings show that even without fine-tuning, a simple zero-shot textual entailment approach outperforms the existing baseline in both in-distribution and out-of-distribution settings. Notably, `flan-ul2` demonstrated strong performance across these scenarios, underscoring its robustness and suitability for such tasks. We also preliminary analyzed attention layer behavior using the smaller `flan-t5-small` model. The results suggest that attention mechanisms could provide valuable insights for improving attribution performance.

We plan to overcome computational limitations for future work by conducting fine-tuning experiments. We aim to use more advanced LLMs to perform a deeper analysis of attention layers. This could provide further actionable insights to refine performance and yield more robust findings.

6 Limitations

Limitation 1: Although fine-tuning could enhance the results beyond zero-shot, it comes with additional computational overhead. Therefore, we restricted our experiments to zero-shot settings in this paper and demonstrated how a straightforward zero-shot textual entailment approach can further improve performance.

Limitation 2: Regarding exploring attention mechanisms to enhance the performance of the attribution task, we were similarly restricted by computational limitations. Consequently, we could not utilize computationally demanding models for this analysis. Instead, the experiments were conducted using a lightweight model, `flan-t5-small`.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tosin Adewumi, Nudrat Habib, Lama Alkhaled, and Elisa Barney. 2024. On the limitations of large language models (llms): False attribution. *Preprint*, arXiv:2404.04631.

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. *Preprint*, arXiv:2212.08037.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. Source-aware training enables knowledge attribution in language models. In *First Conference on Language Modeling*.

Seongmin Lee, Zijie J. Wang, Aishwarya Chakravarthy, Alec Helbling, ShengYun Peng, Mansi Phute, Duen Horng Chau, and Minsuk Kahng. 2024. Llm attributor: Interactive visual attribution for llm generation. *Preprint*, arXiv:2404.01361.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *Preprint*, arXiv:2311.03731.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14919–14935, Bangkok, Thailand. Association for Computational Linguistics.

Vivek Miglani, Aobo Yang, Aram Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using captum to explain generative language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 165–173, Singapore. Association for Computational Linguistics.

Ramakanth Pasunuru, Koustuv Sinha, Armen Aghajanyan, LILI YU, Tianlu Wang, Daniel M Bikel, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Eliciting attributions from llms with minimal supervision.

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. *Journal of machine learning research*, 21(140):1–67.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Wei Zhou, Heike Adel, Hendrik Schuff, and Ngoc Thang Vu. 2024. Explaining pre-trained language models with attribution scores: An analysis in low-resource settings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6867–6875, Torino, Italia. ELRA and ICCL.