# Source Attribution for Large Language Models

**Vipula Rawte[2], Koustava Goswami[1], Puneet Mathur[1], Nedim Lipka[1]**

[1]Adobe Research [2]Adobe Inc.

vrawte@adobe.com

## Abstract

As Large Language Models (LLMs) become more widely used for tasks like document summarization, question answering, and information extraction, improving their trustworthiness and interpretability has become increasingly important. One key strategy for achieving this is **attribution**, a process that tracks the sources of the generated responses. This tutorial will explore various attribution techniques, including model-driven attribution, post-retrieval answering, and post-generation attribution. We will also discuss the challenges involved in implementing these approaches, and also look at the advanced topics such as model-based attribution for complex cases, table attribution, multimodal attribution, and multilingual attribution.

## 1 Introduction

In the context of LLMs, attribution refers to the process of identifying and linking the information generated by the model to its original sources. This involves tracing the content produced by the LLM back to the specific documents, datasets, or other reference materials that informed the response. The goal of attribution is to provide transparency, verify the accuracy of the information, and give credit to the original authors or sources. This is particularly important for ensuring the trustworthiness and accountability of the outputs from generative AI systems.

Attribution methods are pivotal in enhancing the trustworthiness and interpretability of LLMs. They substantiate the model's claims with evidence in the form of references or citations, promoting accuracy and reducing misinformation risk. This ensures each statement produced by the model is backed by appropriate references, establishing a framework for evaluating the completeness and relevance of supporting evidence.

Research in attribution methods for LLMs includes techniques for citation generation, claim verification, and hallucination detection. These techniques aim to improve the quality and trustworthiness of the content generated by LLMs. However, implementing attribution methods in LLMs presents challenges. These include the need for robust validation measures, addressing sources used in reasoning but not present in the final output text, dealing with structured sources or sources in other modalities such as tables or figures and images, and dealing with multi- or cross-lingual sources. Overcoming these challenges is crucial for the successful application of attribution methods in LLMs.

As reliance on AI and machine learning models continues to grow, the need for accountability, transparency, and trustworthiness becomes increasingly important. Attribution methods provide a means to achieve these objectives, making them an essential area of study and application in our community and beyond. This tutorial provides an introduction to the problem space and existing work. We'll dive into areas such as model-based attribution for challenging cases, table attribution, multimodal attribution, and multilingual attribution.

## 2 Outline

1. Background and existing work (see Section 4) (40 mins)
2. Model-based approaches for post-generation attribution (see Section 5) (40 mins)
3. Tabular attribution (see Section 6) (40 mins)
4. Multimodal attribution (see Section 7) (40 mins)
5. Multilingual attribution (see Section 8) (25 mins)
6. Attribution and factuality (see Section 9) (25 mins)



*Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Tutorial Abstract*, pages 1–5
December 20-24, 2025 ©2025 Association for Computational Linguistics

## 3 Target Audience

This tutorial is designed for AI practitioners and researchers who are interested in understanding the landscape of current attribution approaches and designing solutions for generative AI for knowledge workers. The tutorial aims to inspire new research and benchmark creation through the discussion of several open challenges.

To get the most out of this tutorial, attendees should have:

- Basic knowledge about LLMs: Understanding the fundamental concepts of LLMs will help attendees grasp the attribution approaches discussed in the tutorial.
- Familiarity with Information Retrieval: Knowledge of information retrieval techniques will be beneficial as these methods are often used in conjunction with LLMs.

**Approximate count:** 30-50. Additionally, the tutorial is designed to accommodate a wide range of participants, from those new to the field to experienced practitioners and researchers. The tutorial's content will be beneficial to all attendees, regardless of their level of expertise.

## 4 Background and existing work

Attribution methods for LLMs in the field of NLP can be categorized based on their approach: (i) direct model-driven attribution, (ii) post-retrieval answering, and (iii) post-generation attribution (Li et al., 2023). We will provide examples for each approach and discuss their nuances, potential, and challenges.

## 5 Model-based approaches for post-generation attribution

Current post-attribution technologies are challenged by "hard cases" where the generated responses infer new information not explicitly present in the provided content, such as generated opinions, calculation results, logical deductions, comparisons, etc. In response to this challenge, we will explore the "implicit" reasoning within an LLM, investigating attentions and activation patterns to gain insights into how the model processes and generates information. We will also examine the intersection between source attribution and a phenomenon known as "hallucination" in LLMs.

## 6 Tabular attribution

Tables are widely used for handling complex semi-structured data in various domains, including healthcare, finance, and education. Application of LLMs to tabular data presents unique challenges: hierarchical header structures, varying formats (e.g., JSON, HTML, CSV, Markdown), lack of straightforward serialization techniques, noisy content, and ambiguity in raw data (e.g., abbreviations, domain-specific terms) (Sui et al., 2023). Due to the high specificity of table data, attributing table structures at the row/column level in generated answers remains under-explored. Prior methods for post-hoc answer attribution use embedding-based retrievers or LLM prompting that are limited to attributing entire tables rather than fine-grained structures (Huo et al., 2023).

We will introduce a novel task, `Fast-Tab: Fine-grained Attribution over Structured Tables` which involves identifying table rows and columns that directly support claims in a generated answer to a user's question. Next, we will discuss existing baseline methods for this new task. Finally, we will expand on our novel agentic framework – `MATSA: Multi-Agent Table Structure Attribution`, that provides inline citations for generated answers based on table structures by coordinating multiple LLM agents.

## 7 Multimodal attribution

When we talk about long documents, often the documents consist of figures, charts, and images along with long text paragraphs. In many cases, the answers to the asked questions might have a contextual link to localized parts of the images along with a connection to the textual passages. In ideal cases, a multimodal attribution system should be able to cite the sentences back to both images and texts. While textual attribution systems in post-hoc settings are improved with the introduction of high-quality embedding and language models, multimodal attribution is still pretty much new and unexplored. Discussing images is not very straightforward; the images can have noisy content, overlapping contents which make it hard to read, and noisy textual contents in the images. Moreover in the case of charts, depending on the types, the textual contents might be overlapping making it hard to be localized. Prior methodology in text-image space has explored the Referring Expression Segmentation task but does not deal with long textual

contexts to be attributed along with the localized image parts. Thus, we will be introducing a new task called multimodal attribution which supports long documents having charts, info-graphics, and natural images.

With the recent advancements in multimodal models, we will explore the potential of leveraging these capabilities to simultaneously attribute generated text to multiple input modalities.

## 8 Multilingual attribution

We will discuss the challenges faced when dealing with multiple languages, especially those that are under-resourced in terms of data and model training. We will also discuss the complexities that arise in a cross-lingual setup where the languages of the source document, query, and answer are different.

## 9 Attribution and factuality

Attribution can help mitigate hallucinations in LLMs. Ensuring that responses include citations to reliable sources can help verify the information. By referencing specific articles, studies, or databases, the model's outputs can be cross-checked for accuracy.

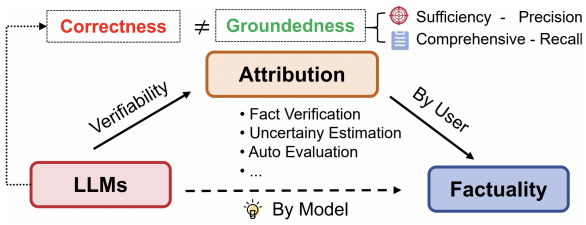


Figure 1: Using attribution for fact-checking (Li et al., 2023).

In the above Fig. 1, using attribution can help developers, and users see the potential sources of an answer and assess its factuality and reliability, enabling them to form their own judgments. Attribution provides a practical approach to reducing hallucinations, as it circumvents the challenge of directly verifying the "truthfulness" of statements, which is often difficult except for the simplest queries.

## 10 Diversity Considerations

The topic of this tutorial is highly relevant to producing verifiable and trustworthy generative outputs that assist knowledge workers and consumers in navigating information. By focusing on attribution approaches, we aim to contribute to responsible and grounded AI solutions.

While the topic is not specifically targeted at a particular underrepresented group, it is universally applicable to all potential participants. However, we will encourage researchers to expand the field to a larger set of languages. The challenges that will be discussed include multi-lingual and cross-lingual attribution, which can be particularly relevant for researchers working with underrepresented languages.

The group of authors of this proposal represents a diverse mix of geolocations (US, India, Germany), roles (academic and industrial), and career stages (Ph.D. candidates, researchers, senior researchers). While we may not necessarily represent minorities, our diverse backgrounds bring a variety of perspectives to the tutorial, enriching the content and its delivery.

## 11 Tutorial Information

**Tutorial Type:** Cutting-edge

**Tutorial Venue:** No preference

**Tutorial Duration:** 3-hour tutorial.

**Reading List**

1. Eliciting Attributions from LLMs with Minimal Supervision (Pasunuru et al., 2023)
2. LLM Attributor: Interactive Visual Attribution for LLM Generation (Lee et al., 2024)
3. Explaining Pre-Trained Language Models with Attribution Scores: An Analysis in Low-Resource Settings (Zhou et al., 2024)
4. Using captum to explain generative language models (Miglani et al., 2023)
5. Source-Aware Training Enables Knowledge Attribution in Language Models (Khalifa et al., 2024)
6. On the Limitations of Large Language Models (LLMs): False Attribution (Adewumi et al., 2024)
7. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models (Bohnet et al., 2023)
8. A Survey of Large Language Models Attribution **[Recommended for preparation before joining the tutorial]** (Li et al., 2023)

9. Automatic Evaluation of Attribution by Large Language Models (Yue et al., 2023)

10. Span Level Attribution: Attribute not just sentences but spans (we just published it and I think it is a new way of defining attribution) Paper:- Peering into the Mind of Language Models: An Approach for Attribution in Contextual Question Answering (Phukan et al., 2024)

**Sharing of Tutorial Materials:** All the tutorial resources will be made publicly available.

## 12 Ethics Statement

The ethical considerations surrounding the use of LLMs and attribution methods are multifaceted. As we continue to rely on these models for tasks like document summarization, question answering, and information extraction, we must ensure the following criteria:

- **AI trustworthiness** Attribution methods aim to enhance the trustworthiness and interpretability of LLMs by substantiating the model's claims with evidence in the form of references or citations.

- **Transparency and Accountability** As the use of AI and machine learning models grows, so does the need for transparency and accountability. Attribution methods can help achieve these objectives by indicating what information has been considered.

- **Inclusivity** Dealing with multi- or cross-lingual sources presents challenges. It's important to ensure that attribution methods are inclusive and considerate of all languages and cultures.

## 13 Presenters

**Vipula Rawte** is a Machine Learning Engineer working at Adobe Experience Cloud. She is a recent Ph.D. graduate from the AI Institute, University of South Carolina, USA, advised by Dr. Amit Sheth. Her primary research interests are in Generative AI and Large Language Models. She has published and given tutorials at EMNLP, LREC-COLING, COLING, and TKDD. She has previously interned at IBM Research Zürich Lab, Dataminr, NYC, and Adobe Research, SJ. Her email is vrawte@adobe.com.

**Koustava Goswami** is a Senior Research Scientist at the Natural Language Processing Group (NLP) at Adobe Research, India. He graduated with a PhD in Computer Science from the National University of Ireland, Galway (now known as University of Galway, Ireland). He has published 30+ papers in NLP, Multimodal Deep Learning, and AI conferences (ACL, EMNLP, EACL, COLING, AAAI, NAACL, ICCV, ECCV, WACV, IEEE Big Data Conference, Frontier Journal, etc.). He also worked as Senior Data Scientist at an MNC and as a visiting researcher at Huawei Research, Bosch Research, etc. He is serving as a PC member at the ACL Rolling Review Paper Submission Process. He was the Organiser PC chair for the AACL-IJCNLP 2023. He is one of the organizers of the workshop SIGTYP happens every year at different NLP conference venues including ACL, EACL, and NAACL. His email is koustavag@adobe.com.

**Puneet Mathur** is a Research Scientist at Adobe. He graduated with a Ph.D. in Computer Science from the University of Maryland, College Park. He has published 35+ research papers on NLP, Multimodal Deep Learning, Computer Vision, Speech, and Artificial Intelligence in top-tier conferences (ACM Multimedia, ACL, NAACL, EMNLP, COLM, CVPR, AAAI, Interspeech, ICASSP, ICWSM, ACM Multimedia, and WACV). He has also worked as a quantitative analyst at a leading hedge fund and as a researcher at Meta (Facebook AI), Microsoft Research, Adobe Research, Verisk AI, and Dataminr. He also served as a senior Program Committee member of AAAI 2023, 2024, and has been the PC Chair for previous Muffin workshops organized at AAAI 2023 and IJACI 2023. His email is puneetm@adobe.com.

**Nedim Lipka** is a Senior Research Scientist at the Document Intelligence Lab of Adobe Research. He has a passion for research in the field of Natural Language Understanding, particularly its applications in conversational services and AI assistants. His research in the field has been published at several international conferences, including COLING, ACL, EMNLP, ICDM, WWW, CIKM, SIGIR, ECIR, etc. In addition to his research, he frequently serves as an area chair in NLP and IR conferences, further demonstrating his commitment and exper-

tise in these areas. Most recently, he has been working on attribution for Adobe's AI assistants. His email is lipka@adobe.com.

## References


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.

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

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. *Preprint*, arXiv:2404.01019.

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.

Vivek Miglani, Aobo Yang, Aram H Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using captum to explain generative language models. *arXiv preprint arXiv:2312.05491*.

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.

Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2024. Peering into the mind of language models: An approach for attribution in contextual question answering. *CoRR*, abs/2405.17980.

Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, and Dongmei Zhang. 2023. Tap4llm: Table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning. *ArXiv*, abs/2312.09039.

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. *arXiv preprint arXiv:2403.05338*.