Unifying Corroborative and Contributive Attributions in Large Language Models

Theodora Worledge* **Judy Hanwen Shen***
Nicole Meister **Caleb Winston** **Carlos Guestrin**[1,2]

Stanford University

Abstract

As businesses, products, and services spring up around large language models, the trustworthiness of these models hinges on the verifiability of their outputs. However, methods for explaining language model outputs largely fall across two distinct fields of study which both use the term "attribution" to refer to entirely separate techniques: citation generation and training data attribution. In many modern applications, such as legal document generation and medical question answering, both types of attributions are important. In this work, we argue for and present a unified framework of large language model attributions. We show how existing methods of different types of attribution fall under the unified framework. We also use the framework to discuss real-world use cases where one or both types of attributions are required. We believe that this unified framework will guide the use case driven development of systems that leverage both types of attribution, as well as the standardization of their evaluation.

1 Introduction

The rapid rise of large language models (LLMs) has been accompanied by a plethora of concerns surrounding the trustworthiness and safety of the LLM outputs. For example, these models can "hallucinate" or fabricate information in response to straightforward prompts [Azamfirei et al., 2023]. Beyond simply verifying that generated content can be trusted, knowing the source from which the output was generated is also crucial in many applications. In fact, Bommasani et. al. [Bommasani et al., 2021] highlight that "*Source tracing is vital for attributing ethical and legal responsibility for experienced harm, though attribution will require novel technical research*". The ubiquitous usage of LLMs in applied settings motivates the development of explanations that provide *both* sources that verify the model output and training sources that are influential in the generation of the output. Unfortunately, attributing an LLM output to sources has been mostly studied in two disjoint fields: citation generation and training data attribution (TDA). Verifying the correctness of model outputs, generally situated in the natural language processing community, includes several different tasks such as fact-checking [Yue et al., 2023], knowledge retrieval [Guu et al., 2020, Gao et al., 2023], attributed question answering [Bohnet et al., 2022], and verifiability in language generation [Rashkin et al., 2021]. Training data attribution, generally situated in the core machine learning community, encompasses a variety of techniques to explain model behavior such as influence functions [Koh and

*Equal contribution. Correspondence to: {worledge, jhshen}@stanford.edu

[1]Chan Zuckerberg Biohub

[2]Stanford Institute for Human-Centered Artificial Intelligence (HAI)

[3]This is an extended abstract of our paper. For a formally defined unified model, properties, detailed survey and case studies, please see our full paper.

Workshop on Attributing Model Behavior at Scale at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Liang, 2017], data simulators [Guu et al., 2023], and data models [Ilyas et al., 2022]. Meanwhile, the term "*attributions*" is used in both fields. When contemplating the two types of attributions, we can think of the former as external validity, which verifies that the output is correct according to external knowledge, and the latter as a certification of internal validity, which provides the source of the generated content. We can easily imagine applications where both types of validity are important for understanding LLM outputs. For instance, a potential criteria to use for identifying a case of model memorization is for a training source to exactly match the model output while also being highly influential in the generation of the output.

In this work, we argue for a unifying perspective of the citation generation and TDA forms of attribution, which we call **corroborative** and **contributive** attributions, respectively. We precisely define each type of attribution and discuss different properties that are desirable in different scenarios. Our work provides a first step towards a flexible, but well-defined notion of language attributions to encourage the development and evaluation of attribution systems capable of providing rich attributions of both types.

2 Motivation: The Necessity of a Unified Perspective

We argue for the study of LLM attributions through a unified perspective of corroborative and contributive attributions. First, we describe the limitations of the current fragmented approach to attributions and then we summarize the case for unification.

2.1 Gaps in existing approach to language model attributions

Misalignment between TDA methods and their use cases Most training data attribution (TDA) papers present their methods as standalone solutions for motivating use cases such as identifying mislabeled data points [Koh and Liang, 2017, Yeh et al., 2018, Pruthi et al., 2020, Schioppa et al., 2021, Kwon et al., 2023], debugging domain mismatch [Koh and Liang, 2017], and understanding model behavior [Grosse et al., 2023]. In the setting of language models, however, TDA methods may not be a comprehensive solution; training sources that are irrelevant to the content of the test example may be flagged as influential by TDA methods [Grosse et al., 2023]. This is undesirable because the semantic meaning of a flagged training source can indicate its importance in generating the semantic meaning of the output. For instance, when searching for misleading training sources in a Question Answering (QA) language model, it is important to understand which of the sources flagged by TDA methods corroborate the misinformation in the output. This is also the case in other practical applications, such as debugging toxicity. Without carefully considering the types of attribution needed in different use cases, we risk investing in methods that, while establishing essential foundations, may not align with practical use.

Citation generation methods do not explain model behavior Corroborative methods (e.g., fact checking [Yue et al., 2023], citation generation [Guu et al., 2020]) are not designed to explain model behavior. For example, the verifying the truthfulness of outputted facts using sources from an external corpus does little to explain why the model generated such an output. When outputted facts are found to be incorrect, there is limited recourse for correcting model behavior. Thus, corroborative attributions alone cannot address all the challenges of explaining the outputs of language models.

Emergent usage of language models require a richer notion of attributions The emerging use of LLMs in domains such as health care and law involves tasks such as document generation and domain-specific QA that require both explanations of whether the output is correct and where the output came from. As an example, in the legal domain, different products based on LLMs such as legal QA, immigration case document generation, and document summarization are currently under development [1]. In this setting, *corroborative* attributions are important to ensure that a generated legal document follows local laws. The sources for such corroborative attributions need not be in the training data. Simultaneously, *contributive* attributions are important for understanding the training documents from which the generated legal document is borrowing concepts. In the legal setting, context and subtle changes in wording matter [Bommasani et al., 2021].

[1] Y-Combinator companies in this area include Casehopper, Lexiter.ai, DocSum.ai, and Atla AI.

2.2 Motivating a unified framework of attributions

Developing a standardized language to describe different types of attribution will improve the (1) **clarity** and (2) **simplicity** of scholarly discussion around attributions. Furthermore, identifying the common components of all attributions provides (3) **modularity** for improving individual components and better (4) **reproducibility** of results. Looking ahead to future work, a unified perspective motivates the (5) **hybrid development** of both corroborative and contributive attributions.

"Attribution" is an overloaded, ambiguous term The term "attribution" is overloaded in machine learning literature. Moreover, recent works have attempted to provide both types of attribution for language models under the vague umbrella term of "attributions" [Bohnet et al., 2022, Park et al., 2023, Grosse et al., 2023]. While existing work recognizes the importance of both corroborative and contributive attribution [Huang and Chang, 2023], comparing these two notions is difficult without precisely delineating between them while also acknowledging their similarities. A unified perspective of both types of attributions improves the **clarity** of technical progress on attributions.

Attribution methods exist concurrently in disjoint fields The two dominant interpretations of attributions for language model outputs come from the natural language processing (NLP) and explainability communities. In NLP literature, attributing a model output to a source generally refers to identifying a source that corroborates the output [Rashkin et al., 2021, Bohnet et al., 2022, Yue et al., 2023, Liu et al., 2023]. We refer to this as *corroborative attribution*. This differs from TDA work, where attributing a model output to a source refers to identifying a training source that highly influenced the model to produce that output [Park et al., 2023, Guu et al., 2023, Lundberg and Lee, 2017, Koh and Liang, 2017]. We refer to this as *contributive attribution*. To the best of our knowledge, there is no established framework that unifies these different types of attributions. Furthermore, methods to achieve both types of attribution and metrics to evaluate them have been developed separately. Our goal is to introduce **simplicity** in understanding the vast landscape of prior work by creating a shared language to discuss attribution methods across different tasks.

Attributions have common components Despite these two types of attribution being studied in different fields, there are commonalities in system components, properties, metrics, and evaluation datasets. For example, fact-checking using corroborative attributions has significant overlap with fact-tracing using contributive attributions, in terms of metrics and evaluation datasets [Akyürek et al., 2022]. Defining the shared components of different types of attributions introduces **modularity** that better enables the improvement of individual components of attribution systems. Furthermore, precise definitions of properties shared across different attributions allow for better **reproducibility** in implementations of attribution systems.

A unifying perspective enables the development of richer attribution systems Because both notions of attribution are relevant to use cases that improve the safety and reliability of language models as information providers, both are often simultaneously relevant in application settings. There are real-world use cases of attribution that require careful reasoning and differentiating between these two interpretations; some use cases even require both notions of attribution. These use cases should motivate the **hybrid development** of methods that provide both citation and TDA for LLM outputs. Furthermore, methods used in one type of attribution may be leveraged to develop other types of attributions.

3 Conclusion and Future Work

3.1 Key Directions for Future work

To conclude, we highlight several promising directions for future work.

Counterfactual contribution to output evaluators In Definition **??**, we outline the possibility of contributive evaluators that are sensitive to semantic changes in the counterfactual output, rather than to changes in the counterfactual loss. The notion of citation to parametric content in [Huang and Chang, 2023] also addresses this potential connection between contributive attribution and the semantic content of the output. To the best of our knowledge, such output-based contributive

attributions for LLMs have not yet been explored. Future work in addressing this challenging technical problem would allow for semantically meaningful contributive attributions.

Contributive attributions with large-scale training data The large scale of data used to train LLMs raises concerns not only about the high resource burdens of TDA methods, but also whether the influence of a single training source is meaningfully noticeable on the loss, not to mention the output. Past work has quantitatively observed that training sources with high influences are more rare than not, but they do exist and in fact largely make up the total influence on an LLM output [Grosse et al., 2023]. Nonetheless, future work may consider extending contributive attributions to notions of influence on a group of training sources, rather than individual training sources [Koh et al., 2019]. Also, the ubiquity of finetuning encourages further work on TDA methods suited for finetuned models [Kwon et al., 2023]. In this case, the attribution domain could be restricted to the finetuning dataset, which is orders of magnitude smaller than the pre-training dataset. This direction is an interesting pursuit in and of itself, especially for model developers interested in debugging fine-tuned models.

Hybrid attribution systems While we present a framework that unifies existing work in both corroborative and contributive attribution literature, developing techniques capable of both types of attributions is left to future work. The area of *fact-tracing* makes a step in this direction by providing contributive attributions in a setting where corroboration matters [Akyürek et al., 2022]. However, the identification and corroboration of facts within the language model output requires further work. Hybrid attribution systems would improve the customizability of attributions, potentially making them useful across a broader range of applications.

Standardized Evaluation From our survey of attribution methods, particularly for corroborative attribution, we observe that evaluation is not standardized between methods. Each attribution method is evaluated on different datasets and often with different metrics. For example, GopherCITE's [Menick et al., 2022] outputs are evaluated on a subset of NaturalQuestions and ELI5 with binary metrics if the answer is plausible and supported by the attribution. On the other hand, WebGPT's [Nakano et al., 2021] outputs are evaluated on a different subset of ELI5 and open-ended dialogue interactions by comparisons to human-generated attributions. More broadly, the utility of an attribution can be expanded beyond correctness to the other properties we introduce.

Use-Case Driven Method Development and Properties-Guided Evaluation In our work, we explore tasks and case studies where attributions are important for industry applications of LLMs. We recommend that attribution system developers choose a use case and then identify the relevant properties for evaluation. This approach of goal-driven development is preferable to strong-arming a developed method to serve a use case. Furthermore, goal-driven development may surface additional settings where corroborative and contributive attributions are needed simultaneously.

3.2 Conclusion

This paper presents a unifying framework for corroborative and contributive attributions in LLMs. We formulate an interaction model to define the core components of attributions and to define their properties. This framework serves as a lens for analyzing existing attribution methods and use cases for attributions. Our analysis elucidates prescriptive suggestions for future research, namely CCO evaluators, the challenges of contributive methods at the scale of LLMs, the value of hybrid attributions systems, the need for standardized evaluation of attribution systems, and goal-driven development. We hope our unifying perspective on the field of attributions leads to improved solutions for misinformation, accountability, and transparency in real-world applications of language models.

References

Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):1–2, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

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

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

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

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*, 2021.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

Kelvin Guu, Albert Webson, Ellie Pavlick, Lucas Dixon, Ian Tenney, and Tolga Bolukbasi. Simfluence: Modeling the influence of individual training examples by simulating training runs, 2023.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

Chih-Kuan Yeh, Joon Sik Kim, Ian E. H. Yen, and Pradeep Ravikumar. Representer point selection for explaining deep neural networks, 2018.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions, 2021.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models, 2023.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale, 2023.

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

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

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446, 2022.

Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. On the accuracy of influence functions for measuring group effects. *Advances in neural information processing systems*, 32, 2019.

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

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.