Explanation in the Era of Large Language Models

Zining Zhu[1,2], **Hanjie Chen**[3,4], **Xi Ye**[5], **Qing Lyu**[6]
Chenhao Tan[7], **Ana Marasović**[8], **Sarah Wiegreffe**[9,10]
[1] Stevens Institute of Technology, [2] University of Toronto, [3] Johns Hopkins University,
[4] Rice University, [5] University of Texas Austin, [6] University of Pennsylvania
[7] University of Chicago, [8] University of Utah,
[9] Allen Institute for AI, [10] University of Washington,
zzhu41@stevens.edu, hchen210@jh.edu,
xiye@cs.utexas.edu, lyuqing@sas.upenn.edu, chenhao@chenhaot.com,
ana.marasovic@utah.edu, wiegreffesarah@gmail.com

Abstract

Explanation has long been a part of communications, where humans use language to elucidate each other and transmit information about the mechanisms of events. There have been numerous works that study the structures of the explanations and their utility to humans. At the same time, explanation relates to a collection of research directions in natural language processing (and more broadly, computer vision and machine learning) where researchers develop computational approaches to explain the (usually deep neural network) models. Explanation has received rising attention. In recent months, the advance of large language models (LLMs) provides unprecedented opportunities to leverage their reasoning abilities, both as tools to produce explanations and as the subjects of explanation analysis. On the other hand, the sheer sizes and the opaque nature of LLMs introduce challenges to the explanation methods. In this tutorial, we intend to review these opportunities and challenges of explanations in the era of LLMs, connect lines of research previously studied by different research groups, and hopefully spark thoughts of new research directions.

1 Outline of Tutorial

This tutorial will take about 3 hours:

- Introduction & Desiderata (30 minutes)
- Free-text, CoT, Structured Explanations (50 minutes)
- Importance Scores (40 minutes)
- Mechanistic, Causal, etc (40 minutes)
- Conclusion & Discussion (20 minutes)

The following subsections list some more detailed content for each section.

1.1 Introduction

Explanation has been an important component in languages and their use. Explanation can reveal the underlying mechanism of the phenomena to be explained (Keil, 2006). Explanation is also a process (Achinstein, 1983). Explanation can be part of an argumentative tool that help humans exploit the uniqueness of societal environment (Mercier and Sperber, 2017), and have profound impacts on the cognition procedures of learning and inference (Lombrozo et al., 2019).

There are many types of explanations. In the literature of philosophy and psychology, one fruitful taxonomy is mechanistic explanations (citing the components and procedures), teleological explanations (citing the goals), and formal explanations (citing the categories) (Lombrozo, 2012). In the NLP and explainable AI literature, there have been many types of explanations as well. Taxonomizing by the nature of the explanandum, we have the explanations towards model predictions vs. the explanations towards other problems (for example, events). Taxonomizing by whether the explanations are produced with the predictions, we have pre-hoc explanations vs. post-hoc explanations. Taxonomizing by the methods to arrive at the explanations, there are many popular methods including free-text, attribution scores, and mechanistic explanations, many of which will be discussed in the next a few sections.

In recent years, the advance of LLM technologies has introduced unique opportunities for explanations. In some application scenarios of education (Khan, 2023; Duolingo, 2023) and commerce (Stanley, 2023), explanations can improve the AI systems. In this tutorial, we will focus on the recent opportunities and challenges introduced by LLMs, which have not been covered by prior tutorials.

1.2 Desiderata of Explanation

What is a good explanation? On a high level, good explanations are the ones that achieve the intended

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts), pages 19–25

communicative goals, which can help developers debug or improve human decisions. On a detailed level, the literature has also identified some desirable properties for measuring the quality of explanations, including but not limited to:

Faithfulness. An explanation should accurately reflect the reasoning process behind the model's prediction (Jacovi and Goldberg, 2020; Lyu et al., 2023a).

Plausibility. An explanation should be understandable and convincing to the target audience (Herman, 2019; Jacovi and Goldberg, 2020).

Usefulness. An explanation should be helpful for the user to achieve a pre-defined goal (Zhou and Shah, 2022; Bansal et al., 2021; Chen et al., 2023).

Minimality. An explanation should only include the smallest number of necessary factors (Halpern and Pearl, 2005; Miller, 2018).

On an implementation level, the procedure to generate explanations has some desirable properties as well. The algorithms should require realistic data and computation resources. Depending on the accessibility of the models, the requirement to access the internal weights of the models can also be noteworthy.

Note that it may be difficult to satisfy all of the properties above at the same time (e.g., minimality and plausibility). One can also argue that these properties are not the "first-order principles" that determine the explanation qualities. We will describe the nuances in this tutorial.

When discussing each desideratum in the tutorial, we will impose a special focus on the challenges and opportunities brought by LLMs. For example, recent studies find that LLM can generate more *plausible* explanations (Marasović et al., 2022; Wiegreffe et al., 2022), which are, however, not necessarily faithful to their internal reasoning mechanism (Turpin et al., 2023; Lyu et al., 2023b).

1.3 Method: Free-Text/CoT

We then proceed with four sections describing the methods to generate explanations. For each category of method, we will also describe the corresponding evaluation criteria and illustrate how well the explanation methods work.

The advancement of LLMs introduces unique opportunities, including the chain-of-thought (CoT) (Wei et al., 2022). There have been various approaches to leverage LLMs' reasoning abilities to explain the problems (Marasović et al., 2022).

Compared to prior, smaller models, larger LMs are able to generate free-text explanations on a zero-shot or few-shot setting. Specifically, the qualities of the generated explanations can be comparable to, and sometimes more preferable than those that were written by humans (Wiegreffe et al., 2022).

The LLMs have the potential to build a special category of models, self-rationalizing models, which outputs both the prediction and the reasons toward that prediction at the same time. The self-rationalizing models can introduce unique advantages. For example, the models themselves may be less susceptible to spurious correlations, making more predictions "right for the right reasons" (Ross et al., 2022). The generated CoT could also be beneficial to "student models" (Wang et al., 2023; Pruthi et al., 2022).

LLMs are also known for "hallucination": they tend to improvise and produce nonfactual content (Ji et al., 2023), so the LLM-produced explanations can be unreliable, even after few-shot demonstrations (Ye and Durrett, 2022). We will describe some recent works to improve this problem, e.g., the approaches of Lyu et al. (2023b). Relatedly, some recent works study prompt writing methods that aim at improving the reasoning qualities, including context faithfulness (Zhou et al., 2023) and help-me-think (Mishra and Nouri, 2023).

1.4 Method: Structured Explanations

Researchers have long wanted to figure out the underlying structures of the explanations. The study of the structures of explanations can be traced back to Hempel and Oppenheim (1948). Explanations can contain various structures. Inductive explanations present observed events that can improve the statistical likelihood that the explanandum event is true (Hempel, 1958). Deductive explanations provide logical arguments that can derive the explanandum event following a set of widely accepted rules (Hempel, 1962). Abductive explanations, on the other hand, aim at making the event more *plausible* while allowing more relaxed structures (Lombrozo, 2012; Zhao et al., 2023).

Wiegreffe and Marasović (2021) listed many structured explanation approaches. They can be presented in graphs (WorldTree (Jansen et al., 2018), OpenbookQA (Mihaylov et al., 2018)), symbolic rules (Lamm et al., 2020), semi-structured texts (Ye et al., 2020), etc.

More recently, many additional structures are found to be useful, for example, Tree-of-thoughts (Yao et al., 2024), Graph-of-thoughts (Besta et al., 2024) and Everything-of-thoughts (Ding et al., 2023). The advance of LLMs allows unprecedented flexibility in controlling the structures and contents of explanations. We will describe some of the new approaches to make these controls possible. We will also describe some ways to evaluate the utility of these new approaches.

1.5 Method: Importance Scores

A category of methods to explain data-driven systems aim at attributing system behavior to the instances in the input data. This category of method is referred to as importance scores. We will discuss some popular importance score-based methods spanning two prominent paradigms (token-wise attribution and instance-wise attribution) in the context of NLP models, especially LLMs.

We will first set up some basics of importance score methods, covering the most commonly used token-level attribution methods (Ribeiro et al., 2016; Lundberg and Lee, 2017; Sundararajan et al., 2017) and instance-wise attribution methods (Koh and Liang, 2017). We plan to give a high-level introduction of these methods. We will omit the technical details, but emphasize on the cost of computation and the requirements on the access to model details for obtaining the interpretations using different methods, so as to better deliver the applicability of these methods on LLMs. We will also introduce the common evaluation protocols that are unique to the importance score methods, such as sufficiency and comprehensiveness (DeYoung et al., 2020).

Next, we will discuss the unique challenges and opportunities of applying the importance score methods on interpreting and developing LLMs. LLMs are associated with extreme scale in both model size and training data size, which can render many previously viable importance score methods prohibitively expensive. We will showcase how importance score methods such as influence function are adapted for interpreting LLMs (Grosse et al., 2023; Piktus et al., 2023), and how they are utilized for gaining deeper understanding of LLMs' behavior (Wu et al., 2023; Madaan and Yazdanbakhsh, 2022) or for improving model performance (Krishna et al., 2023).

1.6 Method: Mechanistic, Causal, Others

Explanations are not the only approaches that help us "open the black boxes". There are many other methods that aim at achieving similar goals. We will briefly mention some of these popular methods, and discuss how they relate to the explanation methods mentioned in our tutorial.

Mechanistic interpretability approaches try to describe the mechanisms of how the DNN-based AI systems work. A representative work in mechanistic interpretability is neural circuits (Conmy et al., 2023). Causal mediation analyses try to apply causal analysis tools to understand the models. Kıcıman et al. (2023) provides an overview of the tools and frontiers related to causal analysis in DNN models.

Model editing provides explanations from a counterfactual aspect: "What would be the output, had this model been modified into the other way?" Some recent works include ROME (Meng et al., 2022) and MEND (Mitchell et al., 2022). Yao et al. (2023) provides a summary on this.

We recommend the readers to check out the EACL tutorial (Mohebbi et al., 2024) and the reviewing article by Ferrando et al. (2024) for more details, especially about Transformer-specific mechanistic interpretability. Our tutorial includes explanation topics that are beyond Transformers.

2 Reading List

In addition to the papers cited in this proposal, we also recommend this reading list on Notion and previous relevant tutorials: Belinkov et al. (2020) presented approaches to interpret the structures and behavior of neural network models; Wallace et al. (2020) described approaches to understanding the predictions of neural network models; Boyd-Graber et al. (2022) focused on the human aspect of explanation evaluation. Compared to the previous tutorials, our tutorial covers some new topics, including free-text / CoT explanations, and structured explanations, etc. We will present perspectives that connect the explanations as model interpretation tools and the explanations as communication procedures.

3 Type of the Tutorial

The tutorial is designed to be at the cutting edge, encompassing advanced technologies for explaining NLP models. In particular, the tutorial will emphasize on explanations in the context of LLMs, including generation and evaluation methods.

4 Target Audience and Prerequisites

Anyone interested in explainable NLP and LLMs is welcome. We anticipate an audience size of approximately 200.

Attendees are expected to have basic knowledge of NLP tasks (e.g., text classification, question answering) and neural language models (e.g., BERT, GPT). We plan to make tutorial materials (e.g., slides, media) public.

5 Breadth and Diversity

Our tutorial is ensured to cover a wide spectrum of explanation topics, ensuring that attendees are exposed to a comprehensive range of concepts, techniques, and advances. We will incorporate seminal works and recent advancements from a wide array of researchers in the field into the tutorial.

The instructors are diverse in terms of gender, nationality, affiliation, and seniority (from PhD students to postdocs to professors). We plan to organize open Q&A sessions to create a space for participants to directly engage with presenters, clarifying doubts and exploring different viewpoints. This format ensures that participants from various backgrounds can contribute to shaping the discussion. In particular, we encourage participants from underrepresented groups to share thoughts and insights and provide feedback.

6 Presenters

Zining Zhu is an incoming assistant professor at the Stevens Institute of Technology. He obtained his Ph.D. in 2024 at the University of Toronto. His research includes model control and interpretability. Zining co-instructed the Natural Language Computing course (CSC401) at UofT in 2023 and 2022, with class size around 200.

Hanjie Chen is an incoming assistant professor at Rice University, and is currently a postdoc at Johns Hopkins University. She obtained her Ph.D. in 2023 at the University of Virginia. Her research focuses on the interpretability/explainability of neural language models. As the primary instructor, she co-designed and instructed the course, CS 6501/4501 Interpretable Machine Learning, at UVA in Spring 2022. She received teaching awards at UVA.

Xi Ye is an incoming assistant professor at The University of Alberta. He obtained his Ph.D. in 2024 at the University of Texas at Austin. His research focuses on leveraging explanations to improve language models for complex textual reasoning tasks. He also works on program synthesis and semantic parsing.

Qing Lyu is a Ph.D. candidate at the University of Pennsylvania, advised by Chris Callison-Burch and Marianna Apidianaki. Her research interests lie in the intersection of linguistics and natural language processing, as well as the interpretability and robustness of language models.

Chenhao Tan is an assistant professor of computer science and data science at the University of Chicago, and is also affiliated with the Harris School of Public Policy. He obtained his PhD degree in the Department of Computer Science at Cornell University and bachelor's degrees in computer science and in economics from Tsinghua University. Prior to joining the University of Chicago, he was an assistant professor at the University of Colorado Boulder and a postdoc at the University of Washington. His research interests include natural language processing, human-centered AI, and computational social science. His work has been covered by many news media outlets, such as the New York Times and the Washington Post. He also won a Sloan research fellowship, an NSF CAREER award, an NSF CRII award, a Google research scholar award, research awards from Amazon, IBM, JP Morgan, and Salesforce, a Facebook fellowship, and a Yahoo! Key Scientific Challenges award.

Ana Marasović is an assistant professor in the Kahlert School of Computing at the University of Utah. Her primary research interests are at the confluence of NLP, explainable AI, and multimodality. Previously, she was a Young Investigator at the Allen Institute for AI and held a concurrent appointment in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. She obtained her PhD in 2019 from Heidelberg University. She received Best Paper Award ar ACL 2023, Best Paper Honorable Mention at ACL 2020, and Best Paper Award at SoCal 2022 NLP Symposium.

Sarah Wiegreffe is a Young Investigator (postdoc) at the Allen Institute for AI, where she is a member of the Aristo team. She also holds a courtesy appointment in the Allen School at the

University of Washington. Her research interests encompass interpretability + explainability of NLP models, with a focus on the faithfulness of generated text to internal LM prediction mechanisms and the utility of model-generated textual explanations to humans. She received her PhD in 2022 from Georgia Tech, advised by Mark Riedl.

7 Technical Equipment

No special requirements. We simply require fundamental technical equipment for our in-person tutorial, including essentials like projectors and screens, microphones, cables and adapters, etc.

8 Ethics Statement

This tutorial aims to provide a comprehensive overview of explanations for NLP, especially the challenges and opportunities in the era of LLMs. We hope the tutorial will provide the audience with a profound understanding of the pivotal role of explanations in enhancing human trust in LLMs, alleviating ethical concerns, and fulfilling societal responsibilities.

Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Peter Achinstein. 1983. *The Nature of Explanation*. Oxford University Press.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–16, New York, NY, USA. Association for Computing Machinery.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32, Seattle, United States. Association for Computational Linguistics.

Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2023. Machine Explanations and Human Understanding.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Ruomeng Ding, Chaoyun Zhang, Lu Wang, Yong Xu, Minghua Ma, Wei Zhang, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2023. Everything of thoughts: Defying the law of penrose triangle for thought generation. *arXiv preprint arXiv:2311.04254*.

Team Duolingo. 2023. Duolingo Max Uses OpenAI's GPT-4 For New Learning Features.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models.

Roger Baker Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamil.e Lukovsiut.e, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Sam Bowman. 2023. Studying large language model generalization with influence functions. *ArXiv*, abs/2308.03296.

Joseph Y. Halpern and Judea Pearl. 2005. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of*

Science, 56(4):843–887. Publisher: The University of Chicago Press.

Carl G. Hempel. 1958. The theoretician's dilemma: a study in the logic of theory construction. Accepted: 2017-02-24T17:47:36Z Publisher: University of Minnesota Press, Minneapolis.

Carl G. Hempel. 1962. Deductive-nomological vs. statistical explanation.

Carl G. Hempel and Paul Oppenheim. 1948. Studies in the Logic of Explanation. *Philosophy of Science*, 15(2):135–175. Publisher: [The University of Chicago Press, Philosophy of Science Association].

Bernease Herman. 2019. The Promise and Peril of Human Evaluation for Model Interpretability. *arXiv:1711.07414 [cs, stat]*. ArXiv: 1711.07414.

Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12):1–38.

Frank C. Keil. 2006. Explanation and Understanding. *Annu Rev Psychol*, 57:227–254.

Sal Khan. 2023. Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access - Khan Academy Blog.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post hoc explanations of language models can improve language models. *ArXiv*, abs/2305.11426.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. QED: A Framework and Dataset for Explanations in Question Answering.

Tania Lombrozo, Daniel Wilkenfeld, T Lombrozo, and D Wilkenfeld. 2019. Mechanistic versus functional understanding. In *Varieties of understanding: New perspectives from philosophy, psychology, and theology*, pages 209–229. Oxford University Press New York, NY.

Tanya Lombrozo. 2012. Explanation and Abductive Inference. In Keith J. Holyoak and Robert G. Morrison, editors, *The Oxford Handbook of Thinking and Reasoning*, 1 edition, pages 260–276. Oxford University Press.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2023a. Towards Faithful Model Explanation in NLP: A Survey.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023b. Faithful Chain-of-Thought Reasoning.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E. Peters. 2022. Few-Shot Self-Rationalization with Natural Language Prompts. In *Findings of NAACL*. arXiv.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 36.

Hugo Mercier and Dan Sperber. 2017. *The Enigma of Reason*. The enigma of reason. Harvard University Press, Cambridge, MA, US.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]*. ArXiv: 1706.07269.

Swaroop Mishra and Elnaz Nouri. 2023. HELP ME THINK: A Simple Prompting Strategy for Non-experts to Create Customized Content with Models.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast Model Editing at Scale. In *ICLR*. arXiv.

Hosein Mohebbi, Jaap Jumelet, Michael Hanna, Afra Alishahi, and Willem Zuidema. 2024. Transformer-specific interpretability. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 21–26, St. Julian's, Malta. Association for Computational Linguistics.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The ROOTS search tool: Data transparency for LLMs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 304–314, Toronto, Canada. Association for Computational Linguistics.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*.

Alexis Ross, Matthew E. Peters, and Ana Marasović. 2022. Does Self-Rationalization Improve Robustness to Spurious Correlations?

Morgan Stanley. 2023. Morgan Stanley wealth management deploys GPT-4 to organize its vast knowledge base.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting.

Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, Online. Association for Computational Linguistics.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-Consistent Chain-of-Thought Distillation.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing Human-AI Collaboration for Generating Free-Text Explanations. In *NAACL-HLT*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasović. 2021. Teach Me to Explain: A Review of Datasets for Explainable NLP. In *Proceedings of NeurIPS*.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. 2023. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *ArXiv*, abs/2307.13339.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing Large Language Models: Problems, Methods, and Opportunities.

Qinyuan Ye, Xiao Huang, Elizabeth Boschee, and Xiang Ren. 2020. Teaching machine comprehension with compositional explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1599–1615, Online. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The Unreliability of Explanations in Few-shot Prompting for Textual Reasoning. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

Wenting Zhao, Justin T. Chiu, Claire Cardie, and Alexander M. Rush. 2023. Abductive Commonsense Reasoning Exploiting Mutually Exclusive Explanations. arXiv.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful Prompting for Large Language Models.

Yilun Zhou and Julie Shah. 2022. The Solvability of Interpretability Evaluation Metrics. ArXiv:2205.08696 [cs].