Tutorial Proposal: Causality for Large Language Models

Zhijing Jin **Sergio Garrido**

Abstract

In this tutorial, we will explore the intersection of causality and large language models (LLMs). Our goal is to provide a comprehensive understanding of how causal inference can enhance the performance, interpretability, and robustness of LLMs. The tutorial will cover foundational concepts in both fields, discuss emerging trends, present three paradigms for causality for LLM research, and corresponding practical applications. We will engage the audience with a comprehensive overview and diverse perspectives.

1 Description

Recently, large language models (LLMs) have achieved remarkable success across many tasks and benchmarks [53, 4, 35, 3, 49, 1], and are being integrated into increasingly complex systems and real-world applications [12, 8, 54]. However, there are still several main technical blockers hindering further technical progress and responsible deployment of LLMs. First, there is still some technical limitations in its reasoning capability, such as a big question over whether it can perform causal reasoning [25, 20, 22], which is a crucial skill before it can be integrated into decision-making systems. Second, even if LLMs largely succeed at many simpler tasks, they still lack transparency of how their decisions can be explained, which calls for timely interpretability research before it can be reliably deployed in downstream use cases [27, 38, 32]. Lastly, its performance also suffers from unwanted variations due to minor perturbations in the input text, so still lacking trust in whether it can perform a task well under dynamic perturbation [39, 31, 55].

On the other hand, causal inference is the study that discovers cause-effect relationships from interventional and/or observational data [37, 47, 52]. Causality started as a philosophical subject [2, 41, 23], and in the recent centuries integrated into statistics with established concepts and tools [11, 40, 46, 37]. However, text as a form of data for causality has been only explored very recently [15, 24, 10], and there is little work providing a summary of how causality can be further connected to the development of LLMs.

This gap indicates enormous new opportunities for connecting causality and LLMs, to enhance LLM capabilities of reasoning, model interpretability, and reliable perfor-

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

mance under perturbations, addressing the aforementioned limitations. Emerging studies at the intersection of causality and LLMs have demonstrated promising results [50, 38, 20, 25, 29, 48].

This tutorial aims to address the synergies across causal inference and LLMs, to summarize and present a unified view of connecting both areas. We will first demonstrate success and development in both areas, covering sufficient technical details of basic definitions and axioms in causal inference. Based on the technical background, we will introduce three types of avenues where causality can benefit LLM development: enhancing causal reasoning skills, enabling interpretability of model behavior, and standardizing robust testing frameworks for LLM evaluation.

Goals of This Tutorial: Our tutorial aims to: (1) provide a comprehensive understanding of the interplay between causality and LLMs, (2) equip attendees with the skills to enhance LLMs using causal inference techniques, (3) foster a discussion on the future of LLMs and causality, highlighting open problems and research opportunities, and (4) engage a broad audience from various subfields of computer science and machine learning by presenting diverse perspectives.

2 Outline

In the following, we provide an outline for the 2.5-hour tutorial, supplied with important references *in italics* that will be covered in each section.

I: Introduction (30 mins)

- Overview of the tutorial
- Basics and recent progress of LLMs
- Basics and technical background of Causal Inference
- Main technical blockers for LLM development
- *References: LLMs: [53, 6, 35, 3, 49, 1], Causality: [45, 44, 10, 21].*

II: Causal Reasoning Skills in LLMs (30 min)

- A taxonomy of different types of causal reasoning
- Testing the existing causal reasoning skills in LLMs
- Enhancing formal causal inference with tool-augmented LLMs
- *References: Causal reasoning types: [37, 47, 36], Causal reasoning for LLMs: [42, 25, 20, 22, 5], Tool-augmented LLMs: [43, 30, 17, 51, 28, 18].*

III: LLM Interpretability by Causal Inference (30 min)

- White-box model interpretability: Using direct intervention and counterfactual prediction to obtain the causal effect of neurons and circuits in LLMs
- Black-box model interpretability: Using do-calculus to obtain the causal effect sizes of certain properties in the input text on the prediction
- *References: Mechanistic (i.e., white-box) interpretability: [33, 9, 34, 32, 16, 14, 29, 13], Behavioral (i.e., black-box) interpretability: [48, 26, 19].*

IV: LLM Robustness Testing by Causal Frameworks (30 min)

- The problem of confounding in existing evaluation practice for LLMs
- Designing test sets with perturbations grounded in the causal graph
- Examples of causal robustness tests for LLMs
- Including 10-min Q&A
- *References: Problems with existing datasets: [26], Designing causally-controlled tests: [39, 31, 48, 7].*

References

[1] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

[2] Helen Beebee, Christopher Hitchcock, Peter Charles Menzies, and Peter Menzies. 2009. *The Oxford handbook of causation*. Oxford Handbooks.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

[5] Roberto Ceraolo, Dmitrii Kharlapenko, Amélie Reymond, Rada Mihalcea, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. 2024. CausalQuest: Collecting natural causal questions for AI agents. *CoRR*, abs/2405.20318.

[6] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

[7] Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Bernhard Schölkopf, Mrinmaya Sachan, Rada Mihalcea, and Zhijing Jin. 2022. Testing occupational gender bias in language models: Towards robust measurement and zero-shot debiasing. *CoRR*, abs/2212.10678.

[8] Cognition. 2024. [link].

[9] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

[10] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

[11] Ronald A. Fisher and E. B. Ford. 1927. The spread of a gene in natural conditions in a colony of the moth panaxia dominula l. *Heredity*, 11:143–174. Early work by Fisher on the application of randomization in agricultural experiments.

[12] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2023. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *arXiv preprint arXiv:2312.11970*.

[13] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *CoRR*, abs/2304.14767.

[14] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[15] Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

[16] Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *CoRR*, abs/2305.00586.

[17] Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. 2023. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

[18] Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. 2024. Autonomous llm-driven research from data to human-verifiable research papers. *CoRR*, abs/2404.17605.

[19] David Jenny, Yann Billeter, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. 2023. Exploring the jungle of bias: Political bias attribution in language models via dependency analysis. *CoRR*, abs/2311.08605.

[20] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: Assessing causal reasoning in language models. In *NeurIPS*.

[21] Zhijing Jin, Amir Feder, and Kun Zhang. 2022. CausalNLP tutorial: An introduction to causality for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 17–22, Abu Dubai, UAE. Association for Computational Linguistics.

[22] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net.

[23] Immanuel Kant. 1781. *Critique of Pure Reason*. Cambridge University Press.

[24] Katherine A. Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5332–5344. Association for Computational Linguistics.

[25] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

[26] Felix Leeb, Zhijing Jin, Luigi Gresele, and Bernhard Schölkopf. 2024. Systematically addressing the monsters under the bench(marks).

[27] Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *CoRR*, abs/2401.12874.

[28] Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Hassan Awadalla, and Weizhu Chen. 2024. Sciagent: Tool-augmented language models for scientific reasoning. *CoRR*, abs/2402.11451.

[29] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *arXiv preprint arXiv:2202.05262*.

[30] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: A survey. *CoRR*, abs/2302.07842.

[31] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

[32] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

[33] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

[34] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895.

[35] OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

[36] Judea Pearl. 2001. Direct and indirect effects. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, pages 411–420. Morgan Kaufmann.

[37] Judea Pearl. 2009. *Causality: Models, reasoning and inference (2nd ed.)*. Cambridge University Press.

[38] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2023, Raleigh, NC, USA, February 8-10, 2023*, pages 464–483. IEEE.

[39] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

[40] Donald B. Rubin. 1980. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

[41] Bertrand Russell. 2004. *History of western philosophy*. Routledge.

[42] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

[43] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *CoRR*, abs/2302.04761.

[44] Bernhard Schölkopf. 2022. Causality for machine learning. In Hector Geffner, Rina Dechter, and Joseph Y. Halpern, editors, *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36 of *ACM Books*, pages 765–804. ACM.

[45] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Towards causal representation learning. *CoRR*, abs/2102.11107.

[46] Peter Spirtes, Clark Glymour, and Richard Scheines. 1993. Causation, prediction, and search.

[47] Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press.

[48] Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schölkopf, and Mrinmaya Sachan. 2023. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

[49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

[50] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[51] Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. 2024. What are tools anyway? A survey from the language model perspective. *CoRR*, abs/2403.15452.

[52] Kun Zhang and Aapo Hyvärinen. 2009. Causality discovery with additive disturbances: An information-theoretical perspective. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20*, pages 570–585. Springer.

[53] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

[54] Huaixiu Steven Zheng, Swaroop Mishra, Hugh Zhang, Xinyun Chen, Minmin Chen, Azade Nova, Le Hou, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2024. Natural plan: Benchmarking llms on natural language planning. *CoRR*, abs/2406.04520.

[55] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.