Discovery, Validation and Editing of Large Language Model Mechanisms: Recent Advances and Future Perspectives

Yinhan He
nee7ne@virginia.edu
University of Virginia
Charlottesville, VA, USA

Wendy Zheng
ncd9cf@virginia.edu
University of Virginia
Charlottesville, VA, USA

Tianyi Zhao
abs4dj@virginia.edu
University of Virginia
Charlottesville, VA, USA

Chen Chen
zrh6du@virginia.edu
University of Virginia
Charlottesville, VA, USA

Jundong Li*
jl6qk@virginia.edu
University of Virginia
Charlottesville, VA, USA

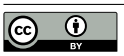
Abstract

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks, yet their internal mechanisms remain largely opaque, making it difficult to understand, predict, or control their behavior. As LLMs are increasingly deployed in high-stakes settings, this lack of transparency raises serious concerns about reliability and safety. Mechanistic interpretability (MI) has emerged as a promising approach to address this challenge, seeking to reverse-engineer the internal computations of LLMs into human-understandable mechanisms, i.e., an approximate high-level algorithm that the LLM implements with a subset of its components (a *circuit*) to complete a certain language task or exhibit a certain behavior. This tutorial provides a comprehensive and up-to-date overview of LLM mechanism discovery, validation, and editing. We begin by introducing foundational concepts, including features, components, computational graphs, and circuits, along with key notation. We then examine mechanism discovery through four methodological families: causal mediation, attribution, sparse decomposition, and optimization-based approaches. Next, we turn to mechanism validation, covering methods for verifying proposed mechanisms and emerging standards for rigorous evaluation. Building on these foundations, we survey mechanistic editing techniques that leverage MI insights to modify behavior at varying granularity, from fine-grained representation-level steering to coarser circuit-level interventions. Lastly, we outline open challenges and future research directions, including scalability of interpretability methods, evaluation benchmarks for mechanistic circuits, and the integration of interpretability with training-time objectives, aiming to inspire continued progress in understanding and governing large language models.

CCS Concepts

• **Computing methodologies** → **Knowledge representation and reasoning**.

*Jundong Li is the corresponding author.



KDD 2026, Jeju Island, Republic of Korea.

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3816458

Keywords

Mechanistic Interpretability; Large Language Models

ACM Reference Format:
Yinhan He, Wendy Zheng, Tianyi Zhao, Chen Chen, and Jundong Li. 2026. Discovery, Validation and Editing of Large Language Model Mechanisms: Recent Advances and Future Perspectives. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD 2026), August 9–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3770855.3816458

1 Introduction

Over the past few years, large language models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks, from question answering and natural language understanding to mathematical reasoning and code generation. However, their internal mechanisms remain largely opaque, making it difficult to understand, predict, or control their behavior. As these models are increasingly deployed in high-stakes settings, this lack of transparency raises serious concerns about reliability, safety, and accountability. To address these challenges, mechanistic interpretability (MI) has emerged as a promising approach that seeks to reverse-engineer the internal computations of LLMs into human-understandable mechanisms. A mechanism is defined as a high-level algorithm that the LLM implements with a subset of its components to complete a certain language task or exhibit a certain behavior. The specific subset of components that implements a mechanism is referred to as a *circuit*; throughout this tutorial, we use *mechanism* to emphasize the algorithmic behavior and *circuit* to emphasize its structural realization in the LLM architecture. As the field continues to evolve, increasing attention is being devoted to rigorously validating proposed explanations and translating interpretability insights into practical model interventions. In response, this tutorial covers the discovery, validation, and editing of LLM mechanisms, organized around six themes: (1) the motivation and significance of MI, highlighting why internal reasoning discovery is critical for high-stakes applications; (2) fundamental notions, including features, components, computational graphs, and circuits; (3) mechanism discovery methods, including causal mediation analysis, sparse feature decomposition, and optimization-based approaches; (4) mechanism validation via rigorous metrics and axiomatic approaches; (5) mechanistic editing, leveraging MI insights for factual correction, reasoning debugging, and bias mitigation; (6) current challenges and

future directions, including scalability and evaluation benchmarks. By promoting more reliable, controllable, and transparent language models, this tutorial aims to support safer real-world deployment and more accountable AI systems.

2 Tutorial Outline

- **I: Introduction: Mechanistic Interpretability in LLMs (25 mins)**
  - **Motivation & Significance:** Reasons why it is critical for LLMs to be applied to high-stakes scenarios and the necessity of internal reasoning discovery.
  - **Challenges:** The failure of traditional model explainability methods in the context of higher demand for LLM internal logics rather than simply explaining from data perspectives.
  - An overview of natural language tasks that have been studied in LLM mechanistic interpretability research.
  - An overview of the applications which can benefit from mechanistic interpretation, verification, and editing.
- **II: Notions and Background (25 mins)**
  - The necessity of moving beyond input-output attribution toward internal component-level understanding.
  - **Fundamental Notions:**
    - Concept of computational graph, consisting of nodes (attention heads, MLP layers, residual streams), and edges representing information flow.
    - Circuits as functional compositions of model components.
    - Important Research Objects in Mechanism Discovery (polysemanticity, superposition, and distributed functionalities)
  - **Causal Foundations:** Interventions, causal effects, and causal mediation analysis.
- **III: Mechanism Discovery (30 mins)**
  - **Causal Mediation:** Measuring component importance via targeted interventions [4, 21**?**].
  - **Attribution Methods:** Estimating component importance using approximations [6, 11, 20].
  - **Sparse Decomposition:** Disentangling activations into interpretable features [5, 13, 15, 17].
  - **Optimization-based:** Finding mechanisms through optimization [2, 3, 8].
- **IV: Mechanism Validation (30 mins)**
  - **Formalizing MI:** Axiomatic approaches [18] and circuit discovery with self-validated guarantees [1, 10].
  - **Validation Metrics:** Faithfulness, Minimality, and Completeness of mechanistic interpretations [**?**].
  - **Validation Tools:** causal inference–based techniques [7, 19] and MI benchmarks [12**?**]
- **V: Mechanism Editing (30 mins)**
  - **Knowledge & Reasoning Editing:** Principled factual correction [16, 23, 24] and reasoning debugging [14].
  - **Model Unlearning:** Mechanistic unlearning for privacy and representation-level steering [9, 22].
  - **Safety & Alignment:** Bias Mitigation [25]
- **VI: Challenges and Future Directions (20 mins)**
  - Summary of presented discovery frameworks, validation methodologies, and editing techniques.
  - Current challenges in mechanistic interpretability, including limitations and gaps in existing approaches
  - Future directions for advancing mechanistic interpretability and improving its practical usefulness and scalability

3 Target Audience and Prerequisites

The target audience includes researchers, industry practitioners, and students interested in data mining, machine learning, natural language processing, and trustworthy AI. Participants with prior experience in large language models will gain the most from the technical discussions, but the tutorial will be structured at an advanced undergraduate or graduate level so that it remains accessible to a broad KDD community. Both academic and industry attendees will be able to follow the key concepts and practical methodologies. All tutorial slides and supporting resources will be made publicly available after the conference. We anticipate 50 to 100 participants, who will gain an up-to-date understanding of recent progress in discovering, validating, and editing mechanisms in large language models, along with insights into their practical implications.

4 Tutors

In-person presenters: Yinhan He, Wendy Zheng, Tianyi Zhao, Chen Chen, and Jundong Li will attend KDD 2026 and present the tutorial in person. **Contributors:** No additional contributors beyond the tutors listed above. **Corresponding tutor:** Jundong Li (jl6qk@virginia.edu).

Yinhan He. Yinhan He is a fourth-year Ph.D. candidate in the Department of ECE at the University of Virginia. His research interests span large language models, agentic AI, interpretable and explainable AI, and graph machine learning, with a particular emphasis on mechanistic interpretability of foundation models. He has been studying LLM mechanism discovery, validation and editing for more than two years, and his work has been published in top-tier conferences. For example, his work on global-level mechanistic interpretability has been accepted to ICML 2025, and collaboration work on reasoning editing has been accepted to ICLR 2026.

Wendy Zheng. Wendy Zheng is a first-year PhD student in the Department of Computer Science at the University of Virginia. Her research focuses on mechanistic interpretability of large language models, seeking to uncover and rigorously evaluate the internal mechanisms that give rise to model capabilities and behaviors. She is broadly interested in developing scalable interpretability methods that remain effective as models grow in size and complexity. She has co-authored multiple works published at top-tier venues, including one work in NeurIPS focusing on implicit reasoning in LLMs and one work in ICML investigating the existence of modular circuits.

Tianyi Zhao. Tianyi Zhao is a first-year PhD student in the Department of Computer Science at the University of Virginia. She is broadly interested in the rigorous development of trustworthy AI systems, with a current focus on leveraging mechanistic interpretability to demystify, evaluate, and refine the internal mechanisms of large language models. Through this lens, she studies fundamental challenges such as hallucination and miscalibration,

while also advancing techniques such as model editing. She has multiple works in mechanistic interpretability in submission.

Chen Chen. Chen is currently an assistant professor in the Computer Science Department at the University of Virginia (UVA). Prior to that, she was a research assistant professor in Biocomplexity Institute at UVA, and a software engineer at Google. Chen got her Ph.D. degree from Arizona State University in 2019. Her research has been focusing on the connectivity of complex networks, which has been applied to address pressing challenges in various high-impact domains, including healthcare, bioinformatics, recommendation, and critical infrastructure systems. Her research has appeared in top-tier conferences (including NeurIPS, ICML, ICLR, KDD, AAAI, IJCAI, SIGIR, WSDM, ICDM, SDM, etc.), and prestigious journals (including PNAS, IEEE TKDE, ACM CSUR, ACM TKDD, KAIS, and SIAM SAM). Chen has received several awards, including "Bests of KDD", "Bests of SDM" and Rising Star in EECS.

Jundong Li. Jundong Li is an Associate Professor at the University of Virginia with appointments in the Department of Electrical and Computer Engineering and the Department of Computer Science. His research interests span data mining, machine learning, and artificial intelligence, with a particular emphasis on graph machine learning, trustworthy and safe machine learning, and large language models. He has published more than 200 papers in high-impact venues, and his work has received over 20,000 citations. He has been recognized with several notable honors, including four early career awards—ICDM Tao Li Award (2025), SIGKDD Rising Star Award (2024), PAKDD Early Career Research Award (2023), and the NSF CAREER Award (2022). He has also received two best paper awards, namely the PAKDD Best Paper Award (2024) and the SIGKDD Best Research Paper Award (2022), as well as multiple industry faculty research awards.

5 Related Tutorials

The related representative tutorials are listed as follows:

(1) Hosein Mohebbi, Jaap Jumelet, Michael Hanna, Afra Alishahi, and Willem Zuidema. "Transformer-specific Interpretability." EACL 2024.
- **Date:** Mar. 21, 2024
- **Location:** Malta
- **Similarities:** It similarly covers causal mediation analysis as a method for circuit discovery in transformers to identify components responsible for model behavior.
- **Differences:** We offer an in-depth analysis of MI, covering a wider range of methods and findings beyond activation patching. We also dedicate coverage to the validation of mechanistic interpretations, the application of MI in model editing, and more recent advances in the field.

(2) Ziyu Yao, and Daking Rai. "Tutorial on Mechanistic Interpretability for Language Models." ICML 2025.
- **Date:** July 14, 2025
- **Location:** Vancouver, Canada
- **Similarities:** This tutorial similarly provides a broad overview of mechanistic interpretability in LLMs, covering foundational concepts and representative works, as well as practical applications of MI.
- **Differences:** Our tutorial incorporates more recent advances in the field. We additionally cover the validation of mechanistic interpretations, a critical and often overlooked step in achieving reliable interpretability, and offer deeper coverage of MI-based model editing.

(3) Eliana Pastor, Eleonora Poeta, André Panisson, Alan Perotti, and Gabriele Ciravegna. "Beyond Input Attribution: A Hands-On Tutorial to Concept-Based Explainable AI and Mechanistic Interpretability." KDD 2025.
- **Date:** Aug. 4, 2025
- **Location:** Toronto, Canada
- **Similarities:** It also covers sparse autoencoders as a key approach in MI, including specific frameworks that apply them to interpret the model's internal representations.
- **Differences:** Our tutorial provides a broader introduction to MI, covering a wider range of approaches beyond sparse autoencoders. We also address the validation of mechanistic interpretations and provide dedicated coverage of MI for model editing, along with recent advances in the field.

(4) Shichang Zhang, Himabindu Lakkaraju, and Julius Adebayo. "Explain AI Models: Methods and Opportunities in Explainable AI, Data-Centric AI, and Mechanistic Interpretability." NeurIPS 2025.
- **Date:** Dec. 2, 2025
- **Location:** San Diego, CA, USA
- **Similarities:** This tutorial introduces MI as a means of attributing LLM behavior to internal components, with a focus on causal mediation analysis as a core technique.
- **Differences:** Our tutorial offers a more comprehensive overview of MI to cover a broader set of methods and findings. We further address the validation of mechanistic interpretations and the application of MI in model editing, as well as the latest developments.

6 Audience Participation and Interactivity

We will be open to participation of all audiences, including allowing the audiences to ask questions throughout this tutorial. Besides, we will schedule a 10-minute break after sections II and IV, which provides opportunities for attendees to refresh, take notes, and interact with the presenters promptly. Moreover, we will also be active in replying any other follow-up questions after the tutorial.

7 Societal Impacts

This tutorial provides a comprehensive overview of recent advances in mechanistic interpretability, focusing on mechanistic discovery, causal verification, and responsible model editing. We aim to encourage the community to move beyond black-box usage toward more interpretable and accountable AI systems. We hope this tutorial will motivate researchers and practitioners to incorporate mechanistic analysis into the development of LLMs. In addition, the tutorial introduces a range of practical techniques for diagnosing, verifying, and modifying model behaviors, which may foster cross-disciplinary collaboration between the data mining, NLP, and AI safety communities. Finally, by promoting more reliable, controllable, and transparent language models, this tutorial can help support safer real-world deployment of LLMs.

References

[1] Alaa Anani, Tobias Lorenz, Bernt Schiele, Mario Fritz, and Jonas Fischer. 2026. Certified Circuits: Stability Guarantees for Mechanistic Circuits. *arXiv preprint arXiv:2602.22968* (2026).

[2] Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. 2024. Finding transformer circuits with edge pruning. *Advances in Neural Information Processing Systems* 37 (2024), 18506–18534.

[3] Tian Bian, Yifan Niu, Chaohao Yuan, Chengzhi Piao, Bingzhe Wu, Long-Kai Huang, Yu Rong, Tingyang Xu, Hong Cheng, and Jia Li. 2025. IBCircuit: Towards Holistic Circuit Discovery with Information Bottleneck. In *International Conference on Machine Learning*. PMLR, 4289–4302.

[4] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems* 36 (2023), 16318–16352.

[5] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems* 37 (2024), 24375–24410.

[6] Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 17432–17445.

[7] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in neural information processing systems* 34 (2021), 9574–9586.

[8] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*. PMLR, 160–187.

[9] Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. 2025. Mechanistic Unlearning: Robust Knowledge Unlearning and Editing via Mechanistic Localization. In *International Conference on Machine Learning*. PMLR, 20964–20992.

[10] Itamar Hadad, Guy Katz, and Shahaf Bassan. 2026. Formal Mechanistic Interpretability: Automated Circuit Discovery with Provable Guarantees. In *The Fourteenth International Conference on Learning Representations*.

[11] Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms. In *First Conference on Language Modeling*.

[12] Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8669–8687.

[13] Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, Vol. 2024. 7827–7845.

[14] Zhenyu Lei, Qiong Wu, Jianxiong Dong, Yinhan He, Emily Dodwell, Yushun Dong, and Jundong Li. 2026. Reforming the Mechanism: Editing Reasoning Patterns in LLMs with Circuit Reshaping. In *The Fourteenth International Conference on Learning Representations*.

[15] Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. 2024. Sparse Crosscoders for Cross-Layer Features and Model Diffing. *Transformer Circuits Thread* (2024). https://transformer-circuits.pub/2024/crosscoders/index.html.

[16] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems* 35 (2022), 17359–17372.

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

[18] Nils Palumbo, Ravi Mangal, Zifan Wang, Saranya Vijayakumar, Corina S Pasareanu, and Somesh Jha. 2025. Validating Mechanistic Interpretations: An Axiomatic Approach. In *International Conference on Machine Learning*. PMLR, 47509–47544.

[19] Alan Sun and Mariya Toneva. 2026. Tracking Equivalent Mechanistic Interpretations Across Neural Networks. In *The Fourteenth International Conference on Learning Representations*.

[20] Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. 407–416.

[21] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems* 33 (2020), 12388–12401.

[22] Hanqi Yan, Hainiu Xu, Siya Qi, Shu Yang, and Yulan He. 2026. When Thinking Backfires: Mechanistic Insights into Reason-induced Misalignment. In *The Fourteenth International Conference on Learning Representations*.

[23] Jiayu Yang, Yuxuan Fan, Songning Lai, Shengen Wu, Jiaqi Tang, Chun Kang, Zhijiang Guo, and Yutao Yue. 2026. ACE: Attribution-Controlled Knowledge Editing for Multi-hop Factual Recall. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=IuWIzmMvKo

[24] Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *Advances in neural information processing systems* 37 (2024), 118571–118602.

[25] Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. 2024. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 7943–7956.

Received 13 March 2026