Fellowship Proposal: Towards a new paradigm for interpretability

Michael Hanna

1 Summary

Interpreting large language models (LLMs)—i.e., making their internal mechanisms understandable to an external observer—promises benefits for a wide variety of stakeholders. Interpretability should allow developers to identify weaknesses in models, users to build trust in models, and scientists to learn about human language from models. However, interpretability does not currently come easy: most interpretability studies rely on bespoke setups, and the most robust interpretability techniques available are brittle and difficult to use, leading users and researchers to rely on less reliable methods. In my PhD, I have worked to solve this problem by growing the *circuits* paradigm of interpretability, a general causal framework for finding the set of components that underlies a model's behavior on a given task. As part of this, I have worked on making circuit-finding more accurate, scalable, and accessible, via papers, tutorials, and contributions to a circuit-finding library and benchmark.

Now, a new technique called *feature circuits* promises to uncover not just the components, but the intermediate concepts and computations that underlie model behavior. This highly general technique could provide users with interpretable explanations, and robustly answer questions that earlier techniques struggled to address, such as whether LLMs plan, or how they generate unfaithful chains of thought; however, it is technically involved and compute-intensive. Going forward, I plan to apply lessons learned from component circuits to advance the emerging feature circuits paradigm, by: 1) Building open-source resources and libraries that facilitate work on feature circuits. 2) Constructing standardized evaluations to ensure correctness of found feature circuits. 3) Integrating with publicly-available user-friendly resources to enable public engagement with interpretability.

2 The challenge of interpretability

From early cognitive scientists trying to understand the mechanisms learned by simple neural networks (Rumelhart et al., 1993), to researchers studying the first LLMs (Rogers et al., 2020), interpretability has long played a central role in artificial intelligence research. Yet despite this long history and potential benefits, the impact of (mechanistic) interpretability, the subfield of interpretability interested in uncovering model mechanisms, has been modest. One challenge that has blunted interpretability's impact is the abundance of methods that produce interpretability illusions; interpretability often produces explanations that look correct, but are misleading, and do not faithfully reflect models' internal mechanisms. (Adebayo et al., 2018; Bilodeau et al., 2024).

In response, causal interpretability techniques have grown in popularity (Vig et al., 2020; Geiger et al., 2021). These methods use causal interventions to manipulate model internals while observing model output, allowing one to causally prove a connection between a hypothesized internal mechanism and external behavior. These techniques are powerful, yet challenging to use: unlike the probing paradigm, which involved a fairly standard setup across research questions, causal interventions are a broad family of techniques, meaning that the appropriate causal experiment for each research question could vary widely. This lack of a unified paradigm for causal interpretability has stalled its adoption by scientists and practitioners who are not interpretability experts (Calderon and Reichart, 2025).

3 Circuits: A causal answer to questions in interpretability

Reacting to this missing paradigm for causal interpretability, *circuits* (Olah et al., 2020) for LLMs have emerged to fill the gap (Wang et al., 2023). Given a model and task, circuit analysis aims to isolate the *circuit*, or set of components (attention heads and multi-layer perceptrons), responsible for the model's behavior on the task. Circuit analyses then causally verify the circuit by ablating all parts

of the model outside the circuit. Crucially, unlike past causal work, the circuits framework is general: given any model, task, and metric satisfying certain conditions, one can find a circuit.

This generality has allowed me to study many different phenomena using circuits. For example, in early work on LLM circuits, I used circuit analysis to explain how GPT-2 small performs the greater-than operation (Hanna et al., 2023), the first application of circuits to math. However, in later work, I helped apply this framework to localize model components responsible for gender bias (Chintam et al., 2023) and how model mechanisms emerge and change over training (Tigges et al., 2024). Most recently, I applied circuit analysis to questions from linguistics, my own field of interest: using circuits to localize model areas responsible for different linguistic abilities, I showed how models' implementations of these abilities may differ from language in the human brain (Hanna et al., 2025).

That said, a general interpretability paradigm is only valuable if it is robust and scalable. To this end, I have worked to not only measure and improve circuit-finding techniques' accuracy (Hanna et al., 2024), but also to boost their scalability, building a library aimed at enabling circuit-finding in modern open-source models; this library forms the backbone of the Mechanistic Interpretability Benchmark's circuit track (Mueller et al., 2025), a new benchmark for comparing circuit-finding methods.

In summary, component circuits are a robust, general, and growing paradigm in causal interpretability, in which I am deeply invested. However, component circuits only determine the components of a model that perform a task, without explaining each component's semantics; they thus fall short of a full mechanistic explanation. For this, we need a new technique: feature circuits.

4 Feature circuits

In the remaining years of my PhD, I plan to focus on feature circuits (Marks et al., 2025), which capture the features model use to produce their behaviors. Unlike larger model components, such as attention heads and multi-layer perceptrons, which play many roles and are hard to interpret, features are monosemantic: they represent single model concepts. Feature circuits thus allow practitioners to glean fine-grained insights into model mechanisms: for example, in recent work, I used feature circuits to study how models build up complex grammatical concepts when processing syntactically ambiguous sentences (Figure 1; Hanna and Mueller, 2025).

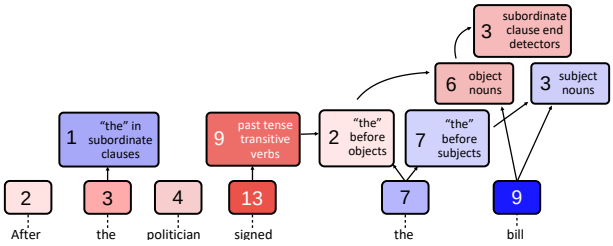

Figure 1: A feature circuit explaining how a model processes the ambiguous input "After the politician signed the bill". Red features push the model to interpret "bill" as a grammatical object, and output ","; blue ones promote reading it as a subject, promoting "*was*". In the coming years, I intend to build on such feature circuits, which, in contrast to component circuits, reveal both the important parts of models and their semantics.

Feature circuits are highly promising. Their ability to capture the structure and semantics of model mechanisms makes them powerful tools, and they can be applied to any LLM input. This not only makes them useful for researchers, but also allows them to provide interpretable explanations for the general public. However, they require expensive auxiliary models, suffer from a lack of standardized evaluations, and require technically complex methods to find. Drawing on my experience with component circuits, I propose the following solutions, to make feature circuits a robust interpretability paradigm for everyone.

4.1 Open-source resources for feature circuits

Finding feature circuits requires open-source resources, because the monosemantic features that comprise them are not normal model components or neurons. Rather, features come from sparse autoencoders (SAEs), auxiliary models that decompose dense LLM representations into higher-dimensional sparse representations; each feature in a feature circuit is an (ideally) interpretable and causally-relevant dimension of such a representation. Unfortunately, SAE training is both computationally expensive and technically challenging, being sensitive to hyperparameters that are often poorly justified and documented. It is thus difficult for most practitioners to train SAEs, even on small models.

In an ongoing collaboration in my capacity as an Anthropic Research Fellow, I aim to alleviate this bottleneck by creating open-source tools for feature circuits. First, we are developing a library for feature circuit finding, using methods recently proposed by Ameisen et al. (2025) and Lindsey et al. (2025). This will allow practitioners with even modest resources to find feature circuits in small models in minutes. Second, my collaborators and I intend to train new auxiliary models (SAEs and related models called transcoders) that will enable researchers to apply feature circuits to state-of-the-art open models—even if they cannot afford to train transcoders on them. However, we also plan to release training code, for those with the resources necessary to train them.

4.2 An evaluation benchmark for feature circuits

Although feature circuits are a promising new technique for interpretability, assessing their correctness is challenging. Existing circuit evaluations, intended for use with coarser-grained component circuits, focus on *faithfulness*, the idea that model performance should remain the same, even when all non-circuit components are ablated. This transfers poorly to much-sparser feature circuits: small, interpretable circuits as in Figure 1 may be causally relevant to model processing, but may not maintain the model's behavior when all other features are ablated. As a result, feature circuits are often validated by intervening on each node of the circuit, and checking that the model's behavior changes in the way implied by the node's semantics (Marks et al., 2025; Hanna and Mueller, 2025). This achieves the goal of validating the circuit, but is a non-standardized, ad-hoc procedure.

I propose to develop a new evaluation benchmark for feature circuits that standardizes their evaluation, building on past work on evaluating component circuits (Gupta et al., 2024; Mueller et al., 2025) and feature-finding methods (Huang et al., 2024; Karvonen et al., 2025). This benchmark will have three parts. The first part, following Gupta et al. (2024), will test feature-circuit-finding methods on toy models trained to use specific feature circuits to solve a given task. This allows us to compare found feature-circuits with our known ground-truth feature circuit. The second part of the benchmark will evaluate techniques that find feature-circuits by estimating the impact that ablating a given feature will have on the model's logits or other features. We can do this by computing the true impact of ablating each model feature; this is expensive, but feasible in small models, for the purpose of benchmarking. The final part of the benchmark will develop a standardized set of tasks for real-world models, and use causal interventions to evaluate whether a proposed feature circuit for those tasks is effective in steering LLM behavior in the way suggested by their semantic annotations.

4.3 Interpretability for everyone

The preceding two proposals target researchers with modest computational resources, studying small models. But what of researchers who have fewer resources, or wish to target larger models; or non-researchers, who cannot engage with scientific tools? I plan to reach these communities by tapping into the rich ecosystem of existing tools for publicly accessible interpretability. To solve the compute issue, I will port the feature circuits library discussed in Section 4.1 to NNsight, an interpretability library that is uniquely integrated with a compute cluster that is freely available to researchers (Fiotto-Kaufman et al., 2025). This is non-trivial, as NNsight still lacks some features necessary for feature circuits, but I am actively collaborating with NNsight developers to add these. NNsight support will allow researchers to utilize feature-circuits without any computational resources of their own.

For members of the public, however, an even more accessible and user-friendly interface for feature circuits is needed. We thus plan to integrate the circuit-tracing library with Neuronpedia (Lin, 2023), a site providing a graphic user interface for visualizing features. The end goal of this integration is to allow users to submit a prompt on Neuronpedia, which will then be run on a remote computing cluster, and visualized on Neuronpedia. This will allow everyday users to obtain feature-circuit-based explanations of model behavior without any technical knowledge.

Finally, having created these resources, I hope to engage with the public, making researchers and laymen alike aware of feature circuits. For component circuits, I have given an EACL tutorial and various tutorial-style talks on how to work with circuits (Mohebbi et al., 2024); as more beginner-friendly interpretability content, I recorded a short video about circuits, along with colleagues. Building on these efforts, by creating more educational content and presenting feature circuits at more public-facing and interdisciplinary venues, will help make interpretability more accessible for all.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9525–9536, Red Hook, NY, USA. Curran Associates Inc.

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Ben Thompson, T., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. (2025). Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Bilodeau, B., Jaques, N., Koh, P. W., and Kim, B. (2024). Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120.

Calderon, N. and Reichart, R. (2025). On behalf of the stakeholders: Trends in NLP model interpretability in the era of LLMs. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 656–693, Albuquerque, New Mexico. Association for Computational Linguistics.

Chintam, A., Beloch, R., Zuidema, W., Hanna, M., and van der Wal, O. (2023). Identifying and adapting transformer-components responsible for gender bias in an English language model. In Belinkov, Y., Hao, S., Jumelet, J., Kim, N., McCarthy, A., and Mohebbi, H., editors, *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 379–394, Singapore. Association for Computational Linguistics.

Fiotto-Kaufman, J. F., Loftus, A. R., Todd, E., Brinkmann, J., Pal, K., Troitskii, D., Ripa, M., Belfki, A., Rager, C., Juang, C., Mueller, A., Marks, S., Sharma, A. S., Lucchetti, F., Prakash, N., Brodley, C. E., Guha, A., Bell, J., Wallace, B. C., and Bau, D. (2025). NNsight and NDIF: Democratizing access to open-weight foundation model internals. In *The Thirteenth International Conference on Learning Representations*.

Geiger, A., Lu, H., Icard, T., and Potts, C. (2021). Causal abstractions of neural networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Gupta, R., Arcuschin, I., Kwa, T., and Garriga-Alonso, A. (2024). Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hanna, M., Belinkov, Y., and Pezzelle, S. (2025). Are formal and functional linguistic mechanisms dissociated in language models?

Hanna, M., Liu, O., and Variengien, A. (2023). How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 76033–76060. Curran Associates, Inc.

Hanna, M. and Mueller, A. (2025). Incremental sentence processing mechanisms in autoregressive transformer language models. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3181–3203, Albuquerque, New Mexico. Association for Computational Linguistics.

Hanna, M., Pezzelle, S., and Belinkov, Y. (2024). Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*.

Huang, J., Wu, Z., Potts, C., Geva, M., and Geiger, A. (2024). RAVEL: Evaluating interpretability methods on disentangling language model representations. In Ku, L.-W., Martins, A., and Srikumar,

V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8669–8687, Bangkok, Thailand. Association for Computational Linguistics.

Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J., Chanin, D., Lau, Y.-T., Farrell, E., McDougall, C., Ayonrinde, K., Wearden, M., Conmy, A., Marks, S., and Nanda, N. (2025). Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability.

Lin, J. (2023). Neuronpedia: Interactive reference and tooling for analyzing neural networks. Software available from neuronpedia.org.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. (2025). On the biology of a large language model. *Transformer Circuits Thread*.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. (2025). Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.

Mohebbi, H., Jumelet, J., Hanna, M., Alishahi, A., and Zuidema, W. (2024). Transformer-specific interpretability. In Mesgar, M. and Loáiciga, S., editors, *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.

Mueller, A., Geiger, A., Wiegreffe, S., Arad, D., Arcuschin, I., Belfki, A., Chan, Y. S., Fiotto-Kaufman, J., Haklay, T., Hanna, M., Huang, J., Gupta, R., Nikankin, Y., Orgad, H., Prakash, N., Reusch, A., Sankaranarayanan, A., Shao, S., Stolfo, A., Tutek, M., Zur, A., Bau, D., and Belinkov, Y. (2025). Mib: A mechanistic interpretability benchmark.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*. https://distill.pub/2020/circuits/zoom-in.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rumelhart, D. E., Todd, P. M., et al. (1993). Learning and connectionist representations. *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, 2:3–30.

Tigges, C., Hanna, M., Yu, Q., and Biderman, S. (2024). LLM circuit analyses are consistent across training and scale. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. (2023). Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.