Discovering Transformer Circuits via a Hybrid Attribution and Pruning Framework

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Interpreting language models often involves circuit analysis, which aims to identify sparse subnetworks, or *circuits*, that accomplish specific tasks. Existing circuit discovery algorithms face a fundamental trade-off: attribution patching is fast but unfaithful to the full model, while edge pruning is faithful but computationally expensive. This research proposes a hybrid attribution and pruning (HAP) framework that uses attribution patching to identify a high-potential subgraph, then applies edge pruning to extract a faithful circuit from it. We show that HAP is 46% faster than baseline algorithms without sacrificing circuit faithfulness. Furthermore, we present a case study on the Indirect Object Identification task, showing that our method preserves cooperative circuit components (e.g. S-inhibition heads) that attribution patching methods prune at high sparsity. Our results show that HAP could be an effective approach for improving the scalability of mechanistic interpretability research to larger models[1].

1 Introduction

Large language models (LLMs) are increasingly being deployed in high-stakes settings, motivating the need to uncover their "black-box" Alishahi et al. [2019] nature and understand how they "think." Hubinger [2020], Zhang et al. [2021] This is a key goal of mechanistic interpretability, a field focused on understanding transformer Vaswani et al. [2017] model behavior by analyzing the interactions between subnetworks of attention heads and multi-layer perceptrons (MLPs)Vig et al. [2020], Sharkey et al. [2025]. The most common approach to mechanistic interpretability is through circuit analysis, which identifies sparse subnetworks, or "circuits", responsible for specific behaviors Olah et al. [2020], Olah, Erdogan [2025]. Manual circuit discovery methods, such as that proposed by Wang et al. [2022], have largely been replaced by automated approaches like Automated Circuit DisCovery (ACDC) Conmy et al. [2023], which uses a greedy search algorithm to ablate edges one by one.

To address the computational cost of ACDC, faster algorithms such as Edge Attribution Patching (EAP) Syed et al. [2023] and Edge Pruning (EP) Bhaskar et al. [2024] have been proposed (see Section 2). However, existing circuit discovery algorithms struggle to scale with larger models without sacrificing performance Hanna et al. [2024], Hsu et al. [2025], Zhang et al. [2025]. EAP uses a first-order Taylor series approximation to ablate all edges simultaneously. Although faster than ACDC, the first-order approximations show low faithfulness to the full model. On the other hand, EP efficiently applies a gradient-based pruning algorithm to discover circuits in parallel. Despite scaling well to larger models while maintaining exceptional circuit faithfulness, EP requires significant compute power.

[1]Our code is available at: `https://anonymous.4open.science/r/HAP-circuit-discovery`

This research proposes a novel Hybrid Attribution and Pruning (HAP) framework to enhance the scalability and maintain the faithfulness of discovered circuits. We leverage EAP to quickly filter out the majority of unimportant edges. This EAP-identified subgraph gives a narrowed search space for EP to find faithful circuits.

In summary, our main contributions are the following:

1. We propose a novel framework (HAP) that improves efficiency and preserves the faithfulness of discovered circuits.
2. We show that HAP matches or outperforms existing methods in efficiency and faithfulness.
3. We demonstrate in an IOI case study that HAP finds the often-missed S-Inhibition heads, preserving the quality of discovered circuits.

2 Related Works

Automated Circuit Discovery Algorithms such as ACDC construct computational graphs where nodes represent model components and edges represent information flow Conmy et al. [2023]. ACDC recursively applies activation patching—replacing activations with those from "corrupted" examples—removing edges that do not degrade task metric performance Syed et al. [2023]. This greedy search can rediscover known circuits, but it is computationally expensive for larger models or datasets due to the requirement for many forward passes, with scalability limited by the number of edges evaluated Conmy et al. [2023].

Edge Pruning and Optimization-based Methods frame circuit discovery as a gradient-based optimization problem, where edges between components of a model's computational graph are pruned using binary masks over edges Bhaskar et al. [2024]. This method allows for finer-grained and more faithful recovery of causal pathways, but requires architectural modifications and additional memory for scalability. EP can parallelize training across multiple GPUs, which enables EP to scale to large models (e.g., CodeLlama-13B) and complex datasets, recovering circuits that are both smaller and more interpretable than those produced by prior methods Bhaskar et al. [2024].

Attribution and Gradient-based Approximations, like EAP, propose gradient-based, first-order approximations to activation patching Syed et al. [2023], enabling simultaneous computation of edge importance scores with one backward and two forward passes. EAP efficiently identifies circuits that align closely with those found by ACDC, as measured by ROC/AUC when compared to manually curated circuit ground-truths, but can miss critical component interactions due to its linear approximation and reduced faithfulness Bhaskar et al. [2024].

3 Methods

The HAP framework operates by leveraging EAP to perform a global search, quickly removing low-importance edges to isolate higher-importance edges. This EAP-identified subgraph gives a narrowed search space for the precise pruning algorithm, EP.

3.1 Step 1: Computational Graph Construction

We start by representing our model as a computational graph following the convention of Bhaskar et al. [2024], where components of the Transformer architecture, namely attention layers and MLPs, are the nodes and the edges between any two nodes represent the connection between the output of one node to the input of the other node. The full model, in our case GPT-2 Small (from Radford et al. [2019], Maintainers [2022]), can be represented at this granularity, and a circuit is a computational subgraph consisting of a set of edges that describe the full model's behavior on a particular task (see Section 4.1.

3.2 Step 2: Edge Attribution Patching

We then use Edge Attribution Patching to quickly get absolute attribution scores that measure the importance of all edges in the computational graph using:

$$L(\mathbf{x} \mid e_{\text{ablated}}) - L(\mathbf{x}) \;\approx\; (e_{\text{clean}} - e_{\text{ablated}})^\top \, \frac{\partial L(\mathbf{x} \mid e_{\text{clean}})}{\partial e_{\text{clean}}} \tag{1}$$

where $L(\mathbf{x})$ is the logit loss, e_{ablated} denotes predictions after ablation of the target edge, and the right side of Equation (1) represents the computed absolute attribution score Syed et al. [2023]. After ranking the scores, we keep the top-k edges for further processing.

3.3 Step 3: Subgraph Selection and Edge Pruning

From here, edges with low attribution scores are masked to produce a high-potential subgraph. The masking threshold balances sparsity against the retention of potentially cooperative but weakly attributed components (e.g. S-inhibition heads). Using the EAP-filtered subgraph, we use the edges found to "jumpstart" the EP training process. EP proceeds by optimizing a binary mask $z \in [0, 1]^{N_{\text{edge}}}$ to minimize output divergence between the original and pruned graphs, under a targeted sparsity constraint:

$$1 - \frac{|H|}{|G|} \geq c \tag{2}$$

This step is performed via gradient-based optimization using clean and corrupted examples Bhaskar et al. [2024].

4 Experiment

4.1 Task Description

The task being studied is defined by a set of prompts that elicit a clearly defined response from the model predictions. We study the Indirect Object Identification (IOI) task, which is in the general format of "*When Dylan and Ryan went to the store, Dylan gave a popsicle to → Ryan*". We use Wang et al. [2022]'s prompt templates to generate an IOI dataset of 200 randomly selected examples with lexical and syntactic diversity, each for training and validation. Our test split involved 36,084 examples as per Bhaskar et al. [2024].

4.2 Experimental Setup

We evaluate our Hybrid Attribution and Pruning (HAP) framework on the Indirect Object Identification (IOI) task using GPT-2 Small (117M). The attribution score threshold in EAP is set very low to preserve possibly cooperative edges that might score low individually. For EP, we use the hyperparameters as detailed in Bhaskar et al. [2024]. All training runs were performed on one NVIDIA H100 GPU. We quantify circuit quality with faithfulness via KL divergence and logit difference between model predictions and circuit predictions, and report standard metrics such as accuracy and runtime.

5 Results

5.1 HAP vs Existing Methods

To compare the performance between different models, we leverage manually discovered circuits in Wang et al. [2022] as a reference to calculate the accuracy of circuits recovered by automatic methods. As shown in Table 1, HAP outperforms EAP in accuracy while having only slightly lower accuracy compared to EP. Similarly, circuits recovered by HAP are much more faithful to the full model compared to EAP (when comparing logit difference), while also maintaining similar faithfulness to EP circuits. It is shown in both KL divergence and logit difference metrics that HAP circuits are only slightly less faithful than EP circuits.

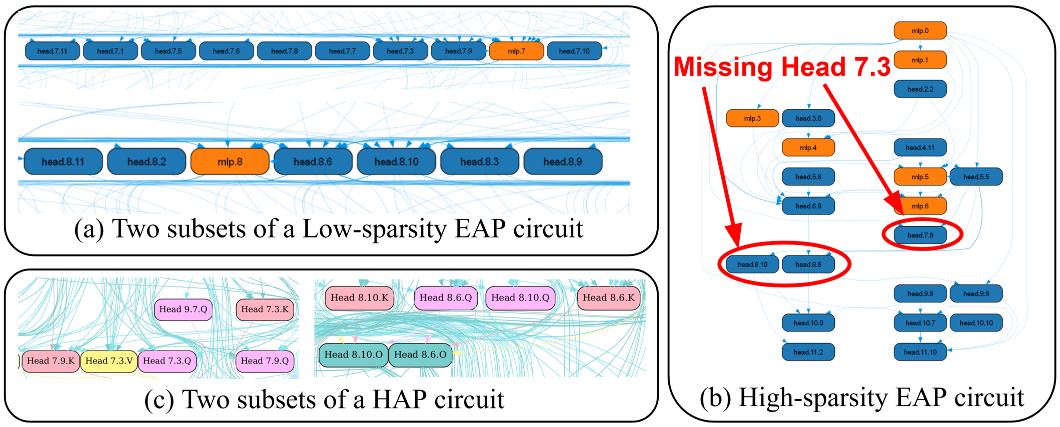
When GPU and target sparsity is controlled, HAP is at least 46% faster than EP while maintaining high accuracy and faithfulness to the full model. This shows that HAP can be a valuable framework for reducing the computational cost of circuit discovery, possibly enabling scalability to larger models.

5.2 Case Study: S-inhibition Heads in IOI

To present the qualitative advantages of our hybrid framework, we present a case study on the IOI task in GPT-2 Small (see Section 4.1). In IOI, the role of S-inhibition heads (or Subject-Inhibition

Algorithm	Sparsity	GPT-2 Small			
		Accuracy ↑	**Logit Diff** ↑	**KL** ↓	**Runtime** (s) ↓
EAP	94±0.5%	0.698	3.13	–	4
EP	94±0.5%	0.772	3.48	0.190	2921
HAP	94±0.5%	0.759	3.42	0.188	1579

Table 1: Efficiency of HAP compared to existing works.



(a) Two subsets of a Low-sparsity EAP circuit

(b) High-sparsity EAP circuit

(c) Two subsets of a HAP circuit

Figure 1: Recovered IOI circuits. While EAP on its own is unable to recover all S-Inhibition Heads at high sparsity ($threshold = 0.002$), HAP preserves S-Inhibition Heads because it only uses attribution patching at low sparsity ($threshold = 4.6 * 10^{-6}$)(See Section 3.2).

Heads) is cooperative: they suppress the Name Mover Heads from incorrectly flagging the subject of a sentence due to their proximity to the verb. Thus S-Inhibition Heads, although critical for accurate task performance, are difficult to detect due to the low individual importance assigned by methods like Syed et al. [2023] at high sparsity.

We found that existing methods do not recover the complete circuit. For example, EAP falls short since S-inhibition heads do not receive high attribution scores, causing them to be undervalued as shown in Figure 1B. In contrast, HAP successfully captures the complete, functional circuit. By first using EAP to define a constrained search space with a generous threshold, we created a "safe zone" that retains these S-inhibition heads despite their low individual scores. The subsequent EP algorithm, operating on this focused and less noisy subgraph, correctly identifies their cooperative importance. As shown in Figure 1C, the Name Mover and S-inhibition heads, including heads 7.3, 7.9, 8.6, and 8.10, are all preserved by HAP. This serves as qualitative evidence that our method is not merely efficient, but also preserves cooperative components that are missed by prior approaches.

6 Limitations

Our experiments are conducted exclusively on the IOI task with the GPT-2 Small model. Although this task is a well-established benchmark for mechanistic interpretability, further evaluation on a broader set of models and tasks is necessary to assess the generality, robustness, and scalability of HAP. Furthermore, the current implementation has not optimized the threshold to select edges during the EAP stage, which will require future hyperparameter tuning. We also acknowledge that variations in the generated training dataset may result in minor performance differences across different runs.

7 Conclusion

We introduce HAP, a hybrid framework that resolves the longstanding speed-faithfulness tradeoff in circuit discovery by strategically sequencing EAP, a fast and approximate algorithm, with EP, a fine-grained and precise one. Our experiments show this approach is not only 46% faster than EP while maintaining comparable faithfulness, but is also qualitatively superior. As demonstrated in our IOI case study, HAP successfully preserves the S-inhibition heads that attribution methods fail to recover in isolation. The results challenge the notion that the speed-faithfulness trade-off is fundamental and provide a simple framework to scale up future mechanistic interpretability research to interpret larger models.

References

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, 25(4):543–557, 2019. doi: 10.1017/S135132491900024X.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding Transformer Circuits with Edge Pruning, April 2024. URL `http://arxiv.org/abs/2406.16778`. arXiv:2406.16778 [cs].

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability, October 2023. URL `http://arxiv.org/abs/2304.14997`. arXiv:2304.14997 [cs].

Ege Erdogan. Automated Circuit Discovery for Mechanistic Interpretability. February 2025.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms, July 2024. URL `http://arxiv.org/abs/2403.17806`. arXiv:2403.17806 [cs].

Aliyah R. Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Y. Odisho, Peter R. Carroll, and Bin Yu. Efficient Automated Circuit Discovery in Transformers using Contextual Decomposition, March 2025. URL `http://arxiv.org/abs/2407.00886`. arXiv:2407.00886 [cs].

Evan Hubinger. An overview of 11 proposals for building safe advanced ai, 2020. URL `https://arxiv.org/abs/2012.07532`.

HF Canonical Model Maintainers. gpt2 (revision 909a290), 2022. URL `https://huggingface.co/gpt2`.

Chris Olah. Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases. URL `https://www.transformer-circuits.pub/2022/mech-interp-essay`.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001, March 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001. URL `https://distill.pub/2020/circuits/zoom-in`.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL `https://api.semanticscholar.org/CorpusID:160025533`.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL `https://arxiv.org/abs/2501.16496`.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution Patching Outperforms Automated Circuit Discovery, November 2023. URL `http://arxiv.org/abs/2310.10348`. arXiv:2310.10348 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL `https://arxiv.org/abs/1706.03762`.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020. URL https://arxiv.org/abs/2004.12265.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. September 2022. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Lin Zhang, Wenshuo Dong, Zhuoran Zhang, Shu Yang, Lijie Hu, Ninghao Liu, Pan Zhou, and Di Wang. EAP-GP: Mitigating Saturation Effect in Gradient-based Automated Circuit Identification, February 2025. URL http://arxiv.org/abs/2502.06852. arXiv:2502.06852 [cs].

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. doi: 10.1109/TETCI.2021.3100641.

A IOI Dataset Generation

In Table 2, we provide the full set of IOI templates from Wang et al. [2022] used to generate our dataset described in Section 4.1. Names were sampled from a list of 100 common English first names, while places and objects were selected from a curated set of 20 frequent options.

IOI prompt templates
Then, [B] and [A] went to the [PLACE]. [B] gave a [OBJECT] to [A]
Then, [B] and [A] had a lot of fun at the [PLACE]. [B] gave a [OBJECT] to [A]
Then, [B] and [A] were working at the [PLACE]. [B] decided to give a [OBJECT] to [A]
Then, [B] and [A] were thinking about going to the [PLACE]. [B] wanted to give a [OBJECT] to [A]
Then, [B] and [A] had a long argument, and afterwards [B] said to [A]
After [B] and [A] went to the [PLACE], [B] gave a [OBJECT] to [A]
When [B] and [A] got a [OBJECT] at the [PLACE], [B] decided to give it to [A]
When [B] and [A] got a [OBJECT] at the [PLACE], [B] decided to give the [OBJECT] to [A]
While [B] and [A] were working at the [PLACE], [B] gave a [OBJECT] to [A]
While [B] and [A] were commuting to the [PLACE], [B] gave a [OBJECT] to [A]
After the lunch, [B] and [A] went to the [PLACE]. [B] gave a [OBJECT] to [A]
Afterwards, [B] and [A] went to the [PLACE]. [B] gave a [OBJECT] to [A]
Then, [B] and [A] had a long argument. Afterwards [B] said to [A]
The [PLACE] [B] and [A] went to had a [OBJECT]. [B] gave it to [A]
Friends [B] and [A] found a [OBJECT] at the [PLACE]. [B] gave it to [A]

Table 2: Templates used in the IOI dataset. The table displays templates following the "BABA" pattern; templates with the "ABBA" pattern were also employed but are omitted here for clarity.

B Connecting EAP to EP

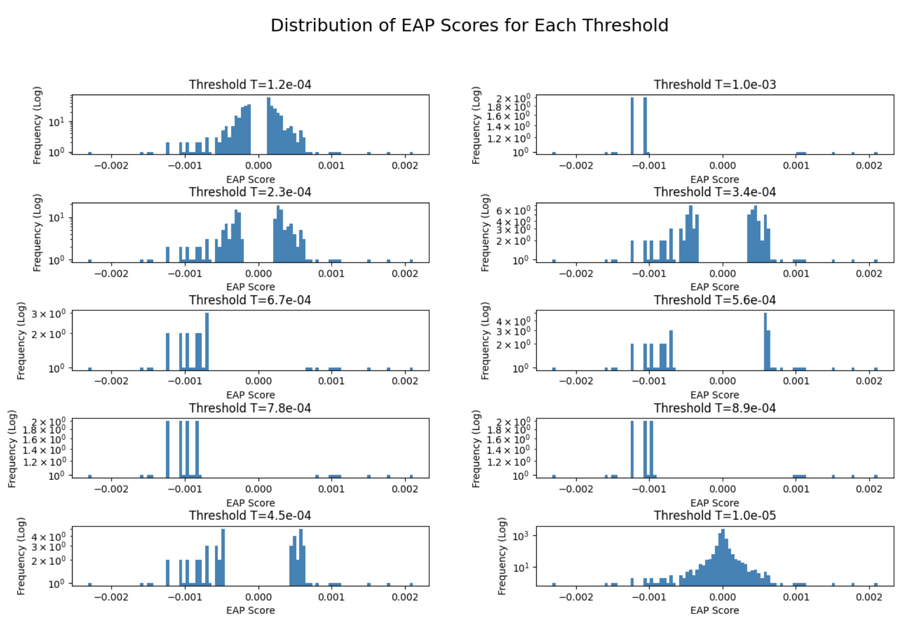


Figure 2: Attribution score distribution over different EAP thresholds.

To map the high-potential edges identified by EAP to the binary masks z of EP, we first show that EAP attribution scores are generally normally distributed (Figure 2). Then, we normalize the output attribution scores to a range $\in [-1, 1]$. To integrate the normalized attribution scores into the binary masks z of EP, we create the initial $log\alpha$ tensor. We then modify the EP initialization by changing the relevant mask parameters using the computed $log\alpha$ tensor. EP then undergoes training.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims are that 1) HAP improves circuit discovery speed by 46% which is supported in Section 5.1 and 2) that HAP maintains all 4 s-inhibition heads which is shown in section 5.2 and figure 1. Also, our claim that HAP could helpful for improving the scalability of circuit discovery in the future is supported by the speed boost mentioned earlier.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations to demonstrated scalability, optimality, dataset variation in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theoretical proofs/claims.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We give the specific conditions to replicate our work in Section 4.1 and 4.2 including dataset size, dataset origin, model, specific hardware setup, and parameters. We also provide a link to our code in the abstract.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide our code + instructions to run in the readme. Code link: `https://anonymous.4open.science/r/HAP-circuit-discovery/README.md`

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify training splits in 4.1 and hyperparameters in 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the experimental setups and design of our baseline comparisons (prominent papers in our field such as Conmy et al. [2023], Syed et al. [2023], Bhaskar et al. [2024]), which don't report error bars in the data they provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: We state the hardware specifications we used, specifically the H100 GPU, in Section 4.2 and provide runtime data in Table 1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our work involves no human subjects or information, the dataset is openly generated from code based on Wang et al. [2022] (not deprecated/outdated) as mentioned in 4.1. Our work aims to make AI more interpretable, safe, and fair, hence does not violate any ethics guidelines.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss how our work can help circuit discovery scale up to larger models, making model interpretability more practical and a viable tool for making AI systems safe and fair.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks, as our dataset is based off Wang et al. [2022] and contains no perosnal information. Also, we don't release any machine learning models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We acknowledge and properly cite Syed et al. [2023] (CC-BY 4.0), Bhaskar et al. [2024] (CC-BY 4.0), Wang et al. [2022] (CC-BY 4.0) for using their code and dataset generator code. We also properly cite Huggingface in the bibliography for their GPT2-small model (MIT License) that we used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide the code to our algorithm in the abstract and provide documentation along with it. The Link is anonymized. All borrowed code in the repo are from CC-BY 4.0 licenses, detailed in question 12 above.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This research does not involve human subjects or crowd-sourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

Justification: LLMs were not part of the core method development of our research. We only used GPT2 as a standard component in the circuit discovery workflow to benchmark novel and existing algorithms.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.