CHARACTERIZING MECHANISTIC UNIQUENESS AND IDENTIFIABILITY THROUGH CIRCUIT ANALYSIS

Ali Sahi*, **Imran Kutianawala***, **Medhansh Beeram, Siwoo Song, Aryan Shrivastava**
Algoverse AI Research; University of Chicago

*Equal contribution

ABSTRACT

As deep-learning frameworks grow more capable, understanding the mechanisms behind their emergent behaviors is critical for safety. Mechanistic interpretability seeks to reverse-engineer the internal representations and algorithms that produce model behaviors. One method for interpretability is 'circuit analysis,' which interprets neural networks through subsets of their nodes and edges that replicate the network's behavior. Recent work has demonstrated that multiple, non-unique circuits can implement the same input–output mappings within a network, making the source of a network's behavior non-identifiable (Méloux et al., 2025). However, the conditions under which functionally equivalent circuits arise remain poorly understood. In this work, we train small multi-layer perceptrons on Boolean tasks, varying architectural and training choices to identify conditions under which multiple, functionally equivalent circuits emerge. We find that methods inducing sparsity, such as L1 regularization and orthogonalization, show a weak correlation with uniqueness, whereas increasing hidden layer size leads to a significant increase in the number of equivalent circuits. Finally, we show that task complexity is a strong predictor of non-uniqueness in networks: reductions in task complexity collapse networks to one causal pathway with increasingly redundant circuits. Our findings offer guidance for researchers in determining when circuit-based explanations can uniquely identify the source of a network's behavior. [1]

1 INTRODUCTION

While recent machine learning research has produced high-level explanations for model behaviors (Ruan et al., 2024; Gadre et al., 2024; Liang et al., 2022; Wei et al., 2022), these methods provide little insight into neural networks' internal computations (Alishahi et al., 2019; Sharkey et al., 2025). Mechanistic Interpretability (MI) addresses this opacity by revealing human-interpretable algorithms within networks, identifying both what algorithm a network implements and where it is instantiated (Olah et al., 2020; Templeton et al., 2024). One method for MI researchers is 'circuit analysis,' which decomposes networks into minimal subgraphs of nodes (e.g. attention heads, MLP layers, or individual neurons) and edges that can perform a task independent of other network components (Olah et al., 2020; Elhage et al., 2021).

Recent work has demonstrated that distinct circuits within a network can implement identical input–output behaviors (Méloux et al., 2025), a phenomenon termed mechanistic non-uniqueness. If multiple circuits produce the same behavior, the underlying mechanism is 'non-identifiable' because researchers can not uniquely recover which circuit is responsible for a given behavior. Notably, while non-uniqueness has been demonstrated and quantified in small MLPs, the conditions under which uniqueness holds remain unclear. This introduces ambiguity for researchers mapping observed behaviors to internal mechanisms. In this paper, we investigate the architectural and training conditions under which networks exhibit mechanistic uniqueness.

In this work, to make brute-force circuit enumeration feasible, we train primarily small MLPs on the Boolean XOR (Méloux et al., 2025) and XNOR tasks as they require non-linear computation,

*Corresponding author: imranmk@bu.edu

[1] All code necessary to reproduce this paper's experiments can be found on GitHub.

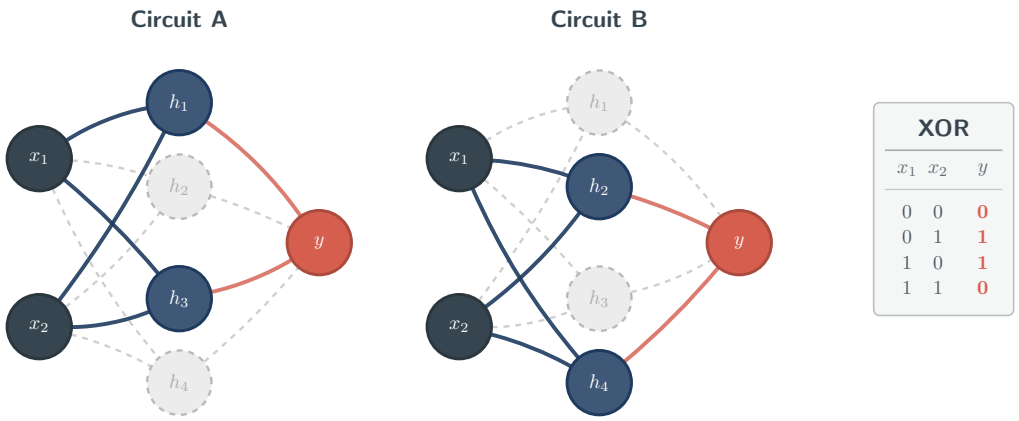

Figure 1: Mechanistic non-uniqueness in a 2-4-1 MLP trained on Boolean XOR, which evaluates to true if exactly one of the two inputs is one. Circuit A (hidden layer neurons h_1, h_3) and Circuit B (hidden layer neurons h_2, h_4) are disjoint sub-networks that independently reproduce the full network's behavior. Active components are shown in solid lines, and pruned components are dashed.

using AND and OR as linear baselines. We then induce sparsity through L1 Regularization and Orthogonalization constraints. Additionally, we vary model capacity by adjusting hidden layer width. For each network, we exhaustively enumerate all possible circuits and test whether multiple valid circuits independently replicate the full network's behavior on the task to a high degree of accuracy.

We find that, counterintuitively, L1 regularization and activation sparsity exhibit weak and inconsistent relationships with uniqueness. Orthogonalization constraints reduce circuit multiplicity in wider networks by penalizing redundancy, though the effect varies with model width. Instead, model capacity is a strong driver of mechanistic uniqueness: hidden layer width shows a strong, positive correlation with the number of valid circuits, even after normalizing for the number of possible circuit subsets. Additionally, models with excess capacity relative to task complexity tend to develop more functionally equivalent circuits. These results indicate capacity plays a larger role than sparsity in determining mechanistic uniqueness.

This work aims to identify when circuit-based analyses can uniquely attribute a behavior to a single circuit. By clarifying when mechanistic uniqueness holds, we hope to help researchers assess when circuit analysis gives a complete picture of model behavior. Broadly, we seek to build upon current efforts to make AI models more transparent. To summarize, our contributions are as follows:

1. L1 regularization and orthogonalization constraints can reduce the number of valid circuits, but the effect is inconsistent across settings.
2. Networks with higher capacity (e.g. wider hidden layers) are more likely to admit multiple valid circuits for the same task.
3. Task complexity predicts non-uniqueness. Tasks with a higher 'effective rank' correlate with a greater number of distinct, unique circuits.

2 Related Works

Circuit-Based Interpretability This work assumes neural networks implement certain functions using subsets of their nodes and edges, a framework introduced by Olah et al. (2020) and closely related to circuits in biological neural networks (Yuste, 2008). Circuit-based interpretability efforts localize where a behavior is implemented and determine what features those components compute (Méloux et al., 2025). This framework has been formalized mathematically (Elhage et al., 2021) and validated empirically across model architectures (Mondorf et al., 2024; Shi et al., 2024). Valid circuits have been found in language models (Wang et al., 2022; Hanna et al., 2023; Stolfo et al., 2023; Nanda et al., 2023; Olsson et al., 2022) and for high-level feature detection in vision models

(Cammarata et al., 2020; Schubert et al., 2021). However, prior work shows that circuit analysis can be highly sensitive to minor input perturbations and experimental choices, with this instability especially pronounced in Transformer architectures (Méloux et al., 2025; Miller et al., 2024).

Circuit Enumeration & Discovery In this work, we test all possible subsets of neurons and edges in small MLPs on Boolean logic datasets. Recent work employed 'activation patching' (Zhang & Nanda, 2024; Vig et al., 2020; Meng et al., 2022; Geiger et al., 2021), a process by which researchers selectively replace neuron activations to identify which components are causally necessary for model behavior. Subsequent work has posited attribution patching (Syed et al., 2023; Kramár et al., 2024; Ferrando & Voita, 2024), a similar process using gradient-based attribution methods to approximate patching effects. Conmy et al. (2023) automated the circuit-discovery process using activation patching to iteratively prune edges from computational graphs.

Redundancy in Neural Networks Our research builds upon previous work demonstrating redundant components in artificial and biological neural networks. Much research has been done proving that redundancy is prevalent in biological systems and neural networks (Edelman & Gally, 2001; Tononi et al., 1999). Subsequent work observed similar phenomena in artificial neural networks as different weight configurations can yield identical input–output mappings (Bushnaq et al., 2024). Relatedly, the hydra effect reveals that ablating circuit components often activates latent backup mechanisms rather than eliminating the behavior (McGrath et al., 2023). Most recently, Méloux et al. (2025) demonstrated that multiple circuits can independently replicate a full network's behavior. Our work creates a preliminary framework for determining when multiple valid circuits arise in a network.

3 METHODS

3.1 WHAT IS A CIRCUIT?

We define a circuit in the same manner as Méloux et al. (2025): a subset of a model's neurons (nodes) and their connecting weights (edges) that can independently perform an interpretable target task. For instance, a circuit implementing the AND function within a network may consist only of the two input neurons, one hidden neuron, and a single output neuron, along with the edges connecting them. In order to isolate circuits from their surrounding models, we ablate extraneous edges by zeroing the weights and keeping all remaining parameters, including biases, fixed. This isolates the contribution of specific pathways without altering the internal computations of participating neurons. Notably, we avoided ablating neurons and focused on edges to avoid any conflation with representational reuse and connectivity reuse. Individual neurons in networks are often reused across multiple representations and downstream pathways. Through neuron-level ablation, all associated representations and pathways through a neuron are removed. In contrast, edge-level ablation preserves learned activations and only selectively disables specific connections. Furthermore, it allows us to isolate the contribution of particular pathways when neurons are shared across mechanisms.

In this work, circuits must meet the following technical criteria:

- Every active neuron has at least one active incoming and outgoing edge (except for input and output neurons, respectively)
- Every active edge connects two active neurons, and likewise, every inactive neuron has no incoming or outgoing edges

These criteria verify that each circuit can accurately process inputs and return outputs. Furthermore, they ensure degenerate or disconnected structures are not considered. In this paper, we classify one of these circuits as 'valid' if it can independently reproduce the full model's behavior to a high degree of accuracy, defined here as a mean squared error below 5%. We consider a valid circuit to be 'minimal' if it contains no neurons and edges that are unnecessary to complete the task (Wang et al., 2022).

3.2 TERMINOLOGY

Here, we define the various terminology used throughout this paper.

Mechanistic Explanation A mechanistic explanation describes the behavior of a circuit and how it contributes to the overall network, providing a human-readable interpretation of the circuit. (e.g. "Circuit $\mathcal{C}$ detects XOR parity")

Mechanistic Uniqueness A network exhibits mechanistic uniqueness when there exists exactly one minimal, valid circuit that implements a given task. Conversely, mechanistic non-uniqueness occurs when multiple distinct circuits produce identical outputs for the same inputs on a task, making them functionally equivalent. We refer to the presence of multiple functionally equivalent circuits within a network as circuit multiplicity.

Identifiability Identifiability is a key component of interpretability, as it enables functional behaviors to be unambiguously attributed to a single circuit. For a network's behaviors to be identifiable, each sub-task must map to a unique, minimal circuit that completes the task. However, if there are multiple distinct circuits that produce the same behavior, this mapping from internal components to function becomes non-identifiable, meaning multiple mechanistic explanations may exist.

3.3 Experimental Setup

In this paper, we investigate circuit multiplicity by using toy tasks in small multilayer perceptrons (MLPs). We choose the Boolean operations XOR, XNOR, OR, and AND tasks as they allow for evaluation over all possible inputs and avoid biases introduced by naturalistic datasets. We particularly work on non-linear tasks (XOR & XNOR) for experimentation. Intuitively, non-linear tasks can often admit multiple intermediate sub-tasks, making trained networks more likely to demonstrate circuit multiplicity. We placed less emphasis on linear tasks, as a single effective direction suffices for linear computations.

Furthermore, we use small MLPs, making brute-force circuit enumeration tractable while retaining enough model capacity to admit multiple internal implementations. We adopt a 2-4-1 MLP (two input units, a four-unit hidden layer, and one output unit). We then vary the width of the hidden layer to examine the effect of model capacity on circuit multiplicity (Section 4.3).

For our experiments, we primarily use the rectified linear unit (ReLU) and hyperbolic tangent (Tanh) activation functions in the hidden layers and a sigmoid function in the output layer. In preliminary experiments, this led to stable optimization in small networks. General training details for experiments are provided in Appendix B.

3.4 Circuit Enumeration Procedure

For each trained model, we enumerate all admissible subsets of nodes and edges consistent with the constraints described above (Section 3.1). For each candidate circuit, we construct the corresponding ablated model—the isolated circuit independent of all other network components—and evaluate it on the full validation dataset. We treat a circuit as valid when the ablated model matches the full model's output with a mean squared error below 5%. This criterion allows small deviations while ensuring the circuit maintains functional equivalence to the original model. Our conception of a valid circuit also corresponds to a sufficient sub-graph.

Complete enumeration is feasible only for small architectures because the number of possible solutions grows exponentially with the number of parameters. Hence, all enumeration experiments are restricted to models with limited width and depth (unless otherwise noted).

To encapsulate, we identify circuits through the following process:

- Enumerate all possible subsets of nodes & edges in the trained model and construct the corresponding ablated model for each subset.
- Evaluate the ablated model on the full validation dataset.
- Declare a circuit as valid if the ablated model achieves near perfect accuracy on validation set, which we define here as a mean squared error below 5%

3.5 METRICS AND DEFINITIONS

To quantify circuit redundancy and representation usage, we measure activation sparsity, normalized multiplicity, and mean overlap, defined below.

Activation Sparsity The frequency with which hidden units are inactive across the entire dataset can be measured using activation sparsity. Activation sparsity refers to the fraction of hidden activations have a value of zero (or almost zero) over the validation set. This metric is taken from neuron activations, and describes the level of selectiveness of a network with respect to its use of hidden representation.

Normalized multiplicity In our experimentation, we take into account differing architectures when calculating the fraction of valid circuits. Normalized multiplicity is defined as the number of valid circuits divided by the total number of allowable circuit subsets (for the same network size). By using this measure, we can make a more thorough comparison of different models (regardless of their size) by accounting for the much larger number of circuits generated by larger networks, since wider networks admit exponentially more possible subgraphs.

Mean Overlap We seek to find how similar valid circuits are to one another. Hence, mean overlap is the average pairwise overlap between valid circuits. It is measured as the fraction of shared active components (nodes and edges) between two circuits. Moreover, a lower mean overlap indicates that valid circuits are more disjoint, whereas a higher overlap suggests that circuits share common structure.

4 RESULTS AND ANALYSIS

Using these methods (Section 3), we analyze potential factors contributing to circuit multiplicity and mechanistic non-uniqueness across our toy models.

4.1 INDUCED SPARSITY

We first investigate whether weight sparsity could force the model to converge on a single, unique circuit. Sparsity was induced by adding an L1 penalty to the training objective. Specifically, we augmented the binary cross-entropy loss with an L1 regularization coefficient term applied to our first-layer weight matrix, scaled by a regularization coefficient $\lambda \in \{0, 10^{-4}, 10^{-3}\}$. For each λ, a separate model was trained. While induced sparsity is helpful in eliminating redundancy, we find that L1 regularization and activation sparsity appear to be weak predictors of mechanistic uniqueness.

Our analysis shows no significant correlation between the strength of regularization and the reduction of circuit multiplicity. While L1 regularization successfully encourages weight sparsity, it does not necessarily resolve the linear dependence of activations that leads to multiple redundant circuits. Our results, presented in Figure 2, show Spearman correlations across various mechanistic metrics to quantify the significance of each relationship. We observe that the correlation between sparsity and the number of valid circuits is weak with $\rho = 0.18$. Additionally, we observe that $\rho = 0.083$ in the relationship between L1 regularization and the number of valid circuits, indicating that L1 regularization has minimal influence on unique circuits. Overall, these values imply that sparsity alone is insufficient to enforce mechanistic uniqueness.

However, while sparsity shows a weak relationship with uniqueness, Figure 2 reveals how multiple non-unique circuits may be organized in a network: We observed $\rho = -0.49$ between mean overlap and normalized multiplicity. This result indicates that as multiplicity increases in a network, the distinct non-unique circuits share fewer components (neurons & edges) between each other. In other words, non-unique circuits become disjoint, a behavior visually represented in Figure 1. Overall, our results suggest that neural networks may prefer to "pack" functionally equivalent circuits into separate subspaces when capacity permits, resulting in modular rather than distributed redundancy.

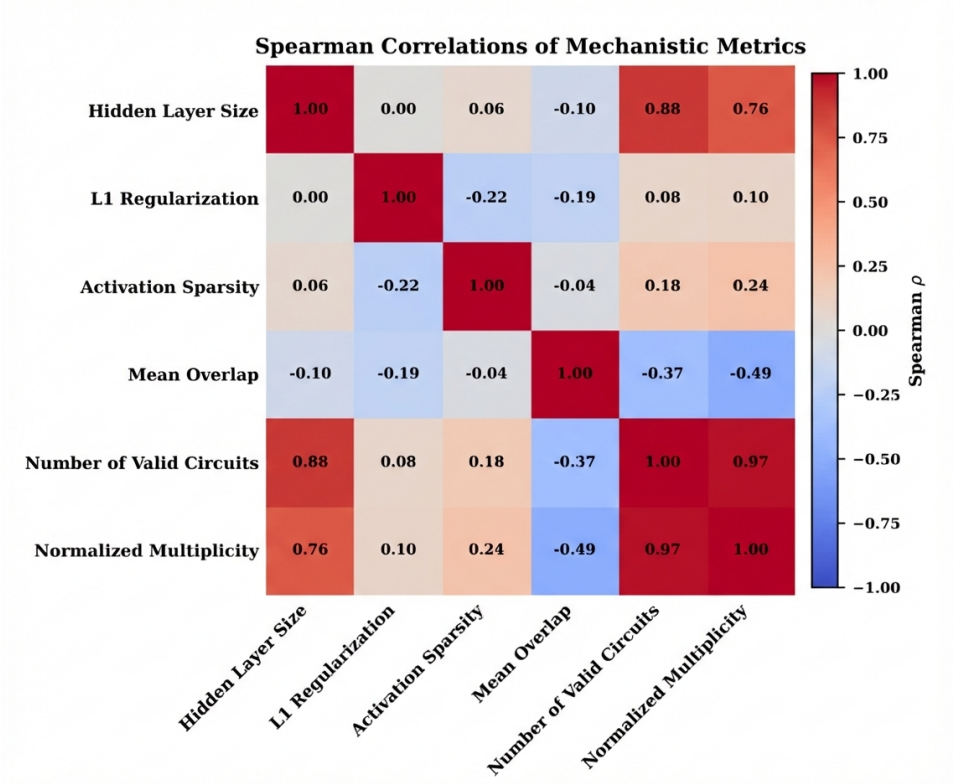

Figure 2: Hidden layer width is strongly correlated with the number of valid circuits ($\rho = 0.88$) and normalized multiplicity ($\rho = 0.76$), while L1 regularization and activation sparsity exhibit weak relationships with multiplicity ($\rho = 0.10$ and $\rho = 0.24$, respectively). Additionally, mean overlap and normalized multiplicity have a strong negative relationship ($\rho = -0.49$).

4.2 Representational Non-Uniqueness and Orthogonalization

To understand why sparsity fails to determine uniqueness, we examine representational non-uniqueness, where different internal configurations encode the same function. This occurs mainly when neurons are linearly independent of one another. To test this, we applied an orthogonality regularization term ($\lambda_{\text{orth.}}$) that encourages hidden-layer activations $H \in \mathbb{R}^{N \times d}$ to be linearly independent between neurons. For a hidden width layer of width d and batch size N, we added the following penalty to the standard binary cross-entropy loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{orth.}} \left\| \frac{H^\top H}{N} - I_d \right\|_F^2$$

Here, I_d is the $d \times d$ identity matrix, $\lambda_{\text{orth.}}$ is a hyperparameter controlling the strength of our orthogonality constraint, and $\| \cdot \|_F$ denotes the Frobenius norm that expands to a sum over diagonal and off-diagonal terms. This penalty allows us to minimize off-diagonal correlations among neuron activations to promote the hidden representations to span independent directions. It penalizes deviations of the empirical covariance of the hidden activations from the identity matrix so multiple neurons encoding the same feature are discouraged.

In Figure 3, the number of valid circuits remains at 2 for all values of $\lambda_{\text{orth.}}$ in low-capacity models with 3 and 4 hidden-layer neurons, indicating that capacity constraints are stronger predictors of uniqueness than weight correlations. Meanwhile, multiplicity grew monotonically for the mid-capacity model (5 hidden-layer neurons) over $\lambda_{\text{orth.}} \in [0.01, 0.10]$, suggesting that orthogonalization does not reduce multiplicity, and may instead lead to additional valid circuits. However, a high capacity model of width 6 neurons converges from 16 to 4 valid circuits when $\lambda_{\text{orth.}} \in [0, 0.01]$ before inflecting back to monotonic growth.

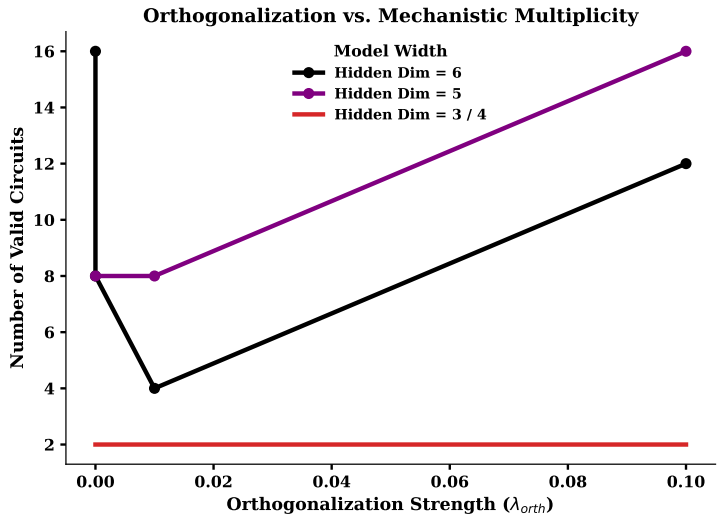

Figure 3: Effect of orthogonalization strength on mechanistic multiplicity across models with hidden layer widths of size 3, 4, 5, and 6. As the orthogonalization coefficient $\lambda_{\text{orth.}}$ increases, models with larger hidden layers exhibit a clear increase in the number of functionally equivalent circuits, while smaller models remain comparatively constrained.

We hypothesize that orthogonalization constraints ($\lambda_{\text{orth.}} > 0.01$) encourage a task's computation to be distributed across multiple decorrelated directions for high-capacity models. The number of task-sufficient subspaces thus increases within the hidden space. Therefore, multiple subsets of orthogonal neurons can independently span the task-sufficient subspace, leading to the number of minimal valid circuits to increase (reintroducing mechanistic non-uniqueness). Compared to high-capacity models, low-capacity models contain fewer distinct computational directions, letting orthogonalization break task performance. Therefore, orthogonalization restricts the directional freedom to redistribute representations to a higher degree. On the other hand, in high-capacity models, the task can be supported by many low-dimensional subspaces spanned by different sets of computational directions. In other words, this could be attributed to orthogonalization not constraining which subspace encodes the task, which leaves many sufficient representations to implement the task. Overall, these results suggest that orthogonalization does not reliably reduce circuit multiplicity (especially for high capacity models).

4.3 Model Capacity

As shown in Figure 2, we observe a strong, positive correlation between model dimensions and the number of functionally equivalent valid circuits, where $\rho = +0.880$. This suggests that non-uniqueness scales with model capacity, especially with hidden layer width. Even when normalizing by the number of possible subsets, we observe that $\rho = +0.764$ in wider models, showing that they exhibit significant functional redundancy.

To understand the effects of scaling, we analyzed the condition number of the hidden layer activation matrices, which let us assess how constrained the internal representation is. We calculate the ratio between the largest and smallest singular values of the hidden-layer weight matrix using singular value decomposition. This allows us to measure whether the computation is concentrated in a small number of well-conditioned directions or spread across many weakly constrained ones.

Models with low multiplicity tend to have lower condition numbers, indicating that the computation is dominated by a small number of stable, well-conditioned directions (Bushnaq et al., 2024). In contrast, higher condition numbers are associated with greater multiplicity, reflecting weight matrices that are close to singular and admit many, nearly equivalent internal solutions. This pattern is shown in Figure 4. For models with width 3, condition numbers range from 10^1 to 10^5 and multiplicity remains near three valid circuits. For models with width 6, the condition number increases

to approximately 10^{15} and multiplicity exceeds 40 valid circuits, indicating that the network implements the same behavior through many weakly constrained directions. Taken together, these results suggest that the hidden-layer condition number is a useful indicator of mechanistic multiplicity.

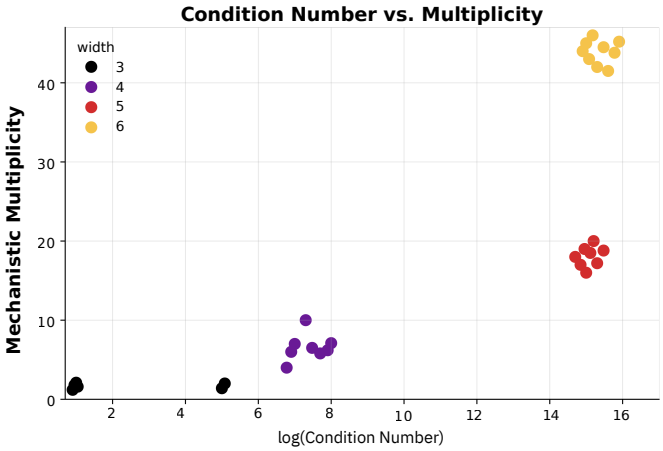

Figure 4: Mechanistic multiplicity increases with the condition number of the hidden-layer weight matrix. Models with smaller hidden dimensions exhibit lower condition numbers and support few valid circuits, while models with larger hidden dimensions exhibit extremely high condition numbers and support many distinct circuit implementations of the same input–output behavior.

4.3.1 Effective Rank & Representational Rank

We characterize tasks by the minimum dimensionality of internal representations required to implement their input–output mapping. Tasks that admit low-dimensional representations impose strong constraints on how the computation can be implemented internally, whereas tasks that require higher-dimensional representations allow more freedom in internal structure.

We formalize this notion using effective rank, which is defined as the minimum dimensionality needed to represent a linearly separable task. Linearly separable Boolean tasks such as AND and OR can be implemented using a single weighted sum and a point-wise nonlinearity, corresponding to a rank-1 internal representation. These tasks require one independent hidden feature, which limits how the computation can be represented. In contrast, tasks such as XOR and XNOR are not linearly separable and require partitioning the input space into multiple regions. Implementing these tasks therefore requires at least rank-2 representations and multiple intermediate features.

We use effective rank as a representational measure rather than a measure of task difficulty or complexity. However, effective rank does not reflect other aspects of difficulty such as optimization behavior, sensitivity to initialization, or data efficiency. It does place a lower bound on the representational degrees of freedom available for implementing the task. Tasks with lower effective rank therefore admit fewer distinct internal implementations, restricting the number of unique circuits that can implement the task.

Precisely calculating effective rank requires complete circuit enumeration, which is intractable on naturalistic tasks, thus motivating our use of synthetic boolean tasks (Section 5.1). However, Roy & Vetterli (2007) show that effective rank scales predictably with network width and remains robust to noise, suggesting the correlations between effective rank and multiplicity identified here scale to larger tasks.

To evaluate the behavior of trained models, we measure the representational rank, defined as the numerical rank of the hidden-layer activation matrix. In this sense, effective rank serves as a proxy for the minimum representational degrees of freedom required by the task, while representational rank reflects the degrees of freedom actually utilized by the trained model. Because effective rank is fixed for a given Boolean task, differences in mechanistic multiplicity must arise from variation in representational rank above this lower bound. When representational rank exceeds effective rank,

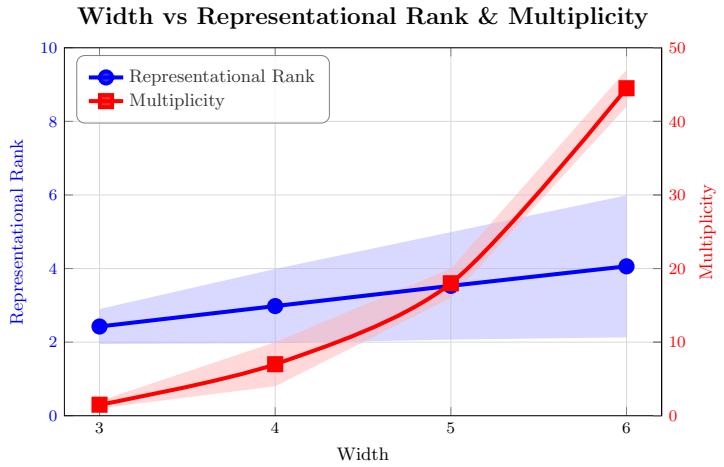

Figure 5: Effect of model width on representational rank and circuit multiplicity. Increasing hidden-layer width leads to a gradual increase in representational rank but a sharp rise in circuit multiplicity, indicating that modest gains in representational capacity are accompanied by a large increase in the number of functionally equivalent circuits.

the model possesses additional degrees of freedom that admit multiple valid circuit configurations. As the hidden dimension increases, we observe a corresponding increase in representational rank, indicating that additional hidden dimensions are actively used by the model. In wider models, this increase reflects the emergence of multiple independent features rather than the reuse of the same representational directions.

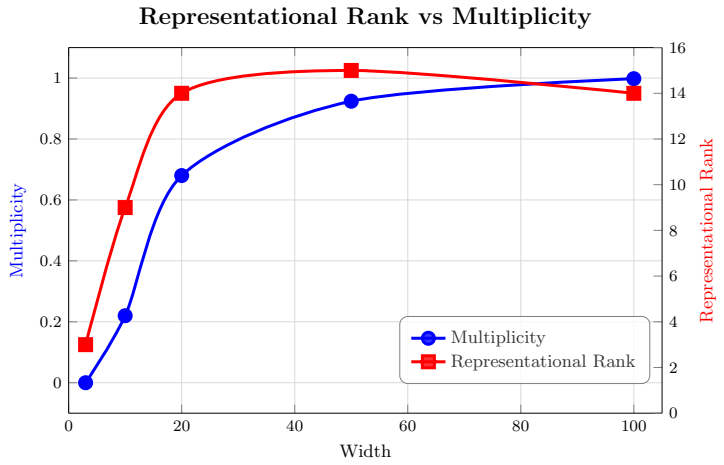

Figure 6: Effect of model width on representational rank and circuit multiplicity. As hidden-layer width increases, both representational rank and the number of valid circuits increase and exhibit closely aligned trends. Both curves increase most rapidly at smaller widths, with representational rank tapering off at later widths.

As demonstrated in Figure 5, representational rank increases from 2.90 to 5.99 as the hidden dimension grows from 3 to 6 neurons. Over the same range, the number of valid circuits increases from 1 to 42, indicating that higher representational rank is associated with increased non-uniqueness in circuit structure. We further examine how representational rank and circuit multiplicity scale in models with higher capacities. As shown in Figure 6, we tracked the representative rank and multiplicity across varying widths. We observed that as the model's capacity allows for higher-rank representations, the mean multiplicity increases as well. At width 3, the model learns a representation with

a rank of 3. In this constrained regime, there is no multiplicity, indicating a unique or highly constrained solution. As width increases to 50 and 100 neurons, the representative rank stays around 14 - 15. The increase in rank coincides with the mean multiplicity approaching 1.0, confirming that higher-rank representations indicate an increase in valid circuits.

5 Discussion

In this work, we characterized the conditions under which a network's behavior can be explained by a unique circuit. Our experiments reveal that model capacity is a dominant predictor of circuit multiplicity, with higher hidden dimension sizes leading to more valid circuits. Specifically, models with capacity exceeding task complexity tend to develop multiple equivalent circuits. Surprisingly, inducements of sparsity, such as L1 regularization, and orthogonality showed weak and inconsistent effects, suggesting that regularization techniques alone cannot guarantee identifiable explanations.

5.1 Limitations & Future Work

This work is not without limitations. To begin, our experiments focus on small-scale multi-layer perceptrons. Our findings suggest that, even in small MLPs, marginal increases in hidden layer width show a strong correlation with increases in multiplicity. Prior work implies the conditions identified in this paper may also scale to larger networks where brute-force enumeration isn't feasible (Méloux et al., 2025). Still, the extent to which these findings generalize to larger-scale neural networks or different architectures (e.g. Transformers) warrants further investigation.

Additionally, due to computational constraints on brute-force circuit enumeration, we only analyze synthetic Boolean tasks. Such tasks were necessary to make precise, ground-truth computations of effective rank and circuit multiplicity under computational constraints. It is possible that natural datasets introduce properties that affect circuit multiplicity in ways not captured by Boolean logic. These limitations motivate future work to employ automated circuit discovery processes, such as those proposed in Conmy et al. (2023), to replicate our experiments on a larger scale with more nuanced datasets.

5.2 Implications

These findings seek to improve the transparency and reliability of circuit-based analyses, helping researchers predict when behaviors can be traced back to one circuit. Furthermore, we believe this work motivates future researchers to cleverly design neural networks to promote identifiability. Whether uniqueness is necessary for robustness or safety remains an open question. When multiple valid circuits exist for emergent behaviors, researchers are unable to attribute those changes to a specific pathway. This ambiguity poses challenges for AI safety efforts that rely on circuit analysis (Sharkey et al., 2025). One might argue that redundancy in artificial networks is beneficial, providing resilience against component failure. However, this could also complicate efforts to remove unwanted capabilities through targeted interventions (Zou et al., 2024), since redundant pathways could preserve the behavior. We hope this research encourages further investigation into the conditions under which circuit-based explanations can reliably inform alignment research.

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.

Lucius Bushnaq, Jake Mendel, Stefan Heimersheim, Dan Braun, Nicholas Goldowsky-Dill, Kaarel Hänni, Cindy Wu, and Marius Hobbhahn. Using degeneracy in the loss landscape for mechanistic interpretability. *arXiv preprint arXiv:2405.10927*, 2024.

Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve detectors. *Distill*, 2020. doi: 10.23915/distill.00024.003. https://distill.pub/2020/circuits/curve-detectors.

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

G M Edelman and J A Gally. Degeneracy and complexity in biological systems. *Proc. Natl. Acad. Sci. U. S. A.*, 98(24):13763–13768, November 2001.

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. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale, 2024. URL https://arxiv.org/abs/2403.00824.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks, 2021. URL https://arxiv.org/abs/2106.02997.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *ArXiv*, abs/2305.00586, 2023. URL https://api.semanticscholar.org/CorpusID:258426987.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. Atp*: An efficient and scalable method for localizing llm behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations, 2023. URL https://arxiv.org/abs/2307.15771.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? *arXiv preprint arXiv:2502.20914*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust, 2024. URL https://arxiv.org/abs/2407.08734.

Philipp Mondorf, Sondre Wold, and Barbara Plank. Circuit compositions: Exploring modular structures in transformer-based language models. *ArXiv*, abs/2410.01434, 2024. URL https://api.semanticscholar.org/CorpusID:273026338.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic interpretability as statistical estimation: A variance analysis of eap-ig, 2025. URL https://arxiv.org/abs/2510.00845.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

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. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. *2007 15th European Signal Processing Conference*, pp. 606–610, 2007. URL https://api.semanticscholar.org/CorpusID:12184201.

Yangjun Ruan, Chris J Maddison, and Tatsunori B Hashimoto. Observational scaling laws and the predictability of langauge model performance. *Advances in Neural Information Processing Systems*, 37:15841–15892, 2024.

Ludwig Schubert, Chelsea Voss, and Chris Olah. High/low frequency detectors. *Distill*, 6, 01 2021. doi: 10.23915/distill.00024.005.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in llms. *ArXiv*, abs/2410.13032, 2024. URL https://api.semanticscholar.org/CorpusID:273403869.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=aB3Hwh4UzP.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery, 2023. URL https://arxiv.org/abs/2310.10348.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

G Tononi, O Sporns, and G M Edelman. Measures of degeneracy and redundancy in biological networks. *Proc. Natl. Acad. Sci. U. S. A.*, 96(6):3257–3262, March 1999.

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. *arXiv preprint arXiv:2004.12265*, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Rafael Yuste. Circuit neuroscience: the road ahead. *Frontiers in Neuroscience*, 2(1):6–9, 2008. doi: 10.3389/neuro.01.017.2008. URL https://www.frontiersin.org/articles/10.3389/neuro.01.017.2008/full.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods, 2024. URL https://arxiv.org/abs/2309.16042.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers, 2024. URL https://arxiv.org/abs/2406.04313.

A SPEARMAN CORRELATION

Table 1: Spearman correlations of mechanistic metrics

MECHANISTIC METRICS	**P-VALUE (P)**
Hidden Dimension vs Num. of Valid Circuits	< 0.001
Hidden Dimension vs Normalized Multiplicity	< 0.001
Mean Overlap vs Num. of Valid Circuits	$= 0.0033$
Mean Overlap vs Normalized Multiplicity	< 0.001
Number of Valid Circuits vs Normalized Multiplicity	< 0.001
L1 Regularization vs Num. of Valid Circuits	$= 0.5307$
Sparsity vs Num. of Valid Circuits	$= 0.1763$

B TRAINING DETAILS

We report the training hyperparameters used for all experiments in Table 2. Our training continued until loss converged to 0.001. Batch size and number of training epochs were varied across experiments to ensure convergence under this criterion.

Table 2: Regular training setup

TRAINING HYPERPARAMETER	**TYPE**
Loss function	Binary Cross-Entropy Loss
Learning rate	0.05
Optimizer	Adam