Causal Promises but *Correlational* Delivers: A Study of Circuit Mechanisms and Model Behaviors

Jenny Kaufmann*
Dept. of Mathematics
Harvard University
jkaufmann@math.harvard.edu

Victoria Li*
Dept. of Computer Science
Harvard University
vrli@college.harvard.edu

Martin Wattenberg
Dept. of Computer Science
Harvard University
wattenberg@seas.harvard.edu

David Alvarez-Melis
Dept. of Computer Science
Harvard University
dam@seas.harvard.edu

Naomi Saphra
Kempner Institute
Harvard University
nsaphra@g.harvard.edu

Abstract

Using a toy balanced parenthesis classification task with an ambiguous rule, we investigate the correspondence between attention patterns and out-of-distribution generalization behavior of small transformer models. We find that observational tools can predict OOD behavior, challenging the common notion among interpretability researchers that causal intervention is the only basis for validating explanations of model behavior.

1 Introduction

"Correlational" has become an insult in the field of machine learning interpretability. Increasingly, we treat causal intervention as the gold standard validation of explanatory methods [7]. Whether intervention is performed on a fully trained model or during model training, the idea is clear: If we have a good explanation of model behavior, we should be able to predict the outcome of specific interventions.

In the natural sciences, however, intervention is not always available—and where it is available, it may not always provide ecologically valid answers. Indeed, intervention-based interpretability approaches are closer to artificial lab environments rather than naturalistic variation. Applying these methods to fully trained models is likely to yield results specific to the settings we are testing them in, and interventions on training—usually conducted by learning on datasets that change model behavior entirely—fundamentally alter the information the model is provided with, possibly eliciting broader changes than we are interested in measuring and producing findings that do not transfer between data settings. *Here, we investigate the benefit of a rich source of naturalistic variation rather than deliberate intervention when attempting to predict model behavior under distribution shift.*

To this end, we study the variability and interpretation of out-of-distribution (OOD) behavior in models with perfect in-distribution (ID) validation performance by introducing naturalistic variation between models. In particular, we vary randomness in model initialization and data order, which do not directly suggest particular OOD generalization behavior or internal circuits, while still creating systematic differences dependent on model depth, weight decay, and random seed.

We train a large population of models on an ambiguous parenthesis sequence dataset, where models can achieve perfect ID accuracy through learning a parentheses **balancing** rule or a parentheses

*Joint first author.

Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

counting rule (i.e., same number of open and closed parentheses sequences), respectively called BALANCED and EQUAL-COUNT. We use an OOD test set to determine which rule each model follows. By correlating external and internal model behavior across a diverse population of models, we gather evidence for when particular internal behaviors relate to rule learning.

- We illustrate how these correlational effects can be used for interpretability by identifying attention patterns that are intuitively and empirically associated with OOD behavior.
- We differentiate circuits that are *causally* necessary during inference from those correlated with a particular behavior. In particular, implementation of BALANCED is associated with either of two types of attention head, but ablation tests show that one type supports the OOD rule while the other suppresses it.
- Although we show that a model's attention patterns can predict its OOD behavior, ablating these circuits barely changes ID behavior. Furthermore, we find no significant correlation between the effect of ablating attention ID and OOD (Appendix A). We argue that some correlational interpretability studies can predict model behavior missed by causal studies, challenging the exclusive gold standard of causal intervention in validating understanding.

2 Related Work

Prior work in mechanistic interpretability has investigated the roles of individual model components [10, 2]. Toy problems such as modular addition [9] have provided a fruitful playground for identifying functions of circuits within models and characterizing how these circuits are used to solve problems. The influence of individual model components on models' behavior is sometimes assessed using causal intervention methods, such as pruning (also known as zero ablation), in which a particular set of weights or activations is set to zero; see e.g. [14] for discussion of these methods and their limitations.

Inductive bias toward hierarchical generalization behavior has been previously explored both in balanced parentheses (Dyck) settings [11, 1, 8, 13] and in natural language settings [5, 8]. In particular, Murty et al. [8] found that training well past in-distribution saturation increased tendencies toward hierarchical OOD generalization behavior. Wen et al. [13] showed that models with attention patterns obeying certain broad constraints can successfully achieve length generalization in a bounded Dyck setting, suggesting that many possible attention patterns can result in ultimately similar behavior.

We note that our approach involves a supervised classification dataset with labels ambiguous between two rules (EQUAL-COUNT and BALANCED), rather than next token prediction for the BALANCED rule only. This allows convenient measurement of the extent to which a model tends to implement one possible strategy versus another.

3 Experiments

We use the experimental setting of Li et al. [4], which we refer to for further details.

Data We consider the task of mapping strings consisting of open and closed parentheses onto a binary classification label. The training data is ambiguous between two possible rules for distinguishing the binary output classes:

- Every *positive* example obeys the BALANCED rule (its parentheses form a valid set of pairings) and so necessarily also obeys the EQUAL-COUNT rule (it has the same number of open and closed parentheses).
- Every *negative* example is a negative example of both rules: it has a different number of open and closed parentheses, and so must implicitly be unbalanced as well (Table 1).

As a result of our training data construction, any unbalanced sequence with the same number of open and closed parentheses is OOD (ex: `))(((()` in Table 1). These OOD examples will be labeled negative under the BALANCED rule but positive under the EQUAL-COUNT rule. Thus, we can study the particular rule each model relies on to generalize OOD by studying how it labels these samples. Our OOD test set consists of 1K datapoints that are labeled negative under BALANCED but positive under EQUAL-COUNT.

	Equal-count	Unequal-count
Balanced	()(())	DNE
Unbalanced	))((()	(())

Table 1: Examples of parentheses sequences in our ID dataset (balanced/equal-count and unbalanced/unequal-count parentheses) and our OOD group (unbalanced/equal-count).

Models and training We train transformers using the minGPT architecture [3] with hidden dimension 64, causal attention, and no dropout. We perform a grid sweep across several hyperparameters, varying weight decay (0.0, 0.001, or 0.01) and number of layers (1, 2, or 3). For each choice of hyperparameter, we run 15 training runs with learning rate 0.0001, using 5 initialization random seeds and 3 dataset ordering shuffle seeds. Our transformer models are trained on a binary classification task, with the classification decoded from the index of the EOS token at the decoder layer. Our task *does not involve language modeling*; the training signal comes from a single boolean prediction value.

4 Methods

Uniform Ablation To investigate the extent to which the model relies on attention patterns, we record the ID and OOD accuracy of models with *uniform* attention ablation: We replace each attention head with a "flat" attention head, in which each token attends equally to each prior token and to itself. This method preserves the existence of any effective bias contributed by attention, while removing any influence of the particular input on the Q and K matrices. It has been shown in [13] that uniform attention is both theoretically and empirically compatible with success at a balanced parenthesis length generalization task, which is similar (though not identical) to our OOD generalization task.

Note that when we ablate heads, we ablate *all* attention in the top layer. Empirically, we find that the stereotyped head attention patterns we analyze occur only in the top layer; by flattening the entire top layer, we therefore control for all head types simultaneously and uniformly across models.

Attention Head Behaviors We observe four distinctive types of attention head. These head types, which are seen only in the final layer of a model, relate to the attention patterns of the EOS token (likely because our classification method relies on decoding the activation at this token).

Some heads preferentially attend to all **open brackets** or all **close brackets**. Other heads attend to tokens on the basis of *depth*. We define $o(i)$ and $c(i)$ as the number of open and closed brackets, respectively, that appear prior to or at index i. Then, we have the *depth*, $d(i) = o(i) - c(i)$, and we observe attention heads which attend preferentially to all **negative depth** or all **nonnegative depth** tokens. Note that *a negative depth token can only occur in an unbalanced string.*

We classify a head as an open-bracket head if, on at least 50% of OOD datapoints, the EOS token preferentially attends to open brackets (i.e., attends more strongly to all open brackets than to *any* close bracket). Close-bracket attention heads are defined similarly. Likewise, we classify a head as a (non)negative depth head if on at least 50% of OOD datapoints, the EOS token preferentially attends to (non)negative depth tokens. Example attention patterns are shown in Figure 1.

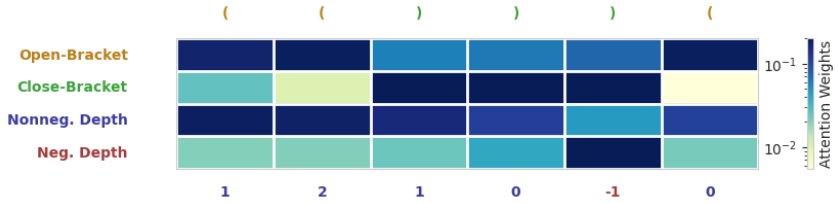

Figure 1: Attention patterns at the EOS token for different example attention heads on input ((()))(. Depths are shown in the bottom row of labels.

5 Results

Although transformer models achieve high ID accuracy, OOD generalization behavior, varies substantially. Some models showed a preference for the BALANCED rule, classifying most OOD strings as negative (since they are unbalanced). Other models learned the EQUAL-COUNT rule, classifying OOD strings as positive (since they have equal numbers of open and close parentheses). Still other models do not adhere closely to either rule.

For consistency, *throughout this paper we measure OOD accuracy with respect to the* BALANCED *rule*, i.e., with 0% corresponding to perfect EQUAL-COUNT and 100% corresponding to perfect BALANCED.

5.1 The BALANCED rule

The hierarchical BALANCED rule is associated with two attention head types: nonnegative and negative depth heads (Figure 2). In other words, models in which either of these head types is present tend to have higher OOD accuracy than models containing neither head type. It is unsurprising that depth-related computations would be more common among models which learn BALANCED, as a string with equal counts of (and) is balanced if its depth at each token index is nonnegative.

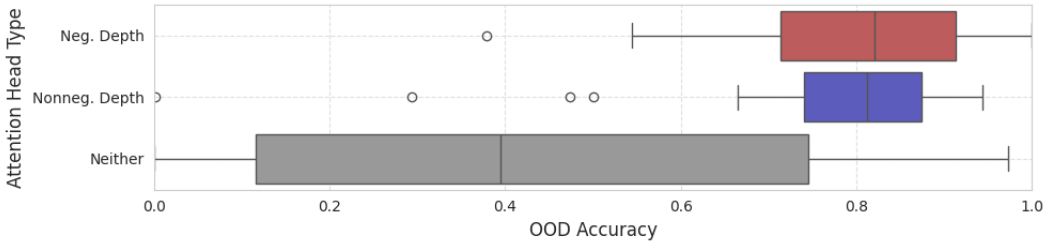

Figure 2: OOD accuracy of 2 and 3-layer models with each head type. Models containing both head types are counted twice.

Causal interventions While both depth-detecting head types are correlated with BALANCED, causal interventions reveal deep differences between these implementations. In our causal experiments, we consider the effect of the attention distribution (that is, the effect of Q and K matrices) rather than of the entire attention layer, so these distributions are ablated by setting all tokens to a uniform weight. Under a typical mechanistic model, if these heads are used to implement BALANCED, ablating them should damage OOD generalization. However, while ablating attention decreases OOD accuracy—as expected—when the ablation includes negative depth heads, it *increases* OOD accuracy when the ablation includes nonnegative depth heads (Figure 3).

Causal verification here conflicts with correlational verification, potentially indicating that nonnegative depth heads are vestigial circuits or spandrels—or used in ways that ablation methods cannot accurately reflect. Previous work indicates that vestigial circuits are pruned by weight decay, but we find that nonnegative depth heads are instead promoted by weight decay (Figure 4).

We therefore posit that these heads either develop as spandrels (side effects of learning BALANCED) or that ablation is reflecting the entanglement of these modules rather than their specific utility in an isolated rule-applying circuit. We leave the evaluation of these hypotheses to future work.

5.2 The EQUAL-COUNT rule

Under the EQUAL-COUNT rule, OOD strings are classified as positive, since they contain an equal number of (and). We say a model learns the EQUAL-COUNT rule if its OOD accuracy is under 10%, meaning it positively labels OOD strings.

One may expect this rule to be associated with the presence of open-bracket and close-bracket heads under the assumption that heads would accumulate counts of each symbol which can be compared.

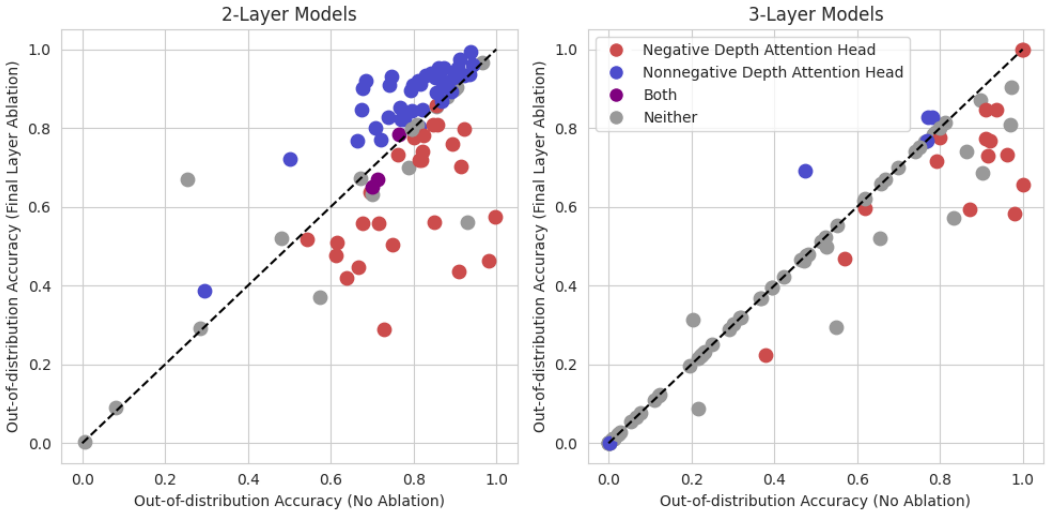

Figure 3: OOD accuracy before and after applying uniform attention ablation the top layer. Although both nonnegative and negative depth heads are associated with BALANCED, only the latter attention pattern supports that rule causally under ablation.

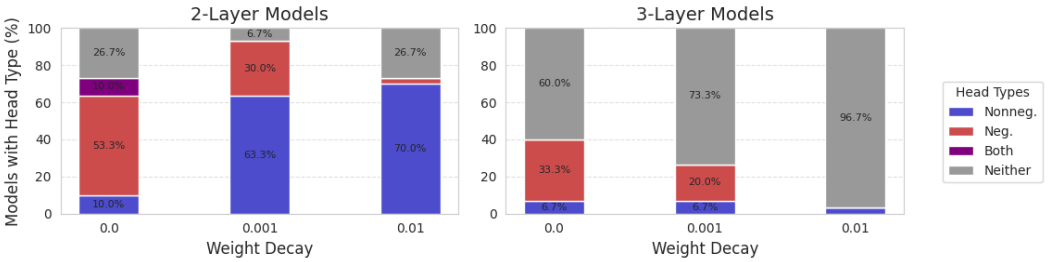

Figure 4: Percentage of 2-layer and 3-layer models containing each head type, by weight decay.

However, among 1-layer models trained without weight decay, these heads are *less* common among models which follow the EQUAL-COUNT rule. Furthermore, we determine via ablation experiments that these models *do not rely* on open- or close-bracket-detecting attention to classify either ID or OOD examples: among models which follow the EQUAL-COUNT rule, uniform attention ablation has negligible effect on ID and OOD accuracy.

Conclusions

Causal tests offer a complex view on how exactly internal mechanisms affect model judgements both in- and out-of-distribution. By using a large number of models to measure correlations between model behavior and internal mechanisms, we find that stereotyped attention patterns can be used to predict model behavior even when causal interventions provide unclear results.

Acknowledgments and Disclosure of Funding

This work was enabled in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. JK is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 2140743. DAM acknowledges support from the Kempner Institute, the Aramont Fellowship Fund, and the FAS Dean's Competitive Fund for Promising Scholarship. MW is supported by the Effective Ventures Foundation, Effektiv Spenden Schweiz, a Superalignment grant from OpenAI, and the Open Philanthropy Project.

References

[1] J. Ebrahimi, D. Gelda, and W. Zhang. How can self-attention networks recognize Dyck-n languages? In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4301–4306, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.384. URL https://aclanthology.org/2020.findings-emnlp.384.

[2] N. Elhage, N. Nanda, C. Olsson, et al. Zoom in: An introduction to circuits, 2021. URL https://distill.pub/2020/circuits/zoom-in/.

[3] A. Karpathy. MinGPT transformer model, 2020. URL https://github.com/karpathy/minGPT.

[4] V. Li, J. Kaufmann, D. Alvarez-Melis, and N. Saphra. Twin studies of factors in OOD generalization. Workshop on Scientific Methods for Understanding Deep Learning (NeurIPS), 2024.

[5] R. T. McCoy, R. Frank, and T. Linzen. Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks. *Transactions of the Association for Computational Linguistics*, 8:125–140, 01 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00304. URL https://doi.org/10.1162/tacl_a_00304.

[6] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT, 2023. URL https://arxiv.org/abs/2202.05262.

[7] A. Mueller, J. Brinkmann, M. Li, S. Marks, K. Pal, N. Prakash, C. Rager, A. Sankaranarayanan, A. S. Sharma, J. Sun, E. Todd, D. Bau, and Y. Belinkov. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability, 2024. URL https://arxiv.org/abs/2408.01416.

[8] S. Murty, P. Sharma, J. Andreas, and C. Manning. Grokking of hierarchical structure in vanilla transformers. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.38. URL https://aclanthology.org/2023.acl-short.38.

[9] N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

[10] C. Olah, N. Cammarata, G. G. Ludwig Schubert, M. Petrov, and S. Carter. Zoom in: An introduction to circuits, 2020. URL https://distill.pub/2020/circuits/zoom-in/.

[11] M. Suzgun, S. Gehrmann, Y. Belinkov, and S. M. Shieber. Memory-augmented recurrent neural networks can learn generalized Dyck languages, 2019. URL https://arxiv.org/abs/1911.03329.

[12] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

[13] K. Wen, Y. Li, B. Liu, and A. Risteski. Transformers are uninterpretable with myopic methods: a case study with bounded Dyck grammars. URL https://par.nsf.gov/biblio/10489627.

[14] F. Zhang and N. Nanda. Towards best practices of activation patching in language models: Metrics and methods, 2024. URL https://arxiv.org/abs/2309.16042.

A Effects of Causal Intervention on Attention

Our causal experiments involve uniform ablation of the attention distribution in all attention heads in the final layer. We chose to ablate the final layer due to our empirical finding that causal intervention on earlier layers had comparatively limited impact on in-distribution performance and on generalization behavior.

We compare the effects of ablating attention on the in-distribution and out-of-distribution accuracy, controlling for number of layers and for use of weight decay. We compute Pearson correlation coefficients to determine whether ablation effects on ID vs OOD accuracy are correlated. We find no statistically significant correlation, though the coefficients are provided in Table 2.

	1 Layer	**2 Layers**	**3 Layers**
With WD	$*$	-0.09	0.21
Without WD	-0.13	-0.13	0.31

Table 2: Correlation coefficients (ρ) for each weight decay and number of layers. Values shown are not statistically significant. 1-layer models trained with weight decay ($*$) are omitted because ablation did not affect ID or OOD accuracy for 59 of 60 models. The presence of a single affected model produced a correlation coefficient of -1.0, but we caution against interpretation of this value as meaningful evidence of anticorrelation.

B Debunking Challenge Submission

B.1 What commonly-held position or belief are you challenging?

Provide a short summary of the body of work challenged by your results. Good summaries should outline the state of the literature and be reasonable, e.g. the people working in this area will agree with your overview. You can cite sources beside published work (e.g., blogs, talks, etc).

Directly modifying a model's behavior, through interventions such as ablation and activation patching, is among the most commonly used techniques for analyzing a model's behavior. These "causal" methods go by numerous names, including causal tracing, activation patching, interchange intervention, and causal mediation analysis [12, 6, 14]. They are intended to localize a model's behavior within a particular circuit by intervening on that circuit and demonstrating effects on the behavior. Causal interventions are frequently upheld as the gold standard in interpretability research [7].

In contrast, observational methods for understanding model behavior are dismissed as too "correlational." It is of course true that observational methods require care to avoid overinterpreting results: as the saying goes, correlation does not imply causation. However, so-called causal methods also suffer from the same problem: caution is necessary to avoid overinterpreting results from a certain activation patching technique that may not transfer to novel settings [14].

B.2 How are your results in tension with this commonly-held position?

As previously highlighted by Zhang et al. [14], interventions that have an effect in distribution might not have similar effects OOD. In fact, we find that there is no significant correlation between the effect of attention flattening ID and OOD. By instead correlating attention distributions with OOD behavior, we show that correlational evidence can be strong enough to validate a connection between internal behavior and external outputs.

We show that attention distributions focused on a single symbol type can predict our FIRST-SYMBOL heuristic rule, whereas attention distributions that track depth predict a BALANCED rule. Furthermore, the head types predict how much ablating Q and K for that head through flattening will damage OOD performance, even though ID interventions cannot predict the same outcome.

B.3 How do you expect your submission to affect future work?

We emphasize that we are by no means opposed the use of causal methods for understanding model behavior: these have been a fruitful source of knowledge in interpretability, and indeed we use them in our own paper. But we challenge the notion that causal methods are the *only* valid interpretability technique.

Although the interpretability field increasingly holds causal intervention to be the gold standard of evaluation, intervention is not always a gold standard in traditional sciences like biology. In genetics, for example, correlational work through comparisons of fraternal and identical twin populations represent a gold standard; by contrast, interventional studies that rely on genetic engineering may actually be *less* informative because single-gene editing can sabotage unrelated processes through complex interactions between genes.

Unfortunately, the large random variation experiment we provide is intractable for many applications. However, claims in this area should depend more on control models that exhibit different OOD behavior, so as to better study the relationship between internal circuits and generalization under distribution shift. In particular, a better understanding of how internal behavior shapes OOD outputs can lead us to a future in which our field takes seriously the relationship between our interpretations and predicting model behavior, which is a key objective of interpretability research that seemingly cannot be achieved through causal interventions alone.