

Can Interpretation Predict Behavior on Unseen Data?

Victoria R. Li¹, Jenny Kaufmann¹, Tian Qin¹,
Martin Wattenberg¹, David Alvarez-Melis¹, Naomi Saphra^{1,2}

¹Harvard University, ²Boston University

Correspondence: vrli@alumni.harvard.edu, nsaphra@bu.edu

Abstract

Interpretability research often predicts model responses to targeted mechanistic *interventions*. But can we predict responses to unseen *input data*? We propose and demonstrate this alternate objective by using model internals to predict their out-of-distribution (OOD) behavior. We train hundreds of Transformers on simple synthetic tasks, where perfect in-distribution accuracy is compatible with multiple OOD generalization rules. We successfully use attention patterns—observed only on in-distribution data—to predict which rule each model follows on OOD data. Our experiments decouple the mechanistic faithfulness of our interpretation from its predictive value; ablations reveal such internal patterns can *suppress* rather than *support* the rule they predict, showing observational analysis can forecast behavior even when causal analysis fails to support a simple cause-effect link. Our findings are a proof-of-concept for a new interpretability objective: understanding model internals to predict behavior and assess reliability under distribution shift.

1 Introduction

When do we understand a system? One standard, that of the classic scientific method (Hepburn and Andersen, 2021), requires *testable predictions of behavior under unseen conditions*. Interpretability research is often evaluated by predicting the effect of a test-time mechanistic intervention (Geiger et al., 2025) such as activation steering (Subramani et al., 2022; Tan et al., 2025; Liu et al., 2024) or patching (Makelov et al., 2024; Kramár et al., 2024; Todd et al., 2024; Vig et al., 2020). In contrast, interpretability researchers rarely infer a model’s rules by predicting its outputs on unseen *data* inputs. We will demonstrate the latter objective.

This work proposes a new mission for interpretability. We argue that progress in model understanding should be measured not only by causal-mechanistic intervention, but also by predicting

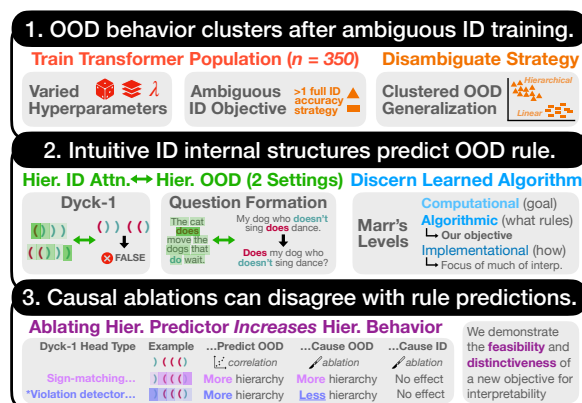


Figure 1: Our findings.

model behavior on unseen data.

If we could predict model behavior under distribution shift, we would unlock new interpretability applications. Well-understood instruments come with engineering tolerances—specified edge cases where the instrument may fail. By providing tolerances for AI models, we might guide training, offer recommendations for reliable use, or flag potential deployment failures. For example, structures associated with specific languages and tasks may provide clues for whether a model can compose them to fluently handle a given task in a particular language. Can we use interpretation to predict a model’s response to unseen data?

Using synthetic tasks, we show that simple analysis of attention patterns can reveal a model’s generalization rule *even if* they are not used straightforwardly in its implementation. In other words, our interpretations may not be *faithful* in a causal, mechanistic sense—but they nonetheless reveal informative traces left by the algorithm. Interpretability-based intuitions then allow us to guess the algorithm being executed, and therefore to predict how models treat unseen data.

We view the model as a whole and use its internals to understand resulting outputs—a *holistic interpretability* approach. We wish to predict a

model’s *out-of-distribution* (OOD) behavior using only its internal representations of *in-distribution* (ID) data. To this end, we train a large set of Transformer-based classifiers with perfect ID validation accuracy and diverse OOD behavior in a synthetic setting. Our model population is trained on data with an ambiguous classification rule: models can achieve perfect ID accuracy using either a parenthesis **counting** rule (EQUAL-COUNT) or a hierarchical parenthesis **nesting** rule (NESTED). We use an OOD test set to determine which rule each model follows. (See Appendix A for a glossary.)

In Marr’s terms (Marr, 2010), our objective targets the *computational* and *algorithmic*—rather than implementational—levels of analysis. We treat internals not as causal mechanisms but as signatures hinting at what structures the model represents for its rule. Full model understanding requires multiple levels of analysis.

Overall, motivated by epistemic concerns and actionable interpretability objectives, we train a population of 270 models to classify Dyck-1 parentheses strings (Sec. 2), with an additional autoregressive English question-formation setting (Appendix L). We find (Figure 1):

- **Independently trained models cluster around systematic generalization rules.** We visualize model OOD judgments, finding clusters that implement linear and hierarchical strategies to varying degrees (Sec. 3). In the Dyck-1 setting, some unregularized training runs learn a heuristic instead of a valid rule, but only apply this heuristic OOD (Sec. 3.1). Models can pass through transient heuristic phases in training, suggesting that vestigial circuits still affect OOD judgments. In addition to weight decay, model depth and seed influence rule learning (Sec. 3.2).
- **OOD generalization rules are intuitively predicted by internal representations of ID data.** We identify heads that encode hierarchical structure in their attention activations (Sec. 4.1). In the Dyck-1 setting, we show that models with these heads usually apply NESTED on unseen OOD data (Sec. 4.2). In the second, autoregressive question-formation setting, we also find a hierarchical internal attention pattern that correspondingly predicts hierarchical OOD generalization (Sec. 4.2.1).
- **Internal structures predict generalization rules even when the rule’s implementation doesn’t rely on them.** We differentiate circuits

causally necessary in implementing a rule from those *correlated* with the rule across a model population. In our autoregressive setting (Appendix L), ablation tests leave OOD behavior unchanged. In our Dyck-1 setting, we find that our interpretation’s predictive power is clearly decoupled from its causal faithfulness. NESTED is associated with multiple types of hierarchical attention heads, but some types suppress—rather than support—the NESTED rule (Sec. 4.3). Furthermore, the effect of ablation is only weakly correlated between ID and OOD data conditions, calling into question the robustness of causal interpretability findings (Sec. 4.3.3).

1.1 Philosophical Motivation

The interpretability of attention is contentious (Bibál et al., 2022): specific attention patterns are rarely necessary for model performance, either in theory (Wen et al., 2023) or in practice (Jain and Wallace, 2019; Serrano and Smith, 2019). At its heart, this debate is philosophical—about what makes an interpretation faithful. We take a different epistemic position: our objective is to predict a model’s response to unseen inputs, not to mechanistic interventions.

Causal interpretability and its limits. In modern interpretability, causality is often held as the gold standard for evaluating understanding. Researchers causally test each module by ablating it and measuring damage to model performance. However, the resulting findings are often compatible with multiple analyses which can only be differentiated by expensive, precise ablations. Moreover, proposed interpretations may not transfer to new settings (Stander et al., 2024; Zhang and Nanda, 2024; Makelov et al., 2024). By focusing on causal intervention alone to test the faithfulness of explanations, the literature on understanding attention has produced skepticism and fierce debate (Bibál et al., 2022; Wen et al., 2023).

Correlational analyses in the sciences. While interpretability often demands causal tests, other sciences frequently use correlational analyses. In genetics, for example, correlational studies can link inherited traits like red hair and painkiller tolerance (Liem et al., 2005). While causal studies may seem epistemically sounder in principle, engineered approaches like single-gene editing can sabotage un-

related processes through complex interactions.¹ Why should it be any easier to isolate mechanisms in highly nonlinear neural networks?

Related work on causal predictions OOD.

Some existing work uses causal analysis to predict model computations on OOD inputs. The visual illusion constructed by Gurnee et al. (2026) validated a causal attribution by proposing a specific failure mode based on inspecting attention patterns. Huang et al. (2025) showed that ID causal features predict errors better than ID non-causal features when used to simulate OOD scenarios. By contrast, our setting demonstrates that some causal experiments can mislead as to the predictive power of representational structures.

Interpretability and scientific realism. Of the many controversies in philosophy of science, few are more central than the rivalry between realism and instrumentalism (Okasha, 2016). Realists claim that concepts used in scientific models, from quarks to gravity, are specific objects and forces acting on the world (Putnam, 1975). Instrumentalists, by contrast, argue that science’s epistemic goal is not to uncover fundamental truths about the world, but to make predictions about their observable outcomes (Duhem, 1954). In the philosophy of mind, instrumentalists may even deny that internal beliefs, desires, or intentions are real phenomena, while still acknowledging that these concepts might help predict a person’s behavior (Churchland, 1981; Dennett, 1991, 1989). Our objective is an instrumentalist one: rather than insisting on a true hierarchical mechanism, we treat hierarchical structure as part of our scientific model of network behavior, sufficient if we can infer it from the traces it leaves.

Like other observational studies, we leverage correlations over model populations to test our understanding. Hidden representations can *proxy* a model’s algorithm even when they are not employed in its implementation, enabling our computational—rather than implementational—level of analysis (Marr, 2010).

¹Haapaniemi et al. (2018) provide a famous example of how targeted gene edits can produce spurious results. They detail how CRISPR gene editing can activate p53, a tumor suppressor. As a result, experimentalists using the method inadvertently select for cells with defective p53 responses, forming a confounder for research on cancer genetics.

2 Methods and Experiment Setup

To predict models’ OOD behavior from their representation of ID data, we first create models with varied OOD generalization to study. We train all these models on a dataset compatible with at least two distinct OOD generalization rules (Sec. 2.1). We study the variation produced by training many models with different hyperparameters and random seeds (Sec. 2.2).

2.1 Data setting

Our setting is inspired by work on systematic generalization from ambiguous training rules (McCoy et al., 2020a,b; Murty et al., 2023). Our primary dataset is based on the Dyck-1 parentheses-balancing task (Suzgun et al., 2019; Ebrahimi et al., 2020; Murty et al., 2023; Wen et al., 2023). Unlike standard parentheses-balancing settings, our training dataset is compatible with either EQUAL-COUNT, an unordered counting rule, or NESTED, a hierarchical parentheses-balancing rule.

Our Dyck-1 models are classifiers, not sequence generators. They verify that the input follows some rule and output a binary class $y \in \{\text{True}, \text{False}\}$. Each input is an n -length sequence $s = s_1 s_2 \dots s_n$, where $s_i \in \{([,])\}$. Using $\mathbf{1}(\cdot)$ as the indicator function, let the number of open $[$ and close $]$ tokens that appear from index 1 to j be $o(j) = \sum_{i=1}^j \mathbf{1}(s_i = [)$ and $c(j) = \sum_{i=1}^j \mathbf{1}(s_i =)$, respectively. Each rule labels s as follows.

- EQUAL-COUNT is True if s has the same number of open and close parentheses:

$$o(n) = c(n) \quad (1)$$

- NESTED is True if s is a recursively nested tree. All NESTED sequences are also EQUAL-COUNT, but in addition to Eq. 1, NESTED s fulfills:

$$\forall j \in \{1, \dots, n\} \quad o(j) \geq c(j) \quad (2)$$

Our training set is compatible with both EQUAL-COUNT and NESTED: each input satisfies *both* or *neither* of Equations 1 and 2. Thus, *a model can perfectly classify ID data by either rule*. We test which rule each model learned using an OOD set of sequences that fulfill Equation 1 but not Equation 2: they have the same number of $[$ and $]$ tokens, but not in a properly nested order. We define accuracy according to NESTED, so a model has 100% OOD accuracy if it labels all OOD examples False.

Implementation We create ID sequences where NESTED and EQUAL-COUNT agree and OOD sequences where they disagree. Each sequence is generated randomly with length sampled $n \sim \text{Binomial}(40, 0.5)$ (see Appx. D). The 1K OOD examples fulfill EQUAL-COUNT but not NESTED (e.g., $\text{((((($)). Our 1M-example train set, and 1K-example ID validation set are label-balanced, containing 50% negative examples which fulfill neither EQUAL-COUNT nor NESTED (e.g., $\text{((($)) and 50% positive examples which fulfill both EQUAL-COUNT and NESTED (e.g., $\text{((($)).

2.2 Models and attention

We train Transformers with causal self-attention and input length $L = 42$. Each sequence s consists of a BOS (beginning-of-sequence) token at s_0 , a sequence of $n \leq 40$ parentheses, an EOS (end-of-sequence) token at s_{n+1} , and $L - n - 2$ padding tokens after EOS. We output a binary class $\hat{y} \in \{\text{True}, \text{False}\}$ at the final layer EOS token.

Because our models output a classification label at the EOS token, we consider attention activations at the EOS index $n + 1$. For a given head’s attention activations $A \in \mathbb{R}^{L \times L}$ on input s , we define $a_{\text{EOS}}(i)$ as attention to token s_i at index $i \in \{1, \dots, n\}$:

$$a_{\text{EOS}}(i) = A_{n+1,i}. \quad (3)$$

Implementation We train a population of classifier models based on the minGPT architecture (Karpathy, 2020) with hidden dimension 64. Following Vaswani et al. (2017), we use reshaping for multi-head attention, so the model’s overall parameter count is the same regardless of per-layer head count W . We set the learning rate $\eta = 0.0001$ with no dropout. All trained models stabilize to an ID validation accuracy of at least 99% after at most 900K training examples (Appx. Fig. 13).

We grid sweep over other hyperparameters to create a diverse model population. We train models with depths $D \in \{1, 2, 3\}$ layers and widths $W \in \{2, 4\}$ attention heads per layer. We also vary optimizer weight decay $\lambda \in \{0, 0.001, 0.01\}$. Per combination, we train models with 5 random seeds for weight initialization and 3 for dataset shuffle order. This sweep results in 15 models per hyperparameter configuration and 270 models in total.

2.3 Additional Question Formation Setting

To complement our classifier experiments, we also study autoregressive language models on English question formation (McCoy et al., 2020a; Linzen

et al., 2016; Qin et al., 2025). Models read a declarative sentence and generate the corresponding yes/no question, which requires moving the *matrix* (root-clause) auxiliary to the front. ID training data is constructed so that the linear “move the first auxiliary” rule and the hierarchical “move the matrix auxiliary” rule produce identical outputs; OOD examples distinguish these strategies (Appendix Table 3). We analyze a population of 79 autoregressive Transformers, trained to perfect ID accuracy. Appendix L includes more detail.

3 Generalization Behavior

All Transformer models achieve high ID validation accuracy, but their OOD behaviors vary.

3.1 Models cluster by systematic rule

We visualize all models according to their OOD output probabilities (Fig. 2). In Dyck-1 (Fig. 2a), one cluster has near-0% OOD accuracy and another has high OOD accuracy, supporting existing claims (Qin et al., 2025; Zhao et al., 2024) that *rule-following is a categorical phenomenon, not a continuum*. Outside these two expected EQUAL-COUNT and NESTED clusters, a third cluster of near-identical judgments emerges—a heuristic-based cluster. In question formation (Fig. 2b), models likewise separate into a linear-rule cluster (low OOD accuracy) and a hierarchical-rule cluster (high OOD accuracy), supporting the categorical nature of rule-following.

3.1.1 The heuristic cluster

In the Dyck setting, the outlier cluster represents a simple heuristic, FIRST-SYMBOL: a sequence is True if its first symbol is (and False if its first symbol is) . This heuristic correctly labels all positive examples, but only half of negative examples (Appx. Table 2). Although some models use this heuristic OOD, they do *not* apply it to ID data or they would not achieve high ID validation accuracy.

We hypothesize that the FIRST-SYMBOL heuristic is implemented by a vestigial circuit, which the model has learned to suppress on ID—but not OOD—data. Vestigial circuits are mechanisms needed early but not later in training. Prior literature (Tessier et al., 2022; Chen et al., 2023; Doshi et al., 2024) suggested that vestigial circuits are pruned by weight decay, which aligns with our observations of the FIRST-SYMBOL cluster: Note the FIRST-SYMBOL circuit only governs 1-layer models trained *without* weight decay, whereas all

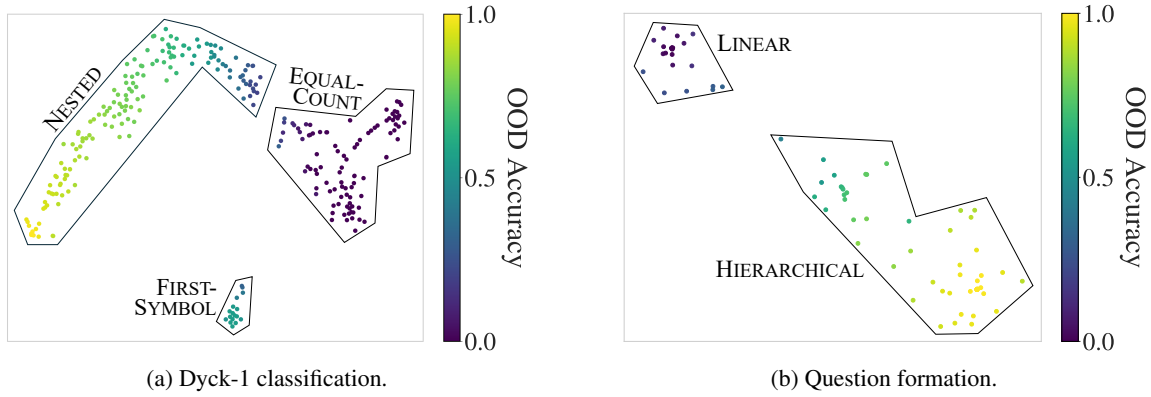


Figure 2: **Independently trained models cluster around systematic generalization rules.** t-SNE of each model’s OOD output probabilities. In Dyck-1 (a), models separate into NESTED, EQUAL-COUNT, and a small FIRST-SYMBOL heuristic cluster. In question formation (b), models separate into a LINEAR cluster (low OOD accuracy) and a HIERARCHICAL cluster (high OOD accuracy).

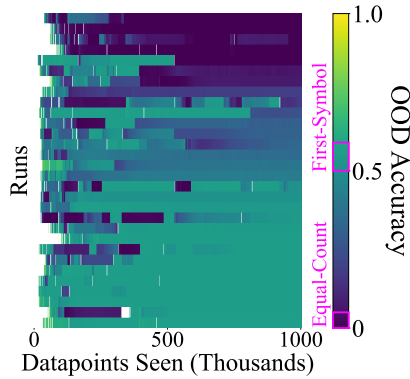


Figure 3: **Unregularized models experience transient heuristic training stages.** Intermediate checkpoints of unregularized 1-layer models, sorted by final OOD accuracy. Checkpoints that cannot model the training distribution (ID accuracy < 0.99) are shown in white cells. The color bar is marked at FIRST-SYMBOL and EQUAL-COUNT accuracy levels. (More analysis in Appx. F.)

1-layer regularized models follow EQUAL-COUNT. As further evidence of vestigiality, 1-layer EQUAL-COUNT models (Fig. 3) often pass through apparent heuristic phases while training, during which OOD accuracy matches that of FIRST-SYMBOL models. These findings suggest that *a vestigial circuit—with no detectable impact on ID behavior—can still affect OOD judgments.*

3.2 Factors in rule selection

Varying hyperparameters for the Dyck setting, we find the dominant factors in model rule selection are model depth and weight decay. Appx. E details other factors, including an extension of existing findings (Abnar et al., 2020; McCoy et al., 2020a; Tran et al., 2018; Saphra and Lopez, 2020) that recurrent architectures favor hierarchical structure.

Model depth Appx. Fig. 9 shows 2- and 3-layer models can learn either NESTED or EQUAL-COUNT, depending on hyperparameters and random seed. In contrast, 1-layer models only learn FIRST-SYMBOL or EQUAL-COUNT. Shallow models do not learn hierarchy, while 2-layer models tend more toward hierarchy than 3-layer models. This inverted U of hierarchical inductive bias by depth mirrors previous work (Murty et al., 2023).

Weight decay Regularization pushes models toward more consistent systematic rules, concentrating OOD behavioral distributions (Appx. Fig. 9). In our setting, non-zero weight decay increases preference for EQUAL-COUNT in 1- and 3-layer models but for NESTED in 2-layer models. These findings imply regularization promotes systematic rules over memorization, but *which* rule is simplest depends on hyperparameters, complicating reductionist narratives of universally simple rules.

4 Interpretability Predicts Behavior

We inspect the attention activations of each head and use intuitive explanations of how ID inputs are processed to predict model decisions on unseen OOD inputs. Models following NESTED display hierarchical attention patterns on ID data; despite shared outputs, they can employ subtly different internal mechanisms with different causal roles. Hierarchical attention patterns predict that a model will follow NESTED *even when* these patterns do not directly implement it. Appx. I confirms that attention patterns add predictive value beyond hyperparameter settings alone.

4.1 Hierarchical attention patterns

Certain heads exhibit interpretable attention patterns at the EOS token. Specifically, these patterns indicate whether a token position j follows the condition in Equation 2 that $o(j) \geq c(j)$. We will connect these intrinsically hierarchical heads to the NESTED generalization rule.

4.1.1 Tracking violations

We identify attention patterns that encode an input’s hierarchical structure. These patterns attend to a token at position j depending on if it violates Equation 2, i.e., if $o(j) < c(j)$. If a sequence contains any such **violation token**, the sequence is not NESTED. In particular, *all* examples in the OOD set—EQUAL-COUNT, but not NESTED—must have at least one violation token. Intuitively, attention patterns which track these violations can signal that a model is closer to NESTED.

There are two ways for an attention output to **track violations** on a given sequence: a head can preferentially attend either to violation or to non-violation tokens. For input s ,

- An attention head favors *violation tokens* on s if there exists threshold $t > 0$ such that:

$$\forall j \in \{1, \dots, n\} : a_{\text{EOS}}(j) \geq t \text{ iff } o(j) < c(j) \quad (4)$$

- An attention head favors *non-violation tokens* on s otherwise, i.e., there exists $t > 0$ such that:

$$\forall j \in \{1, \dots, n\} : a_{\text{EOS}}(j) \geq t \text{ iff } o(j) \geq c(j) \quad (5)$$

Examples of each pattern are shown in Appendix Fig. 6. We say an attention head is **tracking violations on a given input** if it favors *either* violation *or* non-violation tokens on that input.

4.1.2 Hierarchical head types

Some heads reliably track Equation 2 violations on each input, making them hierarchical heads across a dataset. We define a head as a **hierarchical head** on a given dataset if it tracks violations (Appendix Figure 6) on at least 80% of relevant sequences—those containing both violation and non-violation tokens. Heads which track violations on ID examples typically behave as hierarchical heads OOD (i.e., 77% of the time). We find two hierarchical head types:

- **Violation-detecting heads** consistently assign more attention to violations of Equation

2, i.e., positions preceded by more] than [tokens.

- **Sign-matching heads** behave differently depending on whether the final token follows Eq. 2. If the final token violates Eq. 2, the head will attend to violations; if the final token follows Eq. 2 (as in all OOD sequences), the head will attend to non-violations.

Among models with ID hierarchical heads, 81% have sign-matching heads, 9% have violation-detecting heads, and 8% have both; only 2% of models with hierarchical heads have neither subtype. We therefore focus on these subtypes, which cover almost all ID hierarchical heads.

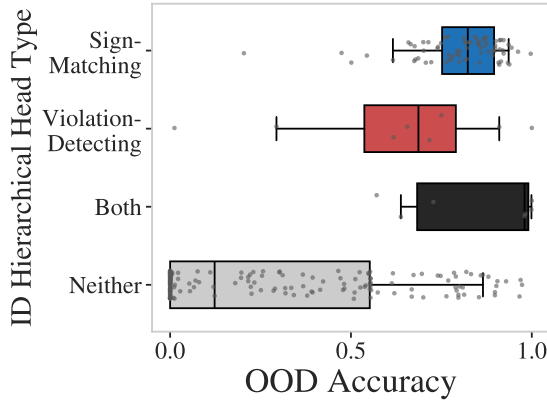
4.2 Hierarchical attention on ID data *predicts* hierarchical rules on OOD data

We return to our overarching question: Can these model internals intuitively suggest which function the model implements, thereby predicting its treatment of the unseen OOD distribution? If we claim to understand a model, we should be able to predict its behavior under unseen conditions. We now demonstrate that our attention-based intuitions can, in fact, reliably predict OOD model behavior.

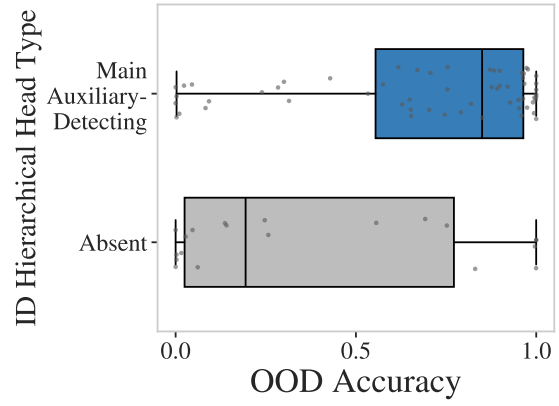
As seen in Fig. 4a, Dyck models which contain at least one ID hierarchical head—of either type—are more likely to follow the NESTED rule. This result holds separately across 2- and 3-layer models (Appx. I). These findings confirm our intuitive hypothesis: ID hierarchical representational structure is associated with OOD hierarchical generalization. These ID internal patterns, along with other ID attention statistics, can predict the OOD rule of held-out models: the hierarchical head is one instance of a broader ID predictive signal (Appx. H).

We find that 1-layer models, which notably *do not* learn the NESTED rule (Appx. Fig. 9), possess no hierarchical heads. In fact, hierarchical heads do not occur in the first layer of any model—possibly explaining why our models only learn NESTED if they have multiple layers.

All OOD sequences end with $o(n) = c(n)$, so a *sign-matching* head attends to non-violation tokens. We do not observe any *ID violation-detecting heads* switch to favoring non-violation tokens on OOD sequences. However, 25% of *ID sign-matching heads* do switch to consistently favoring violations OOD. Given that a mechanism can behave so differently OOD, it would be challenging to describe any internal mechanism in a way that applies across new domains. Nonetheless, these structures pro-



(a) Dyck-1: OOD accuracy of 2- and 3-layer models split by ID hierarchical head type. *Two* subtypes of hierarchical head—sign-matching and violation-detecting—predict higher OOD accuracy than models with neither.



(b) Question formation: OOD accuracy split by presence of an ID *Main Auxiliary-Detecting* head, which attends from the question-formation readout position to the matrix auxiliary. See more details in Appx. L.

Figure 4: **Hierarchical attention patterns predict hierarchical OOD generalization in two settings.** ID hierarchical heads, of either subtype, correlate with hierarchical OOD behavior in both our settings.

vide a *holistic* understanding of the algorithm implemented by the model. This inferred algorithm, unlike mechanistic interpretations tied to particular internal patterns, applies across distribution shift.

4.2.1 Question formation setting

In the question-formation population we also find an ID attention pattern that predicts hierarchy.

- **Main Auxiliary-Detecting heads** concentrate attention on the matrix (root-clause) auxiliary rather than the embedded-clause one.

Concretely, an ID sentence in this setting is “*My unicorn does move the dogs that do wait*”. Here, the ID output is “Does my unicorn move the dogs that do wait?” A Main Auxiliary-Detecting head for this sentence places substantial attention—more than 0.2 of its attention mass—on the matrix auxiliary “does” (the hierarchically-correct target, as opposed to the embedded-clause auxiliary “do”) just before the model outputs the question. We particularly identify heads in the first three layers of each model that exhibit this behavior on at least 80% of ID examples. The presence of such an ID hierarchical head in a question-formation model strongly corresponds to OOD behavior with hierarchical question-formation output (Fig. 4b). Like in the Dyck setting, we can make held-out predictions of OOD generalization based on ID attention-associated signals (Appx. H).

4.3 Hierarchical attention may not *cause* hierarchical rules on OOD data

In mechanistic interpretability, faithfulness is often evaluated through causal intervention. In the Dyck setting, hierarchical heads correlate with the NESTED rule, but correlation alone doesn’t establish them as causal mechanisms. To determine causality, we intervene on attention activations and examine how model performance responds. These experiments demonstrate that an attention pattern might correlate with a systematic rule without supporting it causally—in fact, we will see that the pattern may even, counter-intuitively, suppress the rule. Preventing models from displaying these attention patterns can thus enhance, rather than reduce, the correlated output behavior.

Although our interpretations are not causally faithful, they predict the model’s computation. Our results therefore support the notion that *causal faithfulness* is distinct from our objective of *holistic interpretability* as proposed in Sec. 1.1.

4.3.1 Ablation method

To test whether Dyck models rely on particular attention patterns, we measure accuracy after *uniform attention ablation*—forcing every attention activation to attend equally to all prior tokens. This intervention preserves all other components but strips violation-tracking or other attention patterns. A model could, in principle, still apply NESTED after ablation: Wen et al. (2023) showed uniform attention suffices to implement parentheses balancing. We apply the ablation to *all* heads simultaneously,

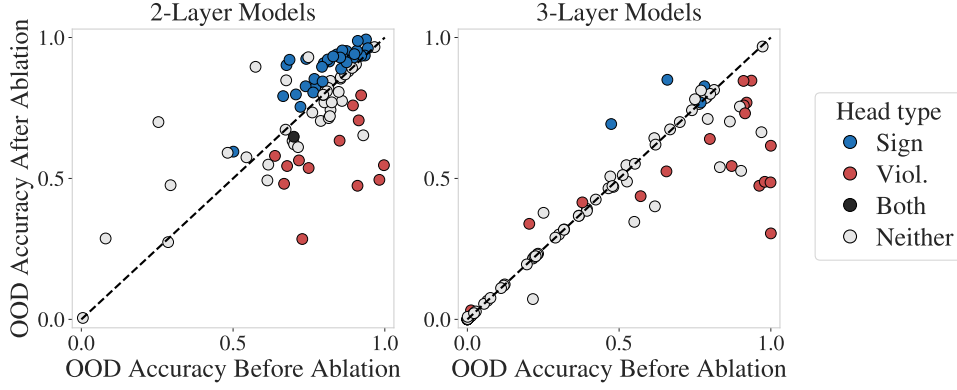


Figure 5: **Some hierarchical attention patterns damage the implementation of hierarchical rules (Dyck-1).** OOD accuracy before and after applying uniform attention ablation. Each point is a single model, colored by presence of an OOD sign-matching and/or violation-detecting head. Ablation damages OOD performance below the diagonal but improves it above. Although both head types are correlated with NESTED, only violation detection causally supports the rule; sign-matching attention *suppresses* the rule.

controlling for confounders from co-occurring head types. Note our results also hold for other ablation experiments ablating heads one at a time or using mean ablation (Appx. J).

4.3.2 Ablation results

We find that certain types of violation-tracking attention—although correlated with high OOD accuracy—actually lower the OOD accuracy of the model when present (Fig. 5). When we ablate the attention of models with OOD *violation-detecting* heads, we damage OOD accuracy, as might be expected if violation tracking is a key mechanism in implementing NESTED. By contrast, when we ablate the attention of models with OOD *sign-matching* heads, we actually *improve* OOD accuracy. The latter type of hierarchical head, although just as correlated with hierarchical generalization as the former type, is not a key mechanism in the rule’s implementation. Instead, it interferes with systematic hierarchical generalization.² Ablating Main Auxiliary-Detecting heads in the question-formation setting does *not* reverse the correlational signal: the ablation leaves OOD behavior essentially unchanged (Appx. L.4). Correlational predictive and causal analyses *can* either converge or diverge; the predictive objective we propose is achievable in either case.

Our findings resist simple narratives about the mechanisms which implement a model’s algorithm. Although all hierarchical heads are *correlated* with hierarchical generalization, not all of them *implement* that rule under causal tests. Therefore, even

simple models can be complex enough to resist causal analysis, but correlational interpretations still provide predictive value. The decoupling is robust to the choice of ablation: *mean* ablation and single-head ablation (Appx. J) reproduce the same qualitative pattern as the uniform-attention ablation shown in Fig. 5—ablating sign-matching heads increases hierarchical OOD accuracy, while ablating violation-detecting heads decreases it.

Unlike the circuit which hypothetically implements FIRST-SYMBOL, OOD sign-matching heads are *promoted* by regularization, occurring more frequently when models are trained with weight decay (Appx. Fig. 15). We therefore reject the position that sign-matching patterns are vestigial. Instead, we conjecture that these heads either develop as spandrels—side effects of learning NESTED—or they malfunction on OOD data. Module generalization failure is a ripe topic for future work.

4.3.3 Causal roles are data dependent

Interpretability research commonly tests proposed explanatory mechanisms by intervening on those mechanisms. We call this approach into question by demonstrating that a model can *respond to an ablation differently on OOD and ID data*. Uniform attention ablation has substantial effects on OOD accuracy (Fig. 5), but leaves ID validation accuracy nearly unchanged, reduced by only 0.6% on average (Appx. K); moreover, effects of ablation on ID and OOD accuracy are only weakly correlated ($\rho = 0.24$, $p < 0.01$). These results suggest that a model might be able to compensate for the loss of a mechanism on ID data, while still heavily relying on the mechanism on OOD edge cases. Although

²Recall that some ID sign-matching heads become violation-detecting heads on OOD data.

a model may be applying the same rule ID and OOD at the output, the distribution shift reveals brittle elements of its implementation. We posit models could contain many redundant backup circuits that compensate for ablation ID, but become unreliable under distribution shift. Without redundancy, model judgments become more sensitive to ablations on the remaining circuits.

These results suggest that a negative ablation result is weak evidence against a proposed mechanism. Our findings mirror recent cases where current interpretability techniques failed to generalize on novel domains. Not only are SAE dictionary features highly data-dependent (Paulo and Belrose, 2025), but steering research shows features that are causal on one distribution may not be causal on new data (Kissane et al., 2024; Kantamneni et al., 2025; Smith et al., 2025). These concerns have spurred new interest in robust interpretability (Alvarez-Melis and Jaakkola, 2018).

5 Discussion and Conclusions

We show that interpretability can be used to predict model behavior on unseen inputs—even when the interpreted artifacts are not causally implicated in this behavior. Our findings suggest a new way of evaluating interpretations: by their ability to predict model behavior on unseen inputs. We intuit OOD generalization rules from ID representations and predict OOD generalization across two distinct settings: Dyck-1 classification and autoregressive English question formation.

In particular, we find that some attention patterns causally support OOD generalization while others limit it. By conducting a *correlational* analysis across a model population, we discover that violation-tracking attention patterns are predictive of hierarchical generalization behavior—even when causal analysis suggests they suppress this same behavior. These desiderata can be added to the current evaluation toolkit, which often focuses on in-distribution ablation response or correlation with data properties.

Causality is often held as the gold standard for mechanistic understanding. But causal findings are often compatible with multiple analyses that only expensive, precise ablations can distinguish, and proposed interpretations frequently fail to transfer to new settings (Stander et al., 2024; Zhang and Nanda, 2024; Makelov et al., 2024). The literature on understanding attention has produced skepticism and fierce debate (Bibal et al., 2022;

Wen et al., 2023) precisely because causal intervention has been treated as the lone arbiter of faithfulness. Our stance is correspondingly instrumentalist (Duhem, 1954): hierarchical structure is part of our scientific model of network behavior, sufficient if we can infer it from its traces.

Predicting model behavior can be valuable independent of whether the underlying representational structure satisfies standard notions of mechanistic faithfulness. Even if perfect prediction remains elusive, flagging edge-case behavior could meaningfully improve evaluation and deployment. Overall, this work argues that the instrumental, *algorithmic*-level goal of predicting model behavior is a worthwhile and tractable interpretability objective.

6 Limitations

While our work provides a proof-of-concept for predictive interpretability, it relies on a controlled synthetic setting with several simplifying assumptions. Our parentheses and question-formation tasks have a clear ground truth for generalization rules, and we train hundreds of models to analyze population-level patterns—conditions that may not hold when analyzing real-world models.

Future work could investigate whether mechanistic analyses can predict generalization in more realistic settings. However, extending our approach faces challenges. First, similar analysis may be difficult in larger models because scale reduces—but does not eliminate—the impact of random variation and the isolated role of individual attention heads. Second, identifying interpretable structure is more difficult outside of synthetic settings with known OOD generalization behaviors.

Nevertheless, we believe representational geometry may provide clues about which capabilities a model has learned and how it may perform under distribution shift. Regardless of whether these geometric interpretations are considered faithful by standard metrics, predicting model behavior can be valuable independent of causal analyses. Even if perfect prediction remains elusive, flagging potential edge case behavior could meaningfully improve model evaluation and deployment decisions.

References

- Samira Abnar, Mostafa Dehghani, and Willem Zuidema. 2020. [Transferring inductive biases through knowledge distillation](#). *Preprint*, arXiv:2006.00555.
- David Alvarez-Melis and T. Jaakkola. 2018. [On the robustness of interpretability methods](#). *ArXiv*, abs/1806.08049.
- D. B. Arnold and M. R. Sleep. 1980. [Uniform random generation of balanced parenthesis strings](#). *ACM Trans. Program. Lang. Syst.*, 2(1):122–128.
- Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900.
- Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. [Data distributional properties drive emergent in-context learning in transformers](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 18878–18891. Curran Associates, Inc.
- Xizi Chen, Rui Pan, Xiaomeng Wang, Fengshi Tian, and Chi-Ying Tsui. 2023. [Late breaking results: Weight decay is all you need for neural network sparsification](#). In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–2.
- Paul M Churchland. 1981. Eliminative materialism and the propositional attitudes. *the Journal of Philosophy*, 78(2):67–90.
- Daniel C Dennett. 1989. *The intentional stance*. MIT press.
- Daniel C Dennett. 1991. Real patterns. *The journal of Philosophy*, 88(1):27–51.
- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. [Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping](#). *Preprint*, arXiv:2002.06305.
- Darshil Doshi, Aritra Das, Tianyu He, and Andrey Gromov. 2024. [To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets](#). In *International Conference on Representation Learning*, volume 2024, pages 3782–3810.
- Pierre Duhem. 1954. *The aim and structure of physical theory*. Princeton University Press.
- Javid Ebrahimi, Dhruv Gelda, and Wei Zhang. 2020. [How can self-attention networks recognize Dyck-n languages?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4301–4306, Online. Association for Computational Linguistics.
- Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. [Causal abstraction: A theoretical foundation for mechanistic interpretability](#). *Journal of Machine Learning Research*, 26:1–64.
- Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarnag, Adam Pearce, Chris Olah, and Joshua Batson. 2026. [When models manipulate manifolds: The geometry of a counting task](#). *Preprint*, arXiv:2601.04480.
- Emma Haapaniemi, Sandeep Botla, Jenna Persson, Bernhard Schmierer, and Jussi Taipale. 2018. Crispr-cas9 genome editing induces a p53-mediated dna damage response. *Nature medicine*, 24(7):927–930.
- Brian Hepburn and Hanne Andersen. 2021. Scientific Method. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2021 edition. Metaphysics Research Lab, Stanford University.
- Jing Huang, Junyi Tao, Thomas Icard, Diyi Yang, and Christopher Potts. 2025. [Internal causal mechanisms robustly predict language model out-of-distribution behaviors](#). *Preprint*, arXiv:2505.11770.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not Explanation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. 2023. Linear connectivity reveals generalization strategies. In *International Conference on Learning Representations*.
- Subhash Kantamneni, Joshua Engels, Senthoran Rajamanoharan, Max Tegmark, and Neel Nanda. 2025. [Are sparse autoencoders useful? A case study in sparse probing](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 29018–29049. PMLR.
- Andrej Karpathy. 2020. [MinGPT transformer model](#).
- Connor Kissane, robertzk, Neel Nanda, and Arthur Conmy. 2024. [SAEs are highly dataset dependent: a case study on the refusal direction](#). *Alignment Forum*.
- János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. [Atp*: An efficient and scalable method for localizing llm behaviour to components](#). *Preprint*, arXiv:2403.00745.
- Edwin B Liem, Teresa V Joiner, Kentaro Tsueda, and Daniel I Sessler. 2005. Increased sensitivity to thermal pain and reduced subcutaneous lidocaine efficacy in redheads. *Anesthesiology*, 102(3):509.

- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. 2024. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32287–32307. PMLR.
- Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. 2024. [Is this the subspace you are looking for? an interpretability illusion for subspace activation patching](#). In *International Conference on Representation Learning*, volume 2024, pages 26486–26515.
- David Marr. 2010. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.
- R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020a. [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.
- R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020b. [BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance](#). In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.
- Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. 2023. [Grokking of hierarchical structure in vanilla transformers](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada. Association for Computational Linguistics.
- Samir Okasha. 2016. *Philosophy of science: very short introduction*. Oxford University Press.
- Goncalo Paulo and Nora Belrose. 2025. [Sparse autoencoders trained on the same data learn different features](#). *arxiv*, abs/2501.16615.
- Jackson Petty, Sjoerd van Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. 2024. The impact of depth on compositional generalization in transformer language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7239–7252. Association for Computational Linguistics.
- Hilary Putnam. 1975. *Mathematics, Matter and Method: Volume 1, Philosophical Papers*, volume 1. cup Archive.
- Tian Qin, Naomi Saphra, and David Alvarez-Melis. 2025. Sometimes I am a tree: Data drives unstable hierarchical generalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11722–11740. Association for Computational Linguistics.
- Naomi Saphra and Adam Lopez. 2020. [LSTMs compose—and Learn—Bottom-up](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2797–2809, Online. Association for Computational Linguistics.
- Sofia Serrano and Noah A. Smith. 2019. [Is attention interpretable?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.
- Lewis Smith, Senthoran Rajamanoharan, Arthur Conmy, Callum McDougall, Tom Lieberum, János Kramár, Rohin Shah, and Neel Nanda. 2025. [Negative results for SAEs on downstream tasks and deprioritising SAE research \(GDM mech interp team progress update #2\)](#).
- Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. 2024. [Grokking group multiplication with cosets](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46441–46467. PMLR.
- Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. [Extracting latent steering vectors from pretrained language models](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.
- Mirac Suzgun, Sebastian Gehrmann, Yonatan Belinkov, and Stuart M. Shieber. 2019. [Memory-augmented recurrent neural networks can learn generalized Dyck languages](#). *Preprint*, arXiv:1911.03329.
- Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. 2025. [Analyzing the generalization and reliability of steering vectors](#). *Preprint*, arXiv:2407.12404.
- Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. [Scale efficiently: Insights from pre-training and fine-tuning transformers](#). In *Proceedings of the 10th International Conference on Learning Representations*.
- Hugo Tessier, Vincent Gripon, Mathieu Léonardon, Matthieu Arzel, Thomas Hannagan, and David Bertrand. 2022. [Rethinking weight decay for efficient neural network pruning](#). *Journal of Imaging*, 8(3):64.

- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. [Function vectors in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Ke Tran, Arianna Bisazza, and Christof Monz. 2018. [The importance of being recurrent for modeling hierarchical structure](#). *Preprint*, arXiv:1803.03585.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. [Investigating gender bias in language models using causal mediation analysis](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.
- K Wen, Y Li, B Liu, and A. Risteski. 2023. [Transformers are uninterpretable with myopic methods: a case study with bounded Dyck grammars](#). In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). In *The Twelfth International Conference on Learning Representations*.
- Rosie Zhao, Naomi Saphra, and Sham M. Kakade. 2024. [Distributional scaling laws for emergent capabilities](#).

A Glossary

Term	Definition and Usage
Depth	Parentheses sequence token characteristic. At index j , the token depth is $o(j) - c(j)$ where $o(j)$ and $c(j)$ are the cumulative counts of ⌈ and ⌋ up to j .
Hierarchical head	Head that tracks a particular type of hierarchical attention pattern. Specifically, in Dyck, we evaluate if the EOS attention weights persistently favor either violation or non-violation tokens (e.g., Equation 4 or 5) on $\geq 80\%$ of sequences containing both violations and non-violations. This head type encompasses violation-detecting heads and sign-matching heads, and it is typically found in 2- and 3-layer models.
EOS	The end-of-sequence token where attention patterns are inspected.
EQUAL-COUNT	Rule that a model can learn. True iff the number of open and close parentheses in a sequence are equal, regardless of their ordering (Equation 1). Marker: Predicts True for all OOD inputs (0% OOD accuracy).
FIRST-SYMBOL	Rule that a model can learn. Predicts True if first token is ⌈ , False otherwise. This rule occurs in 1-layer models with 0 weight decay. Marker: $\sim 55\%$ OOD accuracy.
Depth-tracking attention	Dyck hierarchical attention pattern which reflects the tree structure of the preceding input sequence at an EOS token by indicating which indices violate the condition in Equation 2. The pattern can favor either violation or non-violation tokens.
ID data	In-distribution training and validation data where negative examples satisfy neither the EQUAL-COUNT nor NESTED rules.
NESTED	Rule that a model can learn. Returns True iff the input is properly nested (i.e., satisfies both EQUAL-COUNT and depth non-negativity, Equation 2). Marker: Predicts False on all OOD examples (100% OOD accuracy).
Violation token	Token position that violates the NESTED condition in Equation 2, i.e., a position preceded by more close than open symbols.
Violation-detecting head	Dyck hierarchical head that preferentially attends to tokens that violate Equation 2 on $\geq 80\%$ of sequences containing both violations and non-violations. Behavior: Their presence is correlated with the hierarchical NESTED rule OOD, but ablating this head decreases models' hierarchical behavior.
OOD data	Out-of-distribution test set. For Dyck, negative examples satisfy EQUAL-COUNT but not NESTED. Higher accuracy is associated with NESTED. For question formation, OOD sentences begin with different auxiliaries for linear vs hierarchical rules. Higher accuracy is associated with hierarchical behavior.
Sign-matching head	Hierarchical head that attends to violation tokens if the last token violates Equation 2 and non-violation tokens if the last token does not violate Equation 2, on $\geq 80\%$ of sequences containing both violations and non-violations. Behavior: Their presence is correlated with the hierarchical NESTED rule OOD, but ablating this head increases models' hierarchical behavior.
Main Auxiliary-Detecting head	Question formation analog of a Dyck hierarchical head. At the question-formation readout position (quest), a head in one of the first three layers of the model places attention weight > 0.2 on the matrix-auxiliary token on $\geq 80\%$ of ID sequences. Behavior: Models with such ID heads correlate with higher OOD (hier.) accuracy.
Uniform attention ablation	Flattens attention to a uniform weight distribution. Used to test causal importance of particular attention heads and patterns.
Vestigial circuit	A sub-circuit used early in training, but not necessary for model performance at the end (e.g., FIRST-SYMBOL circuit in unregularized 1-layer models). At the end of training, such circuits may be associated with rules applied OOD but not ID.

Table 1: Glossary of key terms used in this paper.

B Attention Activation Examples

We include a figure to clarify what we mean by violation and non-violation tokens.

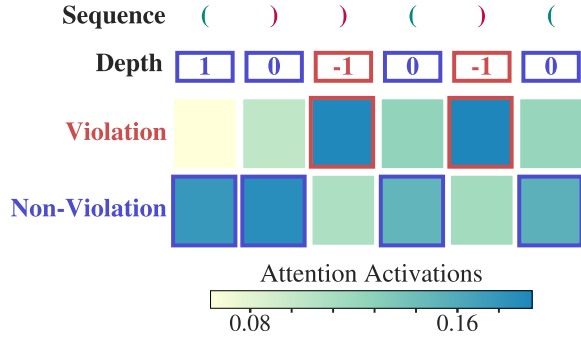


Figure 6: Examples of attention patterns favoring violation (top row) and non-violation (bottom row) tokens at the EOS position. The input sequence is shown along the top; each token’s tree depth $o(j) - c(j)$ is displayed below. Tokens with depth < 0 are violations of Equation 2.

C Reproducibility Statement

All 270 models, data, and code in the Dyck setting are freely available for use without restriction at <https://github.com/vli31/id-predict-ood>. Hyperparameter settings for each run are extensively documented in the main text and appendix of this paper. We recommend that future work experiment further with increasing scale and study the particular interactions between heads, to extend our descriptions of representation to be more granular.

We adapted the minGPT software artifact which has an MIT license (Karpathy, 2020). Packages used for evaluation include SciPy (version 1.14.0). To train models for the question-formation setting, we used the open source output of Qin et al. (2025) available at https://github.com/sunnytqin/hier_gen.

D Dyck Dataset Generation Details

We can think of the sequence length distribution as though we are generating NESTED trees with a 50% probability of recursing at each node, but discarding identical sequences. Each sequence of symbols, sampled uniformly at random, is then sorted according to which rule it follows.

D.1 Dataset parentheses sampling

We create datasets as follows:

1. Sample an even sequence length $n = 2 + 2B$ with $B \sim \text{Binomial}(19, 0.5)$, so that n ranges over even values from 2 to 40. These properties ensure that our samples are concentrated

around a reasonable center, reducing extreme sequence lengths that could occur with other distributions like the Uniform. Also note that our maximum sequence length is 40.

2. Generate a uniformly random parentheses sequence of length n with the desired attributes.
 - To generate a uniformly random sequence that is neither EQUAL-COUNT nor NESTED, we choose each character independently from the set $\{ \text{⌈}, \text{⌋} \}$. If the resulting sequence satisfies EQUAL-COUNT, we discard it.
 - To generate a uniformly random sequence that is EQUAL-COUNT but not NESTED, we randomly permute $n/2$ ⌈ parentheses and $n/2$ ⌋ parentheses. If the resulting sequence is NESTED, we discard it and generate a new one.
 - To generate a uniformly random sequence that is NESTED, we use the algorithm of Arnold and Sleep (1980).

3. If the sequence generated does *not* already appear in the dataset, add it to the dataset.

Thus, each length- n sequence s with the desired attributes is equally likely to be chosen, and it is chosen at most once. Since we discard repeats, the empirical distribution of sequence lengths is skewed towards longer sequences, as short sequences are likely to be repeated.

We tokenize the ⌈ and ⌋ characters in addition to start, end, and padding tokens (BOS, EOS and PAD, respectively, with PAD appended to the end of the sequence) to ensure each sequence for classification has length 42 (including start and end tokens). In other terms, we create a sequence of form:

$$s_0 s_1 \dots s_n \dots s_{41},$$

where s_0 is the beginning-of-sequence token BOS, s_1 through s_n make up the n -length parentheses sequence s , s_{n+1} is the end-of-sequence EOS token, and s_{n+2} through s_{41} are PAD tokens.

Our ID datapoints are randomly split into training and validation datasets. Each ID set contains the same number of True examples (following both EQUAL-COUNT and NESTED) and False examples (following neither EQUAL-COUNT nor NESTED). Our OOD test set consists of parentheses sequences which follow EQUAL-COUNT but

not NESTED, i.e., sequences with the same number of open and close parentheses, but in which the parentheses are not properly nested (e.g., `)))((`). See Table 2 for examples.

Empirically, we find some models classify sequences by a FIRST-SYMBOL heuristic OOD. These models check whether $s_1 = \text{[}$ and label an OOD sequence as True if the first character is `[` and False if it is `]`. (In-distribution, no models follow FIRST-SYMBOL, which would fail to achieve full accuracy ID.)

D.2 Random data order implementation

Our train set contains 200000 distinct datapoints. During training, we repeat this set five times, so all models were exposed to 1 million total parentheses sequences (including 5 repeats of each) across the course of training. The random seed used for data ordering, or “shuffle seed,” determines the order of data within each block of 200000 training examples, so during training, a single model is exposed to the same examples in five different orderings. Models with the same shuffle seed hyperparameter encounter the 1 million total training datapoints in exactly the same order (data within each block is shuffled in a consistent way).

E Factors in rule selection

For consistency, we define OOD accuracy with respect to the NESTED rule. Thus a model achieving 100% OOD accuracy classifies each NON-NESTED, EQUAL-COUNT sequence in our OOD test set as False. Correspondingly, models with 0% OOD accuracy learn EQUAL-COUNT, classifying every OOD example as True (Table 2).

E.1 Architecture

By comparing Transformers to LSTMs, we confirm existing findings (Abnar et al., 2020; McCoy et al., 2020a; Tran et al., 2018; Saphra and Lopez, 2020) that LSTMs are intrinsically hierarchical while Transformers are not. The inductive bias of the LSTM architecture places every trained model at more than 60% accuracy on the OOD generalization set, indicating that none of these models learn EQUAL-COUNT and all are closer to the hierarchical NESTED rule. In contrast, Transformer models exhibit an OOD accuracy distribution with two peaks: one near 0% (indicating perfect application of the EQUAL-COUNT rule) and a smaller one at 90% (indicating a tendency towards NESTED). Overall, 24.4% of Transformers learn

EQUAL-COUNT perfectly, achieving zero OOD accuracy at the last step in training (Figure 7).

These architectures ranged in parameter count with 1-, 2-, and 3-layer Transformers having 53K, 103K, and 153K parameters, respectively, and LSTMs around 250K parameters. Overall, these models took around 25 H100 GPU hours to train.

E.2 Width

Using the non-parametric Mann–Whitney U test to detect differences between distributions, we find that the number of Transformer heads has no significant effect on the distribution of OOD accuracies for any depth of model (Figure 8). This result, which we show holds over randomness in model initialization and data exposure order, adds to a growing body of evidence across settings that changing Transformer width has little effect on model expressivity and OOD generalization (Petty et al., 2024; Tay et al., 2022).

E.3 Depth

In this setting, depth—unlike width—is a significant factor in rule selection, determining the peaks of the OOD behavior distribution. For instance, among the 66 out of 270 Transformers (24%) which achieve perfect EQUAL-COUNT behavior with 0% OOD accuracy (Figure 7), there are 61 1-layer models, no 2-layer models, and 5 3-layer models. (Note our overall model population is evenly split into 90 each of 1-, 2-, and 3-layer models.)

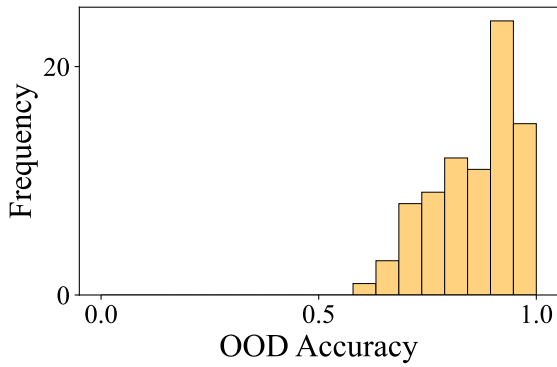
We find that 1-layer models fall into a bimodal distribution centering on two rules: EQUAL-COUNT (characterized by 0% OOD accuracy) and FIRST-SYMBOL (which gives $\sim 55\%$ OOD accuracy). The 18 models that generalize according to FIRST-SYMBOL all exhibit nearly identical judgments on specific examples, following a rule associated with returning a True label if the input begins with `[`. Because we only consider models with at least 99% validation accuracy, the models in question must also have other subroutines that are successfully applied ID but dominated by FIRST-SYMBOL OOD.

We group models via t-SNE (perplexity 12) based on their judgments on the OOD test set (Figures 2 and 11). FIRST-SYMBOL forms an outlier cluster in model judgments of 1-layer models, which otherwise learn EQUAL-COUNT.

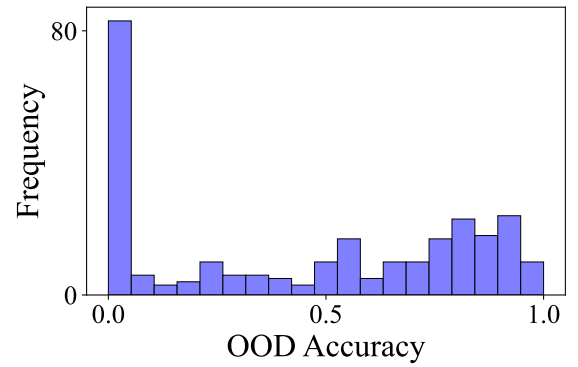
By contrast, only 2% of 2-layer models are EQUAL-COUNT-leaning (determined by $< 20\%$ accuracy OOD) and none learn FIRST-SYMBOL.

Table 2: The EQUAL-COUNT and NESTED rules applied to example parentheses sequences in our ID and OOD test sets. Also, classifications of the same parentheses sequences according to the OOD FIRST-SYMBOL heuristic. Notice that this rule does not achieve perfect accuracy ID.

Dataset	EQUAL-COUNT	NESTED	Possible s_1	Example	FIRST-SYMBOL
ID	True	True	(	(())	True
ID	False	False	()	((()()())	True False
OOD	True	False	()	((())())	True False
—	False	True	—	Does Not Exist	—



(a) OOD-test accuracy for LSTMs.



(b) OOD-test accuracy for Transformers.

Figure 7: Last OOD test accuracy for LSTM and Transformer models that achieve 99%+ ID accuracy. When LSTMs converge to near-perfect ID accuracy, they consistently also apply the NESTED rule to OOD data. Transformers, meanwhile, apply a variety of rules and exemplar-based behavior OOD.

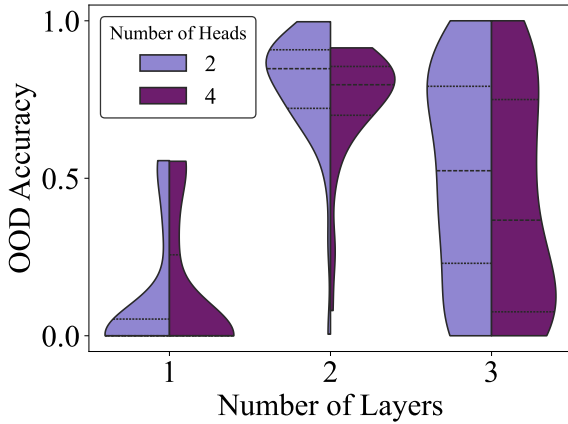


Figure 8: Unlike weight decay and depth (Figure 9), width is not a substantial factor in final OOD rule selection in this setting. The Mann–Whitney U test finds no statistically significant differences in the distribution of last OOD accuracy over width (all $p > 0.05$).

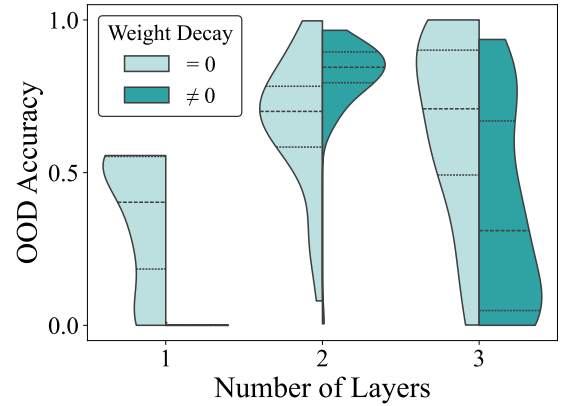


Figure 9: Weight decay and depth substantially impact OOD rule choice. The Mann–Whitney U test finds significant differences in OOD accuracy distribution between the absence and presence of weight decay across depths ($p \ll 0.01$ across layers).

The mode of this model distribution is instead at 90% accuracy, firmly suggesting that most 2-

layer models have approximately learned NESTED. Among 3-layer models, behavior varies enormously, with the distributional mode defined by

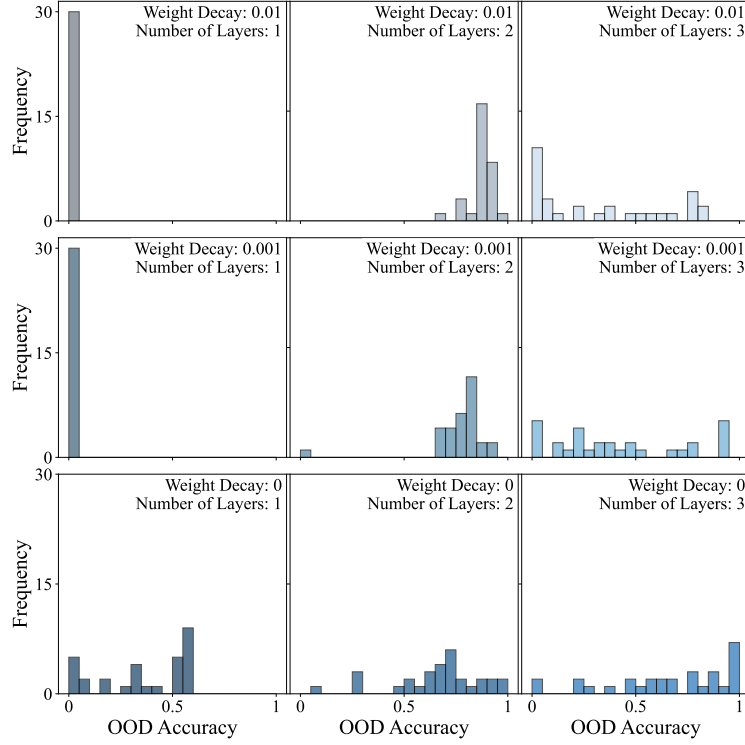


Figure 10: Final accuracy on OOD-test for Transformers of varying depths and weight decays. 1-layer Transformers learn EQUAL-COUNT or FIRST-SYMBOL. Deeper Transformers can learn EQUAL-COUNT or approximate NESTED, with 2-layer Transformers most likely to learn NESTED and 3-layer Transformers instead exhibiting more complex OOD generalization behavior.

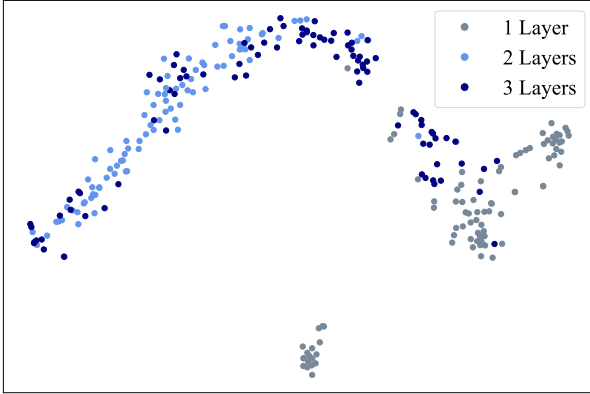


Figure 11: t-SNE of models’ final OOD classifications colored by model depth.

the 20 models with < 0.1 OOD accuracy which learn EQUAL-COUNT. Across all 270 training runs, 10.4% of models achieve at least 90% accuracy, including 15 2-layer and 13 3-layer models.

E.4 Regularization

Weight decay has a significant impact on the distribution of rules models learn. Without any weight decay, models that generalize ID can converge on a variety of OOD generalization rules with wide dis-

tributions. With weight decay, particularly among smaller models, models have more similar OOD behaviors. For example, while 2-layer Transformers show a consistent tendency to prefer NESTED, they achieve higher OOD accuracy more reliably when weight decay is applied (Figure 10, Figure 9).

Among 1-layer models, training with weight decay always results in convergence to the EQUAL-COUNT rule. Among 1-layer models trained without weight decay, 60% (18 of 30) converge to FIRST-SYMBOL with $\sim 55\%$ accuracy on the OOD test set (Figure 9). The presence of all FIRST-SYMBOL-learning models in 1-layer models with weight decay 0 indicates regularization can help prune away vestigial model features unnecessary for ID generalization. The presence of circuits supporting FIRST-SYMBOL may not impact ID performance, but in the absence of regularization, such features significantly decrease OOD performance.

E.5 Randomization in Weight Initialization and Data Order

Existing work (Qin et al., 2025; Chan et al., 2022; Zhao et al., 2024; Juneja et al., 2023) has investi-

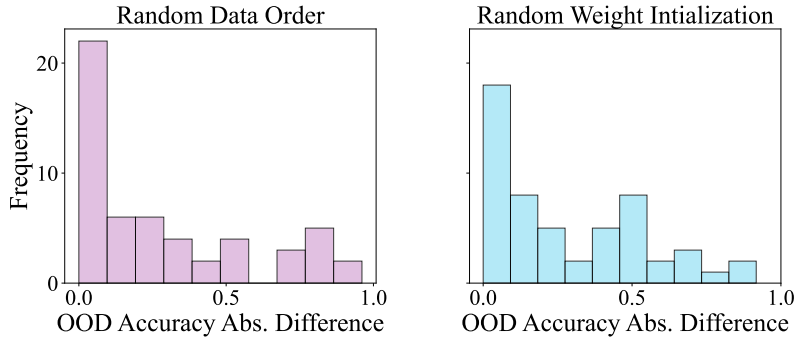


Figure 12: Data order (left) and weight initialization (right) are equally influential random factors in OOD behavior, as shown by the distribution of the ranges of final OOD accuracy between models trained with differing data ordering and weight initialization.

gated the impact of random weight initialization and data order on model performance. Dodge et al. (2020) varied these factors in BERT fine-tuning, finding that modifying either factor had significant, comparable impacts on model performance. We investigate the impact of the two sources of random variation that account for differing model behaviors, and find that while they do not affect ID accuracy, both dataset ordering and model initialization affect generalization behavior.

Since we train models across three data shuffle seeds and five model random seeds, for a fair comparison between ranges of the impact of these factors on final OOD accuracy, we randomly select three random seeds to plot and compare the ranges of performance across shared hyperparameter conditions.

In Figure 12, we show that in a plurality of models trained, OOD performance was impacted by at least 10%, with the maximum difference due to either one of the two random factors reaching an above 90% difference in OOD behavior. The similar distribution in the impact of random initialization and data order is aligned with previous work and indicates both factors are important to determining model OOD performance and should be accounted for in building robust ML systems.

F Training dynamics of different rules

Some OOD generalization rules can converge simultaneously with ID performance, whereas others take longer to learn after the model successfully learns ID. In our setting, models can acquire and stabilize into the EQUAL-COUNT rule, but generally take longer to converge to the NESTED rule.

Although overall, models tend to classify OOD sequences as False at the outset of training—likely

because the False training examples, being sampled uniformly at random, are far more diverse—they rarely stabilize immediately at high rates of False (i.e., equivalent to a NESTED rule). The models that stabilize at a NESTED rule, as seen in Figure 13, often stabilize long after ID convergence. In other words, we see an example of *structural grokking* (Murty et al., 2023). These results support the idea that NESTED is a more difficult rule to fully learn. Indeed, only three Transformer models adhere completely to the NESTED rule by classifying all OOD examples as False. The training dynamics of different rules broken down in detail by model depth are also shown in Figure 14. Notably, deeper models have higher final OOD accuracy variance and also display greater variance during training, possibly because of their higher expressivity.

F.1 Evaluating the Heuristic

The FIRST-SYMBOL heuristic, in contrast to the NESTED and EQUAL-COUNT rules, is not used by any models ID, where it would produce a low accuracy. However, it is used by some 1-layer models OOD, and it produces an OOD accuracy of ≈ 0.55 , reflecting the fact that around 55% of our OOD test examples begin with close parentheses.

As discussed in Section 3.1.1, a majority of 1-layer models appear to pass through a FIRST-SYMBOL heuristic phase, although this heuristic does not persist until the end of training among models trained with weight decay. In this case, we say the models “appear” to pass through this phase, rather than asserting that they do, because we are only able to verify their behavior on individual datapoints at our five saved model checkpoints. However, at those model checkpoints, we are able to

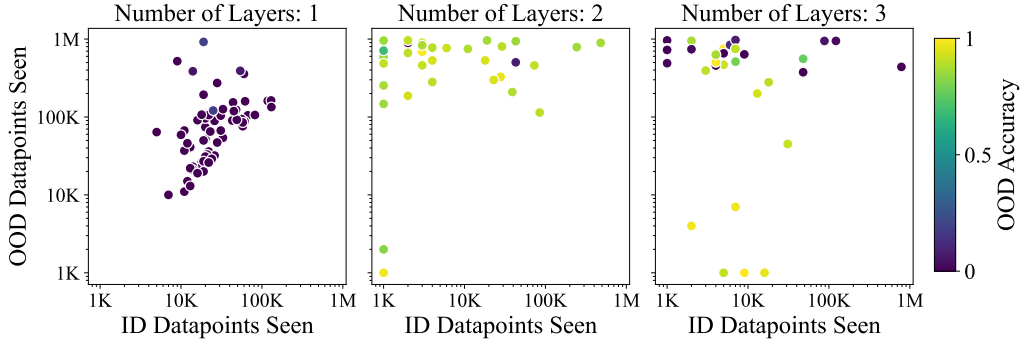


Figure 13: Illustration of generalization rules across training. ID convergence occurs when models achieve $\geq 99\%$ ID accuracy for $> 99\%$ of remaining datapoints seen—all Transformers achieve this metric after 900K datapoints. We define OOD convergence to either EQUAL-COUNT or NESTED as Transformers achieving ≤ 0.2 or ≥ 0.8 accuracy for $> 99\%$ of the rest of the model run, respectively, after seeing at most 975K datapoints—53% of Transformers achieve this metric. Using these metrics, this plot shows the number of datapoints seen before OOD and ID convergence, excluding models that do not converge OOD.

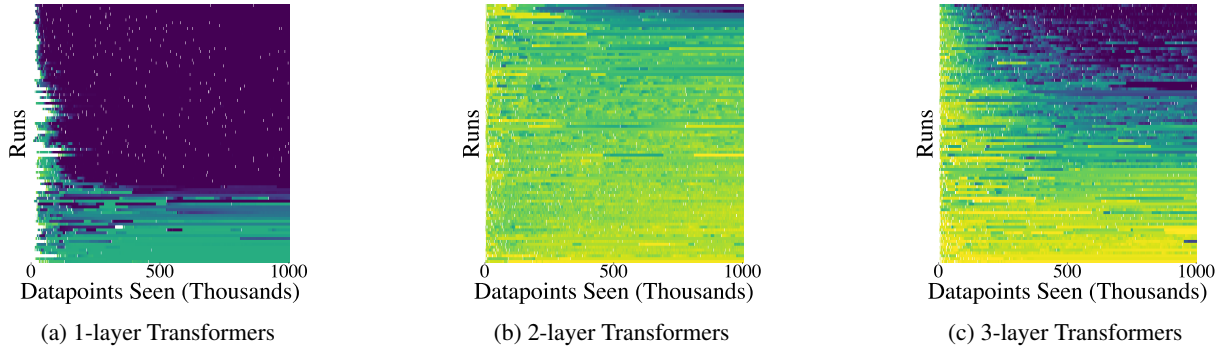


Figure 14: Heatmaps showing model training dynamics broken down by depth, where purple and yellow indicate model adherence to the EQUAL-COUNT and NESTED rule, respectively (using the same scale as Figures 2, 3, and 13). Colored cells indicate the OOD accuracy of a particular run when ID accuracy is at least 0.99.

confirm that 1-layer models whose OOD accuracy is between 0.54 and 0.56 indeed almost always make OOD judgments based on first symbol. We therefore posit that throughout training, instances of OOD accuracy in this range likely reflect the FIRST-SYMBOL heuristic, though we cannot rule out that some such instances reflect some other heuristic which coincidentally produces the same OOD accuracy. Notably, 2- and 3-layer models also reach accuracies in this range while training, but, checking them at saved checkpoints, we find they do not learn FIRST-SYMBOL.

G Attention Heads Classified by OOD Behavior

We break down the presence of types of head across 2- and 3-layer models (Figure 15).

In Section 4.2, we showed that classifying hierarchical heads according to their behavior on the ID validation set is predictive of generalizing according to the NESTED rule. It is also possible to

conduct the same analysis using behavior on the OOD test set. We find that the presence of OOD hierarchical heads is similarly predictive of generalizing according to the NESTED rule (Appx. H). All types of hierarchical head appear to correlate similarly strongly with NESTED generalization.

H ID Hierarchical Heads as a Classifier

We can recast our predictive interpretability claim as a concrete classifier. For Dyck, “model has at least one ID hierarchical head” is a binary prediction of “model implements NESTED on OOD.” Across the 2- and 3-layer Dyck population ($n = 180$), thresholding the true label at OOD accuracy > 0.5 yields **precision** = 0.96, **recall** = 0.68 (F1 = 0.79); the OOD-classified head detector reaches precision 0.94, recall 0.53 (F1 = 0.68). For the question-formation setting ($n = 79$), the Main Auxiliary-Detecting head detector reaches precision 0.76, recall 0.85 (F1 = 0.80) at the same OOD-acc cutoff. The two settings sit at oppo-

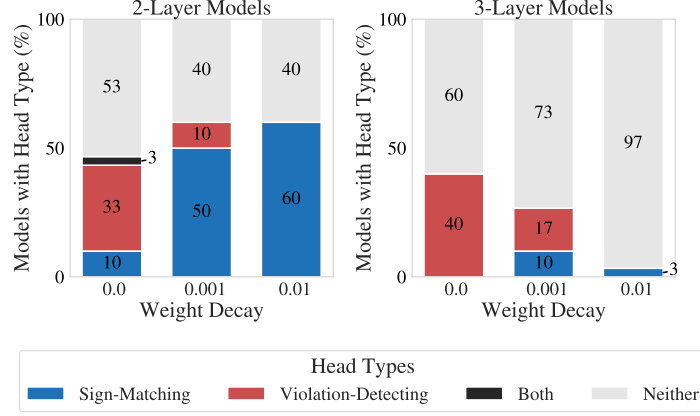


Figure 15: Percentage of 2- and 3-layer models containing each head type, by weight decay. Head types are classified according to their OOD behavior.

site ends of the precision/recall trade-off: in Dyck the head detector is *high-precision, lower-recall*—models flagged as carrying a hierarchical head almost always generalize hierarchically OOD, while some hierarchical-OD models implement the rule without a detectable head (consistent with Wen et al. (2023)’s finding that hierarchical behavior can also arise through nearly uniform attention). In question formation the Main Auxiliary-Detecting head is instead *high-recall*: most hierarchically-generalizing models carry it, at moderate precision. In both settings, carrying the head is a reliable positive signal for hierarchical OOD behavior.

H.1 Held-out evaluation

In the previous section, we fit precision and recall on the population. We also evaluate, on unseen models, predictors of whether a model lies in the NESTED region of Fig. 2a. This predictor sees only ID information. We report AUROC, the probability that a randomly chosen NESTED model is ranked above a randomly chosen non-NESTED model.

Predictors. We use three predictors: (1) The ID hierarchical-head rule in Section 4.2. We associated NESTED behavior with heads that track violations on at least 80% of relevant ID sequences. (2) We also train a random forest based on EOS-attention statistics on 1K ID sequences (not relying on considering Equation 2—in particular, center of mass, mass on the last tenth of the sequence, largest weight, weight on EOS, and mean self-attention over all positions summarized by max, mean and min over heads—15 numbers overall). (3) We further compare to a baseline of hyperparameters alone: a logistic regression on depth, width and

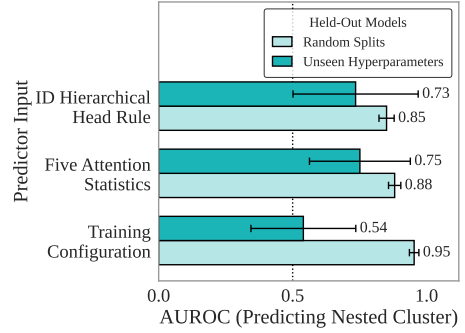


Figure 16: **ID internals predict which unseen models follow NESTED; hyperparameters do not if their setting is unseen.** Each bar is the AUROC for predicting that a held-out Dyck model lies in the NESTED region of Fig. 2a from ID information only: the ID hierarchical-head rule, five attention statistics, or the hyperparameters alone. Light bars: fit on 150 random models, tested on the other 120 (50 splits). Dark bars: all models with one (depth, weight decay) setting held out (six scorable settings). Whiskers: ± 1 SD across splits; dotted: chance.

weight decay.

Evaluation. We evaluate on random splits: 150 training models and 120 held-out models, 50 draws. Unseen hyperparameters are 30 models sharing one (depth, weight decay) setting fit on others; six of the nine settings contain both NESTED and non-NESTED models and can be scored. Figure 16 reports the mean and standard deviation over splits.

Results. In random splits both the ID hierarchical-head rule and the five attention statistics-based predictors reach AUROC 0.85 and 0.88, respectively. The hyperparameters alone do better (0.95), as weight decay and depth strongly

shape rule choice (Appx. E). If we hold out a hyperparameter setting, hyperparameters fall to chance (0.54) while the ID head rule and attention statistics retain predictive power at 0.73 and 0.75 AUROC, respectively. The same two predictors transfer to the question-formation population ($n = 79$; target: the HIERARCHICAL region of Fig. 2b; all models share one configuration, so only random splits apply; 5-fold cross-validation): the Main Auxiliary-Detecting head reaches AUROC 0.76 and the same five statistics, computed at the position preceding the question, 0.72. Hierarchical heads are a readable and interpretable instance of ID internals as a predictive signal of OOD generalization.

I Breakdown of Hierarchical Heads by Layer and Weight Decay

Models with ID hierarchical heads consistently have higher OOD accuracy across both 2- and 3-layer models, indicating that model internals can provide additional predictive power to hyperparameters alone (Figure 17). Particularly for 3-layer models, which have the greatest diversity in OOD performance, the presence of ID hierarchical heads in a model provides additional insight into predicting hierarchical NESTED-like generalization.

Model internals continue to improve predictive power even when fixing both depth and weight decay simultaneously. Some hyperparameter combinations lead to one rule or the other relatively consistently: for example, “1-layer, weight decay > 0 ” is completely predictive of EQUAL-COUNT (see Figure 10, Appendix E.3). However, in hyperparameter settings with diverse OOD behavior, presence of hierarchical heads is predictive of NESTED generalization behavior.

In Figure 18, consider any multilayer setting producing a diverse range of OOD accuracy values, including values both above and below 50% (in other words, any setting except the “2-layer, 0.01 weight decay” setting, where all models learn NESTED). In every such setting, models without hierarchical heads have a median OOD accuracy that falls at or below the bottom quartile of the score for models with such heads. For some settings, the distributions barely overlap at all.

Thus, even if we consider the effect of hyperparameter settings, the presence of a hierarchical (i.e., depth-tracking) head is highly informative. Of course, some of these settings might lead to more hierarchical runs *because* they enable the learning of hierarchical mechanisms, meaning that the ef-

fect of hierarchical heads is far greater than any one setting would suggest.

J Effects of Causal Intervention on Attention

Our causal experiments involve uniform ablation of the attention distribution in all attention heads (Section 4.3). We ablate all attention in order to uniformly and symmetrically intervene on all models. This is in contrast to ablating exclusively hierarchical heads, which would require us to compare models on an “unequal footing,” in the sense that some models would have 0 heads ablated, some models 1 head ablated, and some models 2 or more. Ablating all attention allows us to definitively eliminate all influence from hierarchical heads without introducing asymmetric interventions.

We also examined the effects of ablating individual heads one at a time. We found effects that were generally very similar (if weaker), in comparison with full attention ablation; in particular, in models with 2 or more hierarchical heads, full attention ablation tended to affect OOD accuracy more strongly than one-at-a-time attention ablation. Ultimately, one-head-at-a-time ablation demonstrates the same trend as full attention ablation: ablating violation-detecting heads reduces NESTED behavior, while ablating sign-matching heads increases this behavior. See Figure 19. Mean ablation—replacing every head’s attention pattern with its dataset-mean pattern over the OOD test set, either all heads at once or one head at a time—gives the same result as its uniform counterpart. In particular, across the 180 2- and 3-layer models, models with OOD sign-matching heads become more hierarchical when those heads are ablated, while models with violation-detecting heads become more linear.

K Change in OOD and ID Accuracy After Ablating Heads

Ablating hierarchical heads has little to no effect on ID accuracy (Figures 20 for ID hierarchical heads and 21 for OOD hierarchical heads, left panels). The maximum effect of ablating an ID or OOD head type on the ID data is 77/1000, both for violation-detecting heads, but the median impact is around 10/1000 for all classified head types. Across ID head types, ablation tends to decrease OOD accuracy, but for OOD head types, as seen in Figure 5, ablating violation-detecting heads decreases while ablating sign-matching heads increases OOD accuracy (Figures 20 and 21, right



Figure 17: Last OOD test accuracy of 2- and 3-layer models with and without ID hierarchical heads.

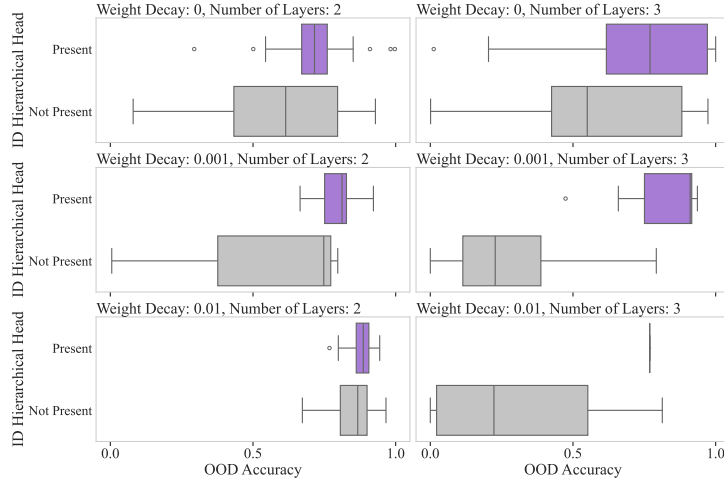


Figure 18: OOD accuracy of 2- and 3-layer models with and without ID hierarchical heads, by number of layers and weight decay value. Over the four populations, (1) 2- and 3-layer models with non-zero weight decay, (2) all 2-layer models, (3) all 3-layer models, and (4) 3-layer models with 0.001 WD, a Mann–Whitney U test shows a significant difference in OOD accuracy distributions.

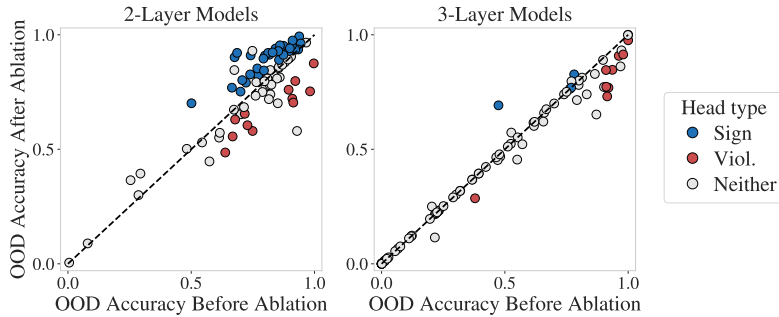


Figure 19: OOD accuracy before and after applying uniform attention ablation to one head in each model. For each model, we plot the head whose ablation affects OOD accuracy most (in absolute value), colored by whether it is a sign-matching or violation-detecting head (or neither). In comparison with full attention ablation (Figure 5), effects on OOD accuracy are smaller. However, our analysis of differing effects of ablating different types of hierarchical heads remains the same as in Section 4.3.2.

panels).

L Question Formation Setting

L.1 Task and Data

Our autoregressive setting is question formation in English (McCoy et al., 2020a; Qin et al., 2025): an autoregressive model reads a declarative sen-

tence followed by the special token marker `quest` and produces the corresponding yes/no question. The correct transformation requires identifying the *matrix* (root) auxiliary and moving it to the front:

My dog does like the ravens that do dance. quest Does my dog like the



Figure 20: For heads classified by type based on their ID behavior, accuracy ID and OOD after ablation subtracted by baseline. Ablation has little impact on ID accuracy, and tends to decrease OOD accuracy across ID sign-matching and violation-detecting heads. The number of models in each category is included in parentheses after the label.



Figure 21: For heads classified by type based on their OOD behavior, accuracy ID and OOD after ablation subtracted by baseline. Ablation has little impact on ID accuracy, but ablating violation-detecting OOD heads decreases and ablating sign-matching OOD heads increases OOD accuracy. The number of models in each category is included in parentheses after the label.

ravens that do dance?

Two rules yield the same output on training and ID validation:

- **Linear rule:** move the *first* auxiliary in the sentence to the front.
- **Hierarchical rule:** move the *matrix* (root-clause) auxiliary to the front.

ID examples are constructed so that the first auxiliary *is* the matrix auxiliary, so both rules predict the same output token; OOD examples place an auxiliary inside a relative clause that precedes the matrix auxiliary, making the two rules diverge. Table 3 contrasts the two settings of this paper.

OOD accuracy measures hierarchical generalization: we report it as $P(\text{hier})$, the mean probability the model assigns to the matrix (hierarchical) auxiliary at the fronting position. 100% means the model consistently moves the matrix auxiliary; 0% means it consistently moves the first auxiliary.

L.2 Models and Population

We analyze 79 distinct autoregressive Transformer language models in this population, following the training procedure of Qin et al. (2025). Each is a 6-layer decoder-only (causal) Transformer with 8 attention heads, model dimension 512, feed-forward dimension 2048, dropout 0.1, and tied input/output embeddings (vocabulary size 71, $\approx 19\text{M}$ parameters). The model reads the declarative followed by the *quest* marker with causal self-attention and

predicts the fronted auxiliary as the next token. All models are trained with Adam: learning rate 10^{-4} , 10,000 warm-up steps, a two-phase learning-rate schedule, no weight decay or label smoothing, gradient-norm clipping at 1.0, and batch size 8. All are evaluated at the 300K-step checkpoint and reach perfect ID validation accuracy. Because the models share architecture and objective and differ essentially only in random seed (weight initialization and data ordering), we can compare their OOD behavior directly. The heads we analyze are causal self-attention heads; “first three layers” refers to the first three of the six layers.

L.3 Main Auxiliary-Detecting ID Head

Pattern. A **Main Auxiliary-Detecting head** is an attention head in the *first three model layers* whose attention from the *quest* token—the final input token, at which the model begins generating the question—lands on the *matrix* (root-clause) auxiliary, i.e. the auxiliary that the hierarchical rule fronts. For an ID sentence “My unicorn does move the dogs that do wait.” the matrix auxiliary is *does* (not the embedded *do*); for the OOD sentence “My unicorn who doesn’t sing does move.” it is still *does* (not the embedded *doesn’t*).

Detection details. For each (model, layer, head) with the layer among the first three, we mark an ID validation example as *detected* if the attention weight $a(\text{quest} \rightarrow \text{matrix_aux}) > 0.2$; a head

		Dyck-1 (classification)	Question formation (autoregressive)
ID	Input		My dog does like the ravens that do dance.
	Output	✓ TRUE ✗ FALSE	Does my dog like the ravens that do dance?
OOD	Input		My dog who doesn't sing does dance.
	EQUAL-COUNT / linear	✓ TRUE	Doesn't my dog who sing does dance?
	NESTED / hierarchical	✗ FALSE	Does my dog who doesn't sing dance?

Table 3: Two diagnostics for *which* rule a model has actually learned. ID examples are consistent with both candidate rules; OOD examples force them apart. In Dyck-1, the linear (EQUAL-COUNT) and hierarchical (NESTED) rules disagree about whether a non-nested-but-equal-count sequence is True. In question formation, they disagree about *which* auxiliary moves: in this example, the embedded **doesn't** (linear) or the matrix **does** (hierarchical).

qualifies as Main Auxiliary-Detecting if it is detected on at least 80% of the ID validation set. A model carries the head type if at least one head in its first three layers qualifies. The two thresholds are robust: sweeping the per-token attention cutoff (from 0.025 to 0.40) and the persistence threshold (from 0.5 to 1.0), the gap in mean OOD hierarchical accuracy between models that carry the head and models that do not stays positive and generally statistically significant. At the reported cutoff in Figure 4 (> 0.2 on $\geq 80\%$ of ID examples), 59 of 79 models (75%) carry at least one Main Auxiliary-Detecting head (a one-sided Mann–Whitney U test rejects equality at $p = 6.9 \times 10^{-3}$).

L.4 Ablation Does Not Surface a Dyck-Style Decoupling

Ablation	mean Δ OOD	median Δ OOD	95% CI	Δ ID	p_{wic}
Mean	+0.002	+0.000	[−0.004, +0.009]	0.000	0.74
Uniform attn.	−0.007	−0.000	[−0.020, +0.004]	0.000	0.23

Table 4: Change in OOD hierarchical accuracy when ablating Main Auxiliary-Detecting heads (all target heads together), for the $n = 59$ models with at least one. Two ablation types (mean, uniform-attention) and one-sided Wilcoxon test. Both effects are within a percentage point of zero, neither reaches $p < 0.05$, and ID validation accuracy is unchanged. One-at-a-time ablation gives the same result.

We ablate Main Auxiliary-Detecting heads in every model that carries at least one ($n = 59$) and measure the change in OOD hierarchical accuracy. To check the result is not an artifact of a specific ablation type, we apply the same ablations we use in the Dyck setting. We consider two ablation types per model, each applied both to all target heads together and to each target head one at a time:

- **Mean ablation:** replace the head’s attention pattern with its dataset-mean pattern, computed on

the evaluation set.

- **Uniform-attention ablation:** replace the head’s attention weights with uniform attention over allowed (causally unmasked) positions.

Table 4 reports the per-type distribution of Δ (OOD hier. acc.) across the 59 ablated models. Both ablation types yield a near-zero mean effect, whether the target heads are ablated together or one at a time. Crucially, ID validation accuracy is unchanged under every intervention—all 59 models retain perfect ID accuracy—so the heads are not load-bearing for the in-distribution task either. The qualitative finding is robust: the Main Auxiliary-Detecting head correlates with hierarchical OOD behavior at ID but does not load-bear under these interventions. The surprising correlational/causal *decoupling* we report in Dyck-1 (where sign-matching attention predicts hierarchy yet ablating it *increases* hierarchy) is not a general result but a proof of existence.

M AI Statement

AI assistance was used for some analysis, wording, and visualizations. All AI generated code was manually verified.