Toward a Theory of Generalizability in LLM Mechanistic Interpretability Research

Sean Trott
Department of Cognitive Science
University of California, San Diego
La Jolla, CA 92093, USA
{sttrott}@ucsd.edu

Abstract

Research on Large Language Models (LLMs) increasingly focuses on identifying mechanistic explanations for their behaviors, yet the field lacks clear principles for determining when (and how) findings from one model instance generalize to another. This paper addresses a fundamental epistemological challenge: given a mechanistic claim about a particular model, what justifies extrapolating this finding to other LLMs—and along which dimensions might such generalizations hold? I propose five potential *axes of correspondence* along which mechanistic claims might generalize, including: functional (whether they satisfy the same functional criteria), developmental (whether they develop at similar points during pretraining), positional (whether they occupy similar absolute or relative positions), relational (whether they interact with other model components in similar ways), and configurational (whether they correspond to particular regions or structures in weight-space). To empirically validate this framework, I analyze "1-back attention heads" (components attending to previous tokens) across pretraining in random seeds of the Pythia models (14M, 70M, 160M, 410M). The results reveal striking consistency in the *developmental trajectories* of 1-back attention across models, while positional consistency is more limited. Moreover, seeds of larger models systematically show earlier onsets, steeper slopes, and higher peaks of 1-back attention. I also address possible objections to the arguments and proposals outlined here. Finally, I conclude by arguing that progress on the generalizability of mechanistic interpretability research will consist in mapping constitutive design properties of LLMs to their emergent behaviors and mechanisms.

1 Introduction

The field of *mechanistic interpretability* aims to uncover the internal structures (e.g., circuits or representations) that give rise to observable behavior in Large Language Models (LLMs) and other neural network-based systems [Olah et al., 2020, Merullo et al., 2023, 2024]. This research has the potential to deliver novel insights about the behavior of LLMs and even help build safer, more aligned models. Yet the scientific study of LLMs has yet to establish firm *epistemological foundations*: although connectionist models have of course been studied for decades [Elman, 1990, McClelland and Rumelhart, 1981], mechanistic interpretability of LLMs is still arguably in a "pre-paradigmatic" stage [Olah et al., 2020, Gurnee et al., 2023, Olah, 2023] and requires further refinement of what constitutes an explanation [Ayonrinde and Jaburi, 2025a,b]. The field thus faces a number of challenges relating to how knowledge is produced and evaluated. Some of these challenges have been discussed in recent literature, e.g., accurately benchmarking model "capabilities" [Raji et al., 2021, Ivanova, 2023, Saxon et al., 2024], but others have received only cursory treatments.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

In this paper, I focus on the question of **generalizability**, specifically of *mechanistic claims* about LLMs. For instance, given a particular claim about the circuits present in a particular model instance, **which aspects of this claim might generalize across model instances and which do not—and what principles can we use to guide those scientific generalizations**? I first argue that the field currently lacks a coherent theory of generalizability, and point to several potential features that might help predict whether two model instances share the same mechanisms (Section 2). This raises the question of what it means for two circuits to be the "same" in the first place (Section 3). Building on recent research, I propose several potential *axes of correspondence* along which mechanistic claims could plausibly generalize (Section 3.1), and also enumerate *functional criteria* that help identify which kinds of mechanisms we might expect to be robust across models (Section 3.2). I then validate the utility of the proposed theoretical framework in an empirical study focusing on the *positional* and *developmental* properties of "1-back attention heads" across random seeds of models in the Pythia suite (14M, 70M, 160M, 410M) [Biderman et al., 2023, van der Wal et al., 2025] (Section 4). I find striking *inter-seed* developmental consistencies within each model; developmental milestones are also highly correlated *across* models, albeit with some subtle differences in timing. Finally, I consider and respond to possible objections (Section 5).

2 Samples, populations, and the generalizability problem

Many scientific disciplines aim to draw general conclusions about the target of inquiry. This approach, sometimes called "nomothetic", contrasts with a more descriptive approach aimed at characterizing individual cases ("idiographic") [Beck, 1953]. In certain fields (e.g., Cognitive Science), this target of inquiry (e.g., human cognition) is too large or abstract to be observed in its entirety: consequently, researchers rely on *samples* to make inferences about the underlying *population of interest*. When drawing generalizations is the goal of scientific inquiry, this *unbiased* sample should be representative of the population of interest.[1]

Research on mechanistic interpretability is arguably nomothetic in nature [Li et al., 2015, Olah et al., 2020, Olah, 2023], i.e., the aim is to produce generalizable or "universalizable" claims about model behaviors or mechanisms. Of course, in practice, interpretability research is not (and cannot be) conducted on the entire "population" of possible LLMs; rather, it is conducted on specific **model instances**. Here, we define a "model instance" $m_{A,\theta,D}$ as a system with some particular architecture A, initialized with some particular parameter set θ, and trained on some particular dataset D.[2] While some recent work has begun to explore the issue of typologizing model instances [Klabunde et al., 2025, Yax et al., 2024], clear principles remain elusive: what would constitute a "representative sample" of LLMs? This makes it difficult for researchers to specify the "population of interest" in any particular study, or to explain the rationale behind which model instances were sampled.

Thus, given a *mechanistic claim* obtained by studying a particular model instance (or instances), what (if anything) have we learned about model instances not in the sample? For instance, if researchers identify a putative "circuit" in GPT-2, what other language models are likely to possess that same circuit—and on what basis could we justify this extrapolation? In principle, the answer to these questions should be informed by the factors known to influence the behaviors and mechanisms of individual model instances, such as: **architectural properties** like model size [Kaplan et al., 2020, Biderman et al., 2023, Riviere et al., 2025, Schrimpf et al., 2021] or depth [Mueller and Linzen, 2023, Petty et al., 2023]; the amount and variety of **training data** [Kaplan et al., 2020, Grieve et al., 2025, Aryabumi et al., 2024, Chang et al., 2024, Conneau et al., 2020, Zhang et al., 2025]; and the **initial parameters** of a model [Bencomo et al., 2025, McCoy and Griffiths, 2023, Marinescu et al., 2024, Hu et al., 2025, Frankle and Carbin, 2018, Belrose and Scherlis, 2024].

Indeed, some recent work has taken exactly this approach [Tigges et al., 2024], mapping the developmental trajectories of multiple circuits (e.g., subject-verb agreement) among models in the Pythia suite [Biderman et al., 2023]; one crucial finding of this work was that these trajectories were relatively *aligned* across model sizes (see also 3.1). Another potentially relevant piece of evidence

[1] This is not always the case: for instance, in Cognitive Science research, English speakers from Western, industrialized countries have long been overrepresented, threatening the external validity of claims made on the basis of those samples [Henrich et al., 2010, Blasi et al., 2022].

[2] Note that additional axes of variation could augment this definition (e.g., particular prompts), which would further complicate the question of identifying a suitable reference class.

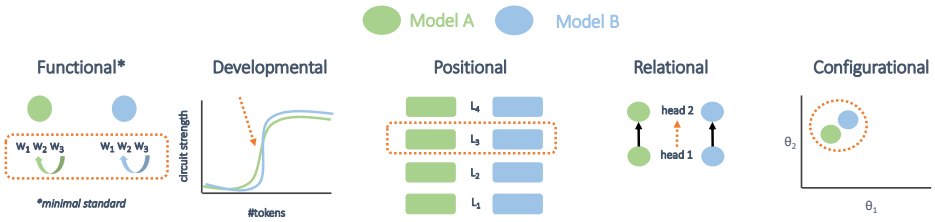


Figure 1: Graphical illustration of the five proposed axes of correspondence.

comes from work on the Platonic Representation Hypothesis [Huh et al., 2024], which predicts that we should observe more representational convergence between larger models trained on larger volumes of data—if this hypothesis is correct, we might also expect larger, better language models to converge on more similar mechanistic solutions.

The question of which model instances will develop a particular kind of circuit is no doubt challenging. Yet this framing points to a further, even more philosophically complicated challenge: what does it even mean to assert that two different models have the "same" circuit?

3 In what ways are two circuits the "same"?

Interpretability research typically involves the application of specific techniques to particular model instances, allowing researchers to determine which model components (e.g., which attention heads) perform a particular function or embed a specific concept [Clark et al., 2019, Wang et al., 2022, Olsson et al., 2022, Merullo et al., 2023, Manning et al., 2020, Merullo et al., 2024, Park et al., 2025, Zhang et al., 2024]. The result of applying such techniques might (for example) consist in a set of *head indices* believed to correspond to that function, e.g., $(L3, H1), (L4, H2)$. If a researcher's goal is idiographic (i.e., characterizing a given model instance), identifying this set might be sufficient: the circuit has been mapped in a particular model. But if a researcher's goal is nomothetic (i.e., drawing generalizations about other models), they face the question of what, in particular, could plausibly be generalized across model instances.

Yet, with some exceptions [Binhuraib et al., 2024], the position of a head *within* a layer in the standard transformer architecture is arbitrary: even among models with the same architecture trained on the same data, there is no intrinsic reason to expect that $(L3, H_{i=1})$ should consistently perform the same function, as opposed to some other head in the same layer $(L3, H_{i\neq 1})$. Thus, what exactly do we mean when we assert that two circuits in different model instances are "the same"?

3.1 Axes of Potential Correspondence

Here, I take inspiration from research in neurophysiology: although mechanistic heterogeneity in biological networks is well-attested [Prinz et al., 2004], researchers nonetheless strive to make generalizations about cell types and cell functions using various axes of correspondence. These include gene expression [Mukamel and Ngai, 2019], putative function [Moser et al., 2008, Knierim et al., 1995, Alexander and Nitz, 2015], temporal patterns [Rivière et al., 2022, Rivière and Rangel, 2017], and anatomical connectivity [Bates et al., 2019, Haber and Schneidman, 2022]. Mechanistic interpretability researchers might therefore identify analogous *axes of potential correspondence* between model instances:

- **Function**: Intuitively, the minimal standard for asserting circuit identity across model instances is whether components in each instance meet certain functional definitions, regardless of where in each model those components are located [Tigges et al., 2024]. Here, a claim might look like: *Attention heads*[3] *performing function X were identified across model instances $m_1, ..., m_n$.*

[3]Note that this framework could in principle be applied to any model component at varying levels of granularity; attention heads are simply used as an illustrative example.

- **Position**: One might also expect certain functions to be performed by components in similar positions across models. Here, researchers might differentiate between *absolute* position (e.g., always layer 3) and *relative* position (e.g., middle layers) [Riviere et al., 2025, Cheng and Antonello, 2024]. Here, a claim might look like: *Attention heads performing function X were identified at a layer depth of 0.5 across model instances $m_1, ..., m_n$.*
- **Developmental**: Just as human development is associated with particular milestones [Murray et al., 2007], some functions might plausibly emerge at similar points in the course of training, e.g., after having encountered a given number of tokens. Indeed, empirical research suggests that multiple specialized circuits begin forming at around 2B-10B tokens [Tigges et al., 2024, Olsson et al., 2022, Rivière and Trott, 2025, van der Wal et al., 2025, Jumelet et al., 2024]; we might also expect such findings to be marked by relative discontinuities or "phase transitions" during development [Chen et al., 2023, Hu et al., Kangaslahti et al., 2025]. Here, a claim might look like: *Attention heads performing function X emerged after $2B$ tokens were observed across model instances $m_1, ..., m_n$.*
- **Relational**: Model components could also be defined in terms of how they interact with other components. For example, induction circuits consist of an "induction head" and a "previous token head" [Olsson et al., 2022, Singh et al., 2024]. Similarly, [Zhang et al., 2024] report analogous circuit structures across model instances trained on different languages (or different combinations of languages). Here, a claim might look like: *Attention heads performing function X were identified in the layer immediately following attention heads performing function Y across model instances $m_1, ..., m_n$.*
- **Configurational**: Finally, particular functions or concepts might correspond to particular geometric configurations (e.g., in weight-space or activation-space). Here, a claim might look like: *Attention heads performing function X consistently occupied Y region of weight-space across model instances $m_1, ..., m_n$.*

This list is not exhaustive, but rather, provides a set of initial organizing principles that help ground claims about which mechanisms might generalize across model instances (and how). For example, this framework suggests that **induction heads** might be particularly promising candidates as generalizable model components. Induction heads participate in *induction circuits*, which are responsible for detecting whether a given token t has appeared earlier in a sequence (e.g., position s), then predicting that the subsequent token will be the one that previously occurred at position $s + 1$ [Elhage et al., 2021, Olsson et al., 2022]. Notably, induction heads satisfy multiple of the criteria discussed above: first, heads meeting the *functional* definition emerge in model instances of various sizes [Olsson et al., 2022, Singh et al., 2024]; second those heads follow similar *developmental* trajectories during training [Singh et al., 2024]; third, they (by definition) share a common *relational* structure with other heads, i.e., they participate in induction circuits [Olsson et al., 2022, Singh et al., 2024]; and fourth, they tend to occupy similar *relative positions* across model instances [Olsson et al., 2022].

3.2 On finding plausible mechanistic candidates

Mechanistic interpretability research aimed at identifying robust, generalizable model components might also benefit from focusing on identifying plausible candidates for mechanistic functions. The relative "success" of induction heads in this regard points to two additional criteria that may prove useful. First, their function is closely tied to the units over which models operate (i.e., token sequences) and was not defined *a priori* in terms of abstract human constructs—in this sense, they may even satisfy the definition of "concept enrichment" explored by Ayonrinde [2025]. Crucially, tracking previous sequences of tokens in the context is an intuitive solution to the problem faced by language models (predicting upcoming tokens); there is thus a clear link between induction head function and the language model training objective.

Second, while this operation is very concrete, it may be amenable to compositional abstraction [Olsson et al., 2022]: in some cases, induction heads may attend not only to exact repetitions of a token but more abstract correspondences (i.e., "types"). This could make them relevant for *in-context learning*, or ICL [Olsson et al., 2022, Singh et al., 2024], which in turn suggests that they could play a useful explanatory role in higher-level accounts of LLM behavior. Although there is debate about the extent to which induction heads are directly involved in ICL [Yin and Steinhardt, 2025, Feucht et al., 2025], efforts to connect *microscopic phenomena* to *macroscopic behavior* [Olah, 2023] can serve as a useful "North Star" for future research.

4 1-back Attention: An Empirical Case Study

If generalizability is to be a realistic ambition, then we should hope to observe some degree of robustness across *minimally different* model instances, such as different random seeds of the same (or similar) architecture trained on the same data.

The current section presents an empirical study exploring this question, focusing on **1-back attention heads**—defined as heads that direct attention from some target token to the immediately preceding token [Clark et al., 2019]. From a definitional perspective, these **1-back heads** satisfy the proposed criterion of closely tracking the actual units over which models operate (i.e., token sequences; see Section 3.2); as with induction heads (see Section 3.2), tracking the immediately preceding token seems intuitively helpful for making predictions about upcoming tokens, therefore tying the putative function of these heads to the overall training objective of the model.

Because 1-back heads are likely very useful—and also quite simple in terms of their behavior—one might expect them to emerge across many model instances, including small models, making them a suitable test case for the proposed axes of correspondence (Section 3.1). For the *reference class*, I limit the analysis to (arguably) the simplest possible "population": different random seeds of four model architectures (Pythia-14M, Pythia-70M, Pythia-160M, Pythia-410M) trained on the same data [Biderman et al., 2023]. Note also that the approach adopted below focuses on characterizing the *behavior* of these heads (in terms of their attention patterns), which is a necessary but not sufficient prerequisite for firmly establishing their function as 1-back heads.

This approach allows us to address three related research questions:

- RQ1 To what extent do we observe *inter-seed* and *inter-model* regularities in terms of the developmental trajectories and relative position of 1-back heads? Here, we find striking *developmental regularities*, consistent with prior work on other model components [Jumelet et al., 2024, Olsson et al., 2022, Tigges et al., 2024]; evidence for positional regularity is more mixed.
- RQ2 What divergences do we observe across model instances, and which (if any) properties allow us to predict these divergences? Here, I find that larger models show an *earlier onset* of 1-back heads than smaller models, a steeper *slope* of 1-back attention over pretraining, and a higher *peak* of 1-back attention.
- RQ3 What predicts *convergences* in developmental trajectories across model instances? Here, different seeds of the same architecture show the strongest correlation; when comparing model instances of different sizes, higher correlations were predicted by the size of each model being compared.

A link to a GitHub repository with code and data required to reproduce these analysis can be found at `https://github.com/seantrott/mechinterp_generalizability`.

4.1 Methods

I selected the Pythia suite of auto-regressive English models [Biderman et al., 2023], focusing on the nine random seeds released for Pythia-14M, 70M, 160M, and 410M [van der Wal et al., 2025]. Each model was assessed at 16 training checkpoints (i.e., all available checkpoints up to and including step 1000, followed by step 1000, 50000, 100000, and 143000). As described in Biderman et al. [2023], each model was trained on approximately 300B tokens. All models were accessed through the HuggingFace *transformers* library Wolf et al. [2020] and run on a 2022 Mac laptop. The Pythia models are licensed under an Apache License, Version 2.0.

Each model at each checkpoint was presented with sentences from the Natural Stories Corpus [Futrell et al., 2021]. The Natural Stories Corpus consists of 10 English-language stories, each containing approximately 1000 words. This served as a repository of naturalistic sentences with which to probe attention head behavior (note that the behavior of attention heads was remarkably consistent across different stories; see Appendix B). The corpus is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The goal was to assess the developmental and positional properties of putative **1-back attention heads**. Here, "1-back attention" was defined as directing attention from a target token to the immediately

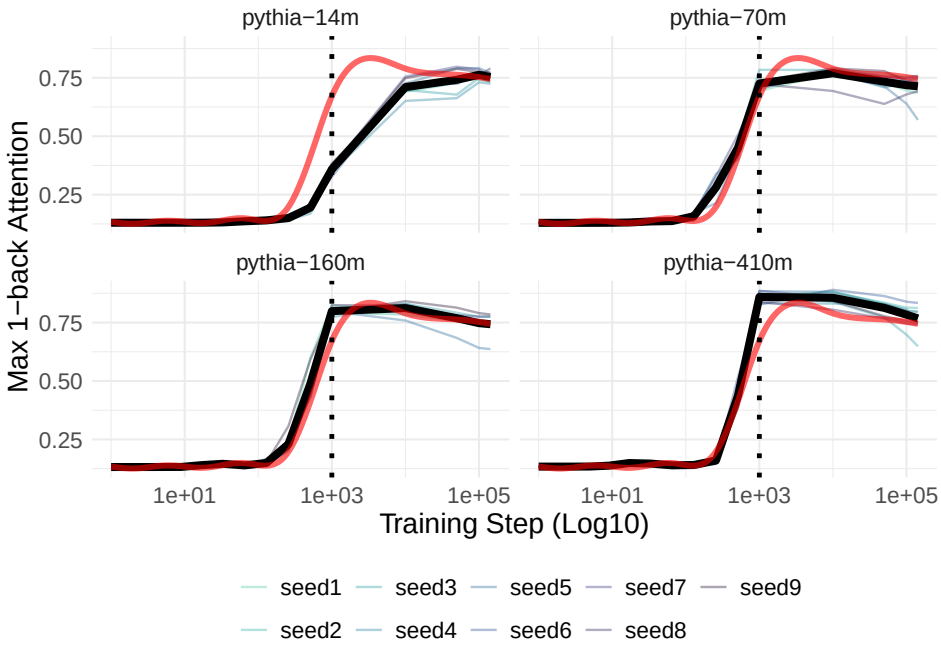


Figure 2: Maximum 1-back attention for each seed of Pythia 14M, 70M, 160M, and 410M. Dark black line indicates average across each seed for that model. Red reference line indicates predictions from a generalized additive model (GAM) fit to all data points (i.e., across models).

preceding token. For each head in each model (at each checkpoint), I calculated the average 1-back attention for each sentence.[4] More precisely, if $A_h(i, j)$ represents the attention assigned by head h from token i to token j, and n represents the number of tokens in a given sentence I calculated:

$$R_h = \frac{1}{n-1} \sum_{i=2}^{n} A_h(i, i-1) \tag{1}$$

Note that i starts at 2 to exclude the first token in the sequence. All analyses and visualizations were conducted in R [R Core Team, 2025].

4.2 Results

4.2.1 RQ1: Inter-seed consistency

First, I asked about the developmental and positional consistency of putative 1-back attention across different seeds of the same architecture. As depicted in Figure 2, different seeds of the same model showed striking regularity in their developmental trajectories. Even different models exhibited remarkably similar patterns: selective 1-back attention tended to emerge around 10^3 training steps, corresponding to roughly $2B$ tokens of exposure. A generalized additive model (GAM) fit to all data points from all models (i.e., the maximum attention at each step for each seed of each model) using Log Training Step as a predictor achieved an $R^2 = 0.95$; as depicted in Figure 2, the GAM's predictions (in red) are consistent with an expected onset in 1-back attention occurring between 512 and 2000 steps.

[4]Note that qualitatively identical results for the developmental analyses (though not the positional analyses) were obtained with alternative operationalizations, such as the ratio between the average 1-back attention and the average self-attention. The two metrics were generally highly correlated within each model at the final step ($r \geq 0.68$ for all models). Average 1-back attention was favored in the final analysis because it did not depend upon arbitrary assumptions about the appropriate baseline; moreover, because the attention scores are normalized, the average 1-back attention can be interpreted as the proportion of attention directed by a given head to previous tokens.

The *position* of these putative 1-back heads exhibited considerably more variance across seeds and especially across models. Figure 3a depicts the maximum 1-back attention at each *layer* of each random seed for each model at the final pre-training step. For $14M$, 1-back heads tended to emerge either in layer 3 or in layer 4[5]; see also Appendix A. Overall, positional consistency across model architectures was relatively low, though there was some evidence for middle layers showing a peak in 1-back attention. A linear mixed effects model predicting average 1-back attention at the final step from each individual head of each model instance revealed a significantly negative effect of Layer Depth [$\beta = -0.11, SE = 0.005, p < 0.001$], i.e., *later* layers were associated with significantly less attention to previous tokens on average.

Together, these results point to a high degree of *developmental correspondence* in putative 1-back attention heads across model instances in the Pythia suite, but also suggest a limited degree of *positional correspondence*: 1-back heads are systematic in *when* but not *where* they appear.

4.2.2 RQ2: Model divergences in timing

Although 1-back attention heads were extremely consistent in *when* they developed across model instances, subtle differences in timing are revealed by comparing the trajectory of each individual model to the fit GAM predictions (Figure 2). Relative to the predictions, smaller models (like $14M$) had a delayed *onset* of 1-back heads, a shallower *slope* of 1-back attention over time, and a reduced *peak*; conversely, larger models (like $410M$) showed a sharper *slope* and a higher *peak*.

To quantify these apparent divergences, I first operationalized each construct (onset, slope, and peak) as follows. The *onset* of 1-back attention for a given seed was defined as the earliest step where the change in maximum 1-back attention relative to the change in log training step (d_{ratio}) exceeded some threshold; in order to avoid dependence on a particular threshold, this was assessed for a range of thresholds $(0.01, 0.3)$ for each seed and averaged across the resulting values for that range. Intuitively, this measure reflects the average *earliest step* at which 1-back attention sharply increased for a given instance.[6] The *slope* was identified by regressing the maximum 1-back attention at each step against Log Training Step and extracting the resulting slope estimate. Finally, the *peak* was defined as the maximum 1-back attention (i.e., across all steps for that seed).

Each measure was then regressed (in a separate linear mixed effects model) against Log Parameters, with seed as a random intercept. Log Parameters was significantly related with each dependent variable in the expected direction. That is, larger models displayed a reduced *onset* of 1-back attention [$\beta = -0.18, SE = 0.05, p < .001$], increased *slope* [$\beta = 0.01, SE = 0.003, p < .001$], and a higher *peak* [$\beta = 0.07, SE = 0.01, p < 0.001$]. These relationships are also depicted in Figure 3.

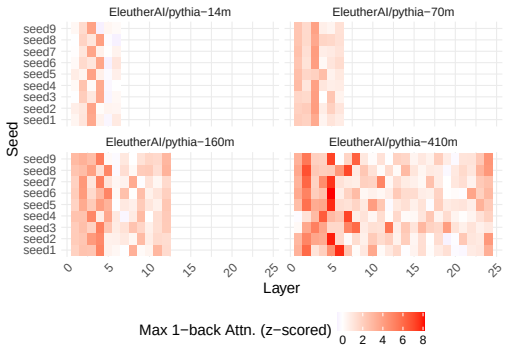
4.2.3 RQ3: What predicts convergence?

A central question concerning generalizability is what, if anything, predicts that two model instances will belong to the same "reference class" with some respect to some *axis of correspondence* (Section 3.1). As described in Section 4.2.1, I observed striking convergence in the developmental trajectories of 1-back attention across random seeds of the same model, and even relatively strong temporal alignment between models of different sizes. Here, still focusing on the *temporal axis*, I asked a related question: which properties predict higher temporal convergence between two model instances?

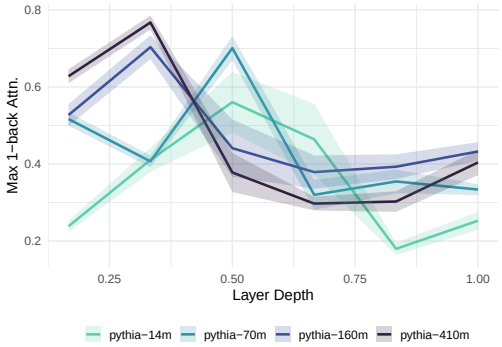
First, I calculated the Pearson's correlation between every pair of 1-back attention trajectories (i.e., for each pair of model instances). Unsurprisingly, random seeds of the same *architecture* exhibited substantially higher correlation ($M = 0.99, SD = 0.04$) on average than instances from different architectures ($M = 0.95, SD = 0.003$). I then constructed a linear regression with correlation between each pair of model instances ($r_{i,j}$) as a dependent variable; predictors included a factor indicating whether the two instances were the Same Model, as well as the Number of Parameters (Log10) of model m_i and model m_j. Consistent with the descriptive results, Same Model positively predicted higher $r_{i,j}$ [$\beta = 0.05, SE = 0.002, p < .001$]; holding Same Model constant, both Number of Parameter predictors were also positively related with higher $r_{i,j}$ [$\beta = 0.03, SE = 0.002, p <$

[5]This echoes other work revealing *bimodal* distributions in the mechanisms and behaviors that emerge across random seeds [Zhao et al., 2025], and suggests that there might be multiple "attractors" in weight-space with the respect to 1-back attention head emergence.

[6]Note that the results are robust to different thresholds, as well as different operationalizations of 1-back attention onset.

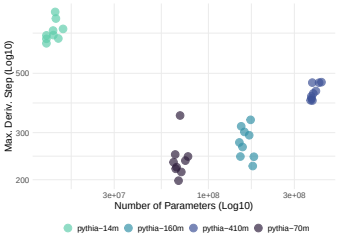


(a) Z-scored 1-back attention for each head at the final step of each Pythia-14M random seed. Dotted red line represents 3 standard deviations from the mean for that seed.

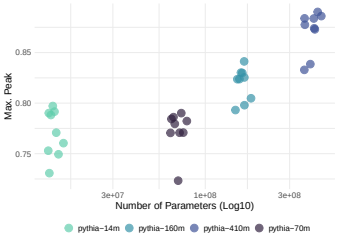


(b) Maximum attention to previous tokens at each *binned layer depth* across models (shading indicates one standard error calculated across seeds).

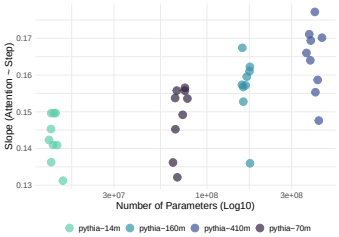
Figure 3: Different random seeds of the same architecture showed considerable variance in where putative 1-back attention heads formed.



(a) Earliest *onset* of 1-back attention heads in each seed of each model.



(b) Max *peak* of 1-back attention heads in each seed of each model.



(c) *Slope* of 1-back attention (against pretraining step) in each seed of each model.

Figure 4: Larger Pythia models tended to show a slightly earlier *onset* of 1-back attention, a higher *peak*, and a steeper *slope*.

.001]. That is, stronger temporal convergence was observed among instances belonging to the *same* architecture and among instances of *larger* (different) architectures; similar results are reported in the Appendix (Section C).

4.3 Discussion

The primary goal of this empirical case study was as a *demonstration of viability* for the proposed *axes of correspondence* (Section 3.1)—specifically focusing on the positional and temporal axes of model similarity. 1-back attention heads were selected because they were a well-established phenomenon [Clark et al., 2019] that would likely emerge in even small models, and which also satisfied the proposed criteria for plausible mechanistic candidates (Section 3.2).

Concretely, the case study addressed three research questions. First, I observed considerable *inter-seed* and *inter-model* convergence along the temporal axis (see Figure 2), though there was less positional consistency: 1-back heads were highly systematic in *when* but not *where* they emerged. Second, I observed subtle timing differences as a function of model size: larger models showed an earlier onset, steeper slope, and higher peak of 1-back attention (see Figure 3). Finally, inter-seed temporal convergence was (unsurprisingly) highest among instances of the same architecture; among instances of different architectures, temporal convergence was higher when both instances were larger (Section 4.2.3), possibly consistent with the Platonic Representation Hypothesis [Huh et al., 2024]. Together, these results suggest that at least when it comes to putative 1-back attention heads, model mechanisms are more constrained by *developmental* features than *positional ones*, which is itself informative about the nature of various constraints operating over head specialization.

A key limitation of this work is that candidate 1-back heads were defined in terms of their attentional behavior—their *functional* role in model predictions was not investigated. Future work could conduct further analyses to investigate the functional axis specifically, perhaps connecting these heads to broader circuits in which they may or may not participate; the latter goal would also connect to other potential proposed axes of correspondence, such as the *relational* structure of heads to other model components.

5 Objections and Limitations

Thus far, I have argued: first, that *generalizability* is a central epistemological challenge in the science of LLM interpretability; and second, that progress could be made by considering particular *axes of correspondence* among model instances. I have also presented a case study illustrating the value of using these proposed axes to guide empirical investigation. Here, I consider possible objections to the key arguments in the paper, as well as potential replies to each objection.

Objection: Interpretability *is* idiographic, not nomothetic.

Reply: If interpretability is idiographic, then generalizability is indeed not a concern. That said, interpretability research does (arguably) appear to be nomothetic in nature, with identifying "universal" circuits as a high-level goal [Olah et al., 2020, Olah, 2023]. Further, as argued in Olah et al. [2020], (near-)universality would also make interpretability a much more *tractable* field:

> In the same way, the universality hypothesis determines what form of circuits research makes sense. If it was true in the strongest sense, one could imagine a kind of "periodic table of visual features" which we observe and catalogue across models. On the other hand, if it was mostly false, we would need to focus on a handful of models of particular societal importance and hope they stop changing every year. There might also be in between worlds, where some lessons transfer between models but others need to be learned from scratch. [Olah et al., 2020]

Objection: Automated interpretability techniques reduce the need for generalization, as putative circuits can quickly be discovered in new model instances.

Reply: In a world of fully automated interpretability [Conmy et al., 2023], a suite of methods would simply be applied to each model instance as necessary, perhaps reducing the need to speculate about "unobserved" instances. However, my position is that a coherent theory of generalizability would still be of value in such a world for two reasons. First, a theory of generalizability is, at root, an articulation of what makes two model instances similar or different along some dimension; it is not only *useful* for conducting research but could also be seen as a *progress marker* of research. Second, in such a world there would presumably be vast empirical data characterizing large numbers of individual model instances—a theoretical framework such as the one proposed here (Section 3.1) would give researchers a *lexicon* [Kuhn, 1982, Contreras Kallens and Dale, 2018] to describe observed convergences and divergences between model instances and make sense of this landscape.

Objection: The case study is limited—1-back heads are simple, and a narrow reference class was chosen.

Reply: The goal of the case study was to investigate the viability of identifying correspondences across model instances, which is why a simple phenomenon (1-back heads) was selected, as well as a relatively narrow reference class (different seeds or sizes of the same underlying architecture trained on the same data). The underlying logic was that if generalizability is to be a goal of interpretability research, we should be able to establish it in at least this case, which perhaps reflects a "lower bound" of tractability. The relative success of the case study suggests these axes of correspondence may serve as fruitful guides for future work, which could expand to more complex circuits or mechanisms.

Objection: The functional axis is sufficient to establish circuit identity.

Reply: The functional axis of correspondence (Section 3.1) is arguably the *minimal standard* for establishing whether components of different model instances "do the same thing". My argument is not that other axes are *necessary*, but rather that they are additional organizing principles for assessing similarities and differences between circuits. It is informative to be able to assert that two circuits satisfy the same functional criteria but emerge at different timepoints in different models—

particularly if those divergences can be related to other points of departure between those models (see Section 4.2.2). Moreover, finding correspondences along these other dimensions might also strengthen our confidence in the robustness of a particular result or in the identity of circuits across model instances: intuitively, two circuits seem more conceptually and mechanistically similar if they not only satisfy the same functional definition [Wang et al., 2022] but also exhibit similar *developmental trajectories* (or converge along other axes of correspondence). That said, functional alignment is probably the only strictly necessary axis to assert that two circuits are implementing the same function; as in evolutionary biology, different systems often achieve the same goals or perform the same computations in different ways.

Objection: Random seeds (and pretraining checkpoints) are not available for most models, making generalizability too difficult to investigate.

Reply: Indeed, the problem is even worse—the space of *actual* model instances is also not itself a representative sample of the distribution of *possible* models. With some exceptions [Biderman et al., 2023], available models are driven by specific research or commercial interests and not necessarily with the goal of exhaustively characterizing the space of possible models. However, the fact that generalizability will be hard to investigate is not a refutation of its epistemological importance for the field of interpretability. Fully addressing the gap is beyond the scope of this paper (and will likely require large-scale coordination between institutions in multiple sectors), but individual researchers do still have options available to them, such as the Pythia suite [Biderman et al., 2023, van der Wal et al., 2025] and OLMo 2 [OLMo et al., 2025]. Moreover, not every interpretability study needs to address every axis of correspondence: a study investigating developmental convergences across seeds might focus on the Pythia suite [Biderman et al., 2023], but a study focused on positional or relational consistency would not necessarily need pretraining checkpoints. Nonetheless, the field would clearly benefit from a larger number of open-source models subjected to controlled training regimes.

6 The Path Forward

The shift from a pre-paradigmatic stage of research [Olah et al., 2020, Olah, 2023] to more established research practices and inferential principles will require a combination of theoretical and methodological refinement. In this paper, I have argued that *generalizability* is a major epistemological gap in the scientific study of LLM mechanisms. Yet enumerating challenges is often straightforward; identifying markers of progress on those challenges can be much more difficult. In that spirit, I have drawn on the growing body of existing research [Tigges et al., 2024, Zhang et al., 2024, Olsson et al., 2022] to propose *axes of correspondence* that might serve as organizing principles with which to guide questions about generalizability. I have also presented the results of a case study validating the utility of this framework in identifying areas of convergence and divergence between model instances. Moving forward, one marker to look for is the construction of a theoretically legible typology (or "phylogeny") that makes clearly articulable predictions about which pairs of model instances will share similar mechanisms along which axes of correspondence.

7 Reproducibility Statement

A link to a GitHub repository with code and data required to reproduce these analyses can be found at `https://github.com/seantrott/mechinterp_generalizability`.

Acknowledgments and Disclosure of Funding

Thank you to Pamela Rivière, Kola Ayonrinde, Cameron Jones, and Oisín Parkinson-Coombs for comments on an earlier version of this draft. The work relied on the HuggingFace *transformers* package [Wolf et al., 2020], the open-source Pythia suite [Biderman et al., 2023, van der Wal et al., 2025], the Natural Stories Corpus [Futrell et al., 2021], and the open-source R computing environment [R Core Team, 2025].

References

Andrew S Alexander and Douglas A Nitz. Retrosplenial cortex maps the conjunction of internal and external spaces. *Nature neuroscience*, 18(8):1143–1151, 2015.

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? exploring impact of code in pre-training. *CoRR*, abs/2408.10914, 2024. URL `https://doi.org/10.48550/arXiv.2408.10914`.

Kola Ayonrinde. Position: Interpretability is a bidirectional communication problem. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*, 2025. URL `https://openreview.net/forum?id=O4LaRH4zSI`.

Kola Ayonrinde and Louis Jaburi. Evaluating explanations: An explanatory virtues framework for mechanistic interpretability. 2025a. Forthcoming.

Kola Ayonrinde and Louis Jaburi. A mathematical philosophy of explanations in mechanistic interpretability: The strange science part i.i, 2025b. forthcoming.

Alexander Shakeel Bates, Jasper Janssens, Gregory Sxe Jefferis, and Stein Aerts. Neuronal cell types in the fly: single-cell anatomy meets single-cell genomics. *Current opinion in neurobiology*, 56: 125–134, 2019.

Samuel J Beck. The science of personality: Nomothetic or idiographic? *Psychological Review*, 60 (6):353, 1953.

Nora Belrose and Adam Scherlis. Understanding gradient descent through the training jacobian. *arXiv preprint arXiv:2412.07003*, 2024.

Gianluca Bencomo, Max Gupta, Ioana Marinescu, R Thomas McCoy, and Thomas L Griffiths. Teasing apart architecture and initial weights as sources of inductive bias in neural networks. *arXiv preprint arXiv:2502.20237*, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/biderman23a.html`.

Taha Osama A Binhuraib, Greta Tuckute, and Nicholas Blauch. Topoformer: brain-like topographic organization in transformer language models through spatial querying and reweighting. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. Over-reliance on english hinders cognitive science. *Trends in cognitive sciences*, 26(12):1153–1170, 2022.

J Douglas Carroll and Phipps Arabie. Multidimensional scaling. *Measurement, judgment and decision making*, pages 179–250, 1998.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.236. URL `https://aclanthology.org/2024.emnlp-main.236/`.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. *arXiv preprint arXiv:2309.07311*, 2023.

Emily Cheng and Richard J Antonello. Evidence from fmri supports a two-phase abstraction process in language models. *arXiv preprint arXiv:2409.05771*, 2024.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, page 276. Association for Computational Linguistics, 2019.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747/`.

Pablo Contreras Kallens and Rick Dale. Exploratory mapping of theoretical landscapes through word use in abstracts. *Scientometrics*, 116(3):1641–1674, 2018.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Sheridan Feucht, Eric Todd, Byron Wallace, and David Bau. The dual-route model of induction. *arXiv preprint arXiv:2504.03022*, 2025.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018.

Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. The natural stories corpus: a reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77, 2021.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411, 2025.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Adam Haber and Elad Schneidman. Learning the architectural features that predict functional similarity of neural networks. *Physical Review X*, 12(2):021051, 2022.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

Michael Y Hu, Angelica Chen, Naomi Saphra, and Kyunghyun Cho. Latent state models of training dynamics. *Transactions on Machine Learning Research*.

Michael Y Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. Between circuits and chomsky: Pre-pretraining on formal languages imparts linguistic biases. *arXiv preprint arXiv:2502.19249*, 2025.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *CoRR*, abs/2405.07987, 2024. URL https://doi.org/10.48550/arXiv.2405.07987.

Anna A Ivanova. Running cognitive evaluations on large language models: The do's and the don'ts. *arXiv preprint arXiv:2312.01276*, 2023.

Jaap Jumelet, Lisa Bylinina, Willem Zuidema, and Jakub Szymanik. Black big boxes: Do language models hide a theory of adjective order? *arXiv preprint arXiv:2407.02136*, 2024.

Sara Kangaslahti, Elan Rosenfeld, and Naomi Saphra. Hidden breakthroughs in language model training. *arXiv preprint arXiv:2506.15872*, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

James J Knierim, Hemant S Kudrimoti, and Bruce L McNaughton. Place cells, head direction cells, and the learning of landmark stability. *Journal of Neuroscience*, 15(3):1648–1659, 1995.

Thomas S Kuhn. Commensurability, comparability, communicability. In *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association*, volume 1982, pages 668–688. Cambridge University Press, 1982.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *Feature Extraction: Modern Questions and Challenges*, pages 196–212. PMLR, 2015.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

Ioana Marinescu, R Thomas McCoy, and Tom Griffiths. Distilling symbolic priors for concept learning into neural networks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

James L McClelland and David E Rumelhart. An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review*, 88(5):375, 1981.

R Thomas McCoy and Thomas L Griffiths. Modeling rapid language learning by distilling bayesian priors into artificial neural networks. *arXiv preprint arXiv:2305.14701*, 2023.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. *arXiv preprint arXiv:2310.08744*, 2023.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 61372–61418. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/70e5444e5f331f7f5431f302110b97af-Paper-Conference.pdf.

Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31(1):69–89, 2008.

Aaron Mueller and Tal Linzen. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11237–11252, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.629. URL https://aclanthology.org/2023.acl-long.629/.

Eran A Mukamel and John Ngai. Perspectives on defining cell types in the brain. *Current opinion in neurobiology*, 56:61–68, 2019.

Graham K Murray, Peter B Jones, Diana Kuh, and Marcus Richards. Infant developmental milestones and subsequent cognitive function. *Annals of neurology*, 62(2):128–136, 2007.

Chris Olah. Interpretability dreams. *Transformer Circuits*, 2023. https://transformer-circuits.pub/2023/interpretability-dreams/index.html.

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

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL `https://arxiv.org/abs/2501.00656`.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Yein Park, Chanwoong Yoon, Jungwoo Park, Minbyul Jeong, and Jaewoo Kang. Does time have its place? temporal heads: Where language models recall time-specific information. *arXiv preprint arXiv:2502.14258*, 2025.

Jackson Petty, Sjoerd van Steenkiste, Fei Sha, Ishita Dasgupta, Dan Garrette, and Tal Linzen. The impact of depth and width on transformer language model generalization. 2023.

Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit parameters. *Nature neuroscience*, 7(12):1345–1352, 2004.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. URL `https://www.R-project.org/`.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL `https://openreview.net/forum?id=j6NxpQbREA1`.

Pamela D Rivière and Lara M Rangel. Spike-field coherence and firing rate profiles of ca1 interneurons during an associative memory task. In *Association for Women in Mathematics Research Symposium*, pages 161–171. Springer, 2017.

Pamela D Rivière, Gabriel Schamberg, Todd P Coleman, and Lara M Rangel. Modeling relationships between rhythmic processes and neuronal spike timing. *Journal of Neurophysiology*, 128(3): 593–610, 2022.

Pamela D Riviere, Anne L. Beatty-Martínez, and Sean Trott. Evaluating contextualized representations of (Spanish) ambiguous words: A new lexical resource and empirical analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8322–8338, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.422. URL `https://aclanthology.org/2025.naacl-long.422/`.

Pamela Rivière and Sean Trott. Start making sense(s): A developmental probe of attention specialization using lexical ambiguity. *arXiv preprint*, 2025.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. Benchmarks as microscopes: A call for model metrology. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=bttKwCZDkm.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Aaditya K. Singh, Ted Moskovitz, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. Llm circuit analyses are consistent across training and scale. *arXiv preprint arXiv:2407.10827*, 2024.

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, pages 1–25, 2025.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

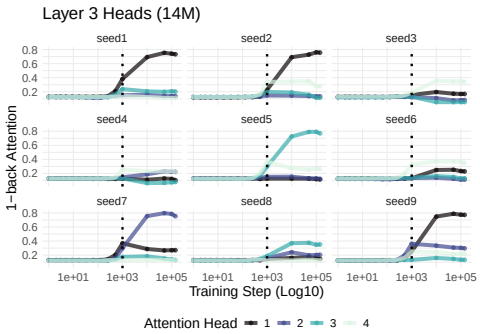
Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. Phylolm: Inferring the phylogeny of large language models and predicting their performances in benchmarks. *arXiv preprint arXiv:2404.04671*, 2024.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=C7XmEByCFv.

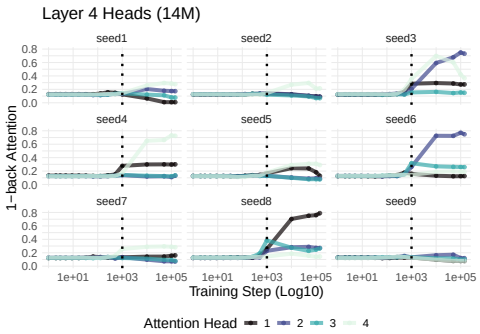
Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. The same but different: Structural similarities and differences in multilingual language modeling. *arXiv preprint arXiv:2410.09223*, 2024.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NCrFA7dq8T.

Rosie Zhao, Tian Qin, David Alvarez-Melis, Sham Kakade, and Naomi Saphra. Distributional scaling laws for emergent capabilities. *arXiv preprint arXiv:2502.17356*, 2025.



(a) 1-back attention for layer 3 heads across Pythia-14M seeds and pre-training checkpoints.



(b) 1-back attention for layer 4 heads across Pythia-14M seeds and pre-training checkpoints.

Figure 5: In each random seed of Pythia-14M, 1-back attention heads emerged at roughly similar training checkpoints.

A Individual head trajectories in 14m

The developmental trajectories depicted in the primary manuscript collapsed across heads and layers for the purpose of illustrating temporal patterns across all models tested. Here, I depict the trajectories of 1-back attention for each individual head in each random seed of Pythia-14M for layer 3 (Figure 5a) and layer 4 (Figure 5b).

B Consistency across stories

One question that arises is about the generalizability of the results reported in the primary manuscript across *input sentences*. That is, how dependent is the developmental trajectory observed upon the corpus used to assess attention head function? To address this, I plotted the maximum 1-back attention for each random seed of Pythia-14M for each *story* from the Natural Stories Corpus. The developmental trajectories are extremely similar, with changes in attention beginning around 10^3 training steps for all random seeds, for all stories.

C Multi-dimensional scaling of seeds

In the primary manuscript, I reported the average correlation between seeds belonging to the same architecture and between seeds belonging to different architectures; a key finding was among seeds of *different architectures*, stronger temporal convergences were observed when both instances being compared were larger.

Another perspective on this result comes from embedding the correlation matrix of all model instances in a $2D$ space using multi-dimensional scaling (MDS) [Carroll and Arabie, 1998]. As depicted in Figure 7, the first MDS component appeared to track model size. Moreover, seeds of smaller models ($14M$) exhibited tight clustering and relatively larger separation from larger models.

I quantified this trend by first calculating the centroid among each set of seeds (i.e., for 14M, 70M, etc.). I then calculated the Separation Ratio, defined as the mean distance of each centroid to other centroids, divided by the mean distance of each point within a model class to the centroid of that class. It decreases systematically with model size, reflecting the combination of wider internal spread and closer inter-cluster proximity in larger models. The Separation Ratio systematically declined with model size: 14M was the largest (11.4), followed by 70M (3.37), and 160M (2.42). The results are nuanced, however, as 410M exhibited a larger Separation Ratio than 70M and 160M (4.0), though still considerably smaller than 14M; see also Figure 2.

Together, this suggests that at least along the temporal axis, smaller model instances of the same architecture are relatively different from other models—conversely, larger models of different architectures are relatively more similar to each other.

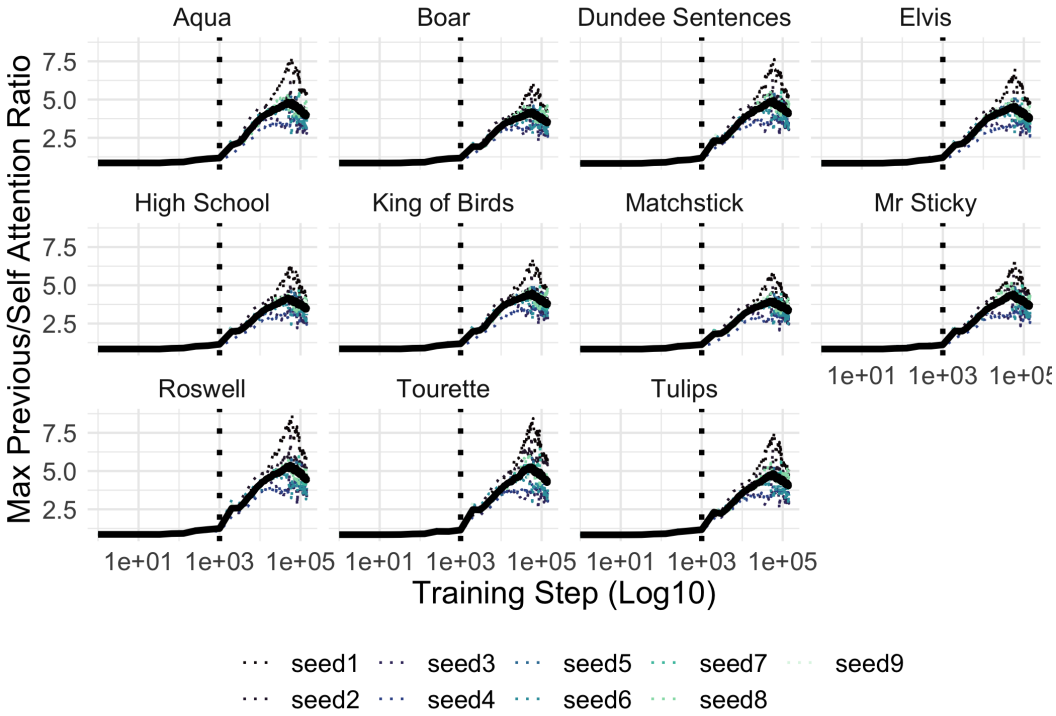


Figure 6: Developmental trajectory of putative 1-back attention heads across random seeds of Pythia-14M and across the story corpora used to assess this behavior. Across all heads at a given checkpoint for a given model, the maximum 1-back attention was calculated.

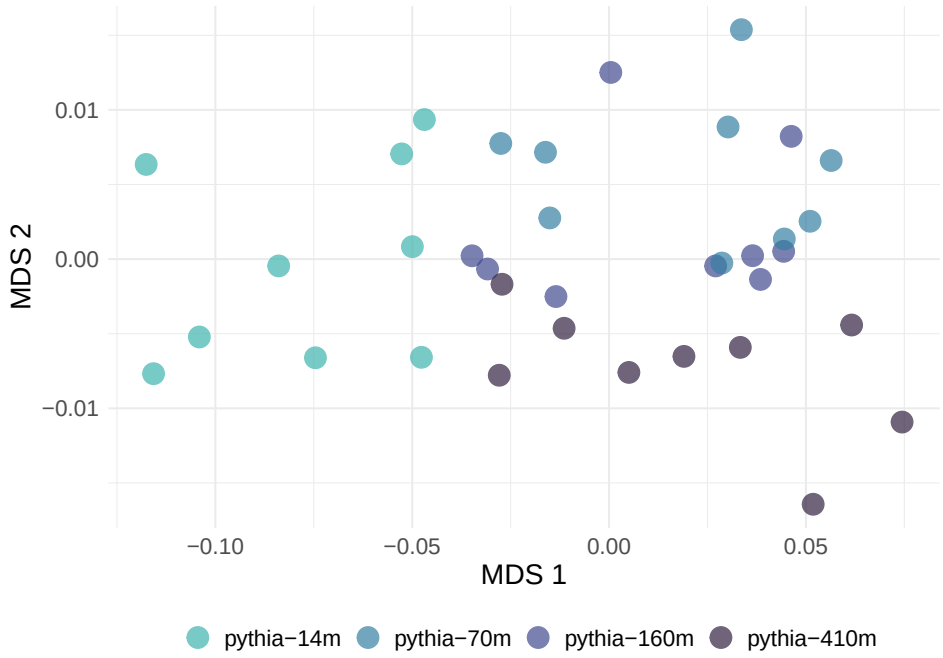


Figure 7: Results of applying multi-dimensional scaling (MDS) to the correlation matrix of all model instance pairs. Each MDS point represents a given seed of a particular model.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The goal of the paper is to argue that generalizability is a major epistemological challenge in the study of LLMs, propose a theoretical framework for investigating it, and present the results of a novel case study; the abstract and introduction summarizes this.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Yes, there is an Objections and Limitations section devoted to identifying limitations, particularly in the theoretical framework proposed (which is the central point of the paper).

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [NA]

Justification: The results in the paper are empirical results and I have attempted to motivate and contextualize them (including the assumptions behind various operationalizations, etc.); there are no theoretical results depending on a proof.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The Methods section discloses the models run and data used as input to the models; the Results section describes the details of specific operationalizations and statistical analyses. Additionally, a link to a GitHub repository with code and data required to reproduce these analysis can be found at `https://github.com/seantrott/mechinterp_generalizability`.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: A link to a GitHub repository with code and data required to reproduce these analysis can be found at `https://github.com/seantrott/mechinterp_generalizability`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [NA]

   Justification: No model training was performed.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Statistical results are reported where relevant, including with standard errors and significance. Additionally, figures displaying means also include confidence intervals reflecting the standard error; other figures display the "raw" data.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The Methods section includes a sentence describing the compute environment in which models were run.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Human-subjects data was not used, and all corpus data presented to the model was extracted from a publicly available research corpus.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [No]

Justification: The answer is "yes" if this includes the impact on the scientific community, given that the goal of the work is to shape the research conversation around mechanistic interpretability. But it is "no" otherwise, as the paper does not discuss broader societal implications at length.

Guidelines:

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

11. **Safeguards**

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

Answer: [No]

Justification: No new corpus data or language models were created in this research; the data outputs reflect the behavior of specific model components (attention heads) and are unlikely candidates for misuse.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The language models and corpus data used in the paper are cited, and the licenses for each are also described.

Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: All data and code is accompanied by documentation (though new assets are minimal).

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

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

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

Answer: [NA]

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

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: The *focus* of the research is LLMs, but LLMs were not used in this research.

Guidelines:

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