Start Making Sense(s): A Developmental Probe of Attention Specialization Using Lexical Ambiguity

Pamela D. Rivière
University of California, San Diego
pdrivier@ucsd.edu

Sean Trott
Rutgers University - Newark
sean.trott@rutgers.edu

Abstract

Despite an in-principle understanding of self-attention matrix operations in Transformer language models (LMs), it remains unclear precisely how these operations map onto *interpretable* computations or functions—and how or when individual attention heads develop specialized attention patterns. Here, we present a pipeline to systematically probe attention mechanisms, and we illustrate its value by leveraging lexical ambiguity—where a single word has multiple meanings—to isolate attention mechanisms that contribute to word sense disambiguation. We take a "developmental" approach: first, using publicly available Pythia LM checkpoints, we *identify* inflection points in disambiguation performance for each LM in the suite; in $14M$ and $410M$, we identify heads whose attention to disambiguating words covaries with overall disambiguation performance across development. We then *stress-test* the robustness of these heads to stimulus perturbations: in $14M$, we find limited robustness, but in $410M$, we identify multiple heads with surprisingly generalizable behavior. Then, in a *causal analysis*, we find that ablating the target heads demonstrably impairs disambiguation performance, particularly in $14M$. We additionally reproduce developmental analyses of $14M$ across all of its random seeds. Together, these results suggest: that disambiguation benefits from a constellation of mechanisms, some of which (especially in $14M$) are highly sensitive to the position and part-of-speech of the disambiguating cue; and that larger models ($410M$) may contain heads with more *robust* disambiguation behavior. They also join a growing body of work that highlights the value of adopting a developmental perspective when probing LM mechanisms.

1 Introduction

Transformer-based language models (LMs) (Vaswani et al., 2017) have proven adept both at the primary language modeling objective and at a variety of downstream tasks. A core innovation of Transformers is the attention block, which consists of a series of parallel "attention heads": sets of Query, Key, and Value matrices that each transform token vectors (or "embeddings") according to linguistic context. Yet despite the success of these systems, there remains limited mechanistic understanding of how specific model components give rise to observable behaviors or implement particular functions; developing such understanding is the central goal of research on model interpretability (Sharkey et al., 2025; Geiger et al., 2022; Mueller et al., 2025). In particular, a growing body of work suggests that different attention heads learn to "specialize" (ie., over the course of pre-training) in which kinds of context they attend to (Olsson et al., 2022; Wang et al., 2023; Merullo et al., 2024; Park et al., 2025).

Here, we investigate contextualization mechanisms in the Transformer LM architecture by combining careful psycholinguistic *experimental design* and targeted *editing of LM weights* at multiple pre-training LM checkpoints. The former allows us to identify key "developmental milestones" associated with improvements in contextualization over the course of pre-training, and also isolate candidate attention heads that integrate information from disambiguating cues. The latter allows us to assess the causal influence of these components in disambiguation performance.

Specifically, we leverage a dataset of human relatedness judgments of (English) ambiguous words (Trott and Bergen, 2021) to probe which components of Pythia-$14M$ and Pythia-$410M$ (Biderman et al., 2023) contribute to

contextualization—and when these mechanisms develop throughout pretraining. As noted below (Section 2), ambiguity is a useful probe for several reasons: it is exceedingly common (Rodd et al., 2004), word sense disambiguation is an established task of interest in the field of natural language processing (NLP) (Haber and Poesio, 2020), and the process of disambiguation is a specific sub-case of the more general challenge of *contextualization*—one that importantly allows us to vary the surrounding context while keeping the target token identity the same. This work extends the study of self-attention and linguistic contextualization, and underscores the value of deploying carefully designed variations in stimuli— coupled with causal manipulations of isolated attention heads—to sketch any given head's *functional scope* and assess its contributions to model behavior.[1]

2 Related Work

Prior work has successfully adopted a "developmental approach" to investigate the emergence and nature of Transformer attention head specialization (Chen et al., 2024). Focusing on the evolution of self-attention over the course of pre-training, Olsson et al. (2022) identified "induction heads": attention heads that selectively track repetitions of a target token in the preceding context—as well as the tokens that follow—in order to produce correct next-token completions. The emergence of inductive attentional patterns coincides with marked improvements to in-context learning (Olsson et al., 2022), with subsequent work establishing a causal role for induction heads in successful completions within tasks that require tracking previously-occurring token sequences (Wang et al., 2023; Merullo et al., 2024; Zhang et al., 2025).

2.1 Motivation for Current Work

However, not all next-token predictions can be made (successfully) by simply tracking and copying previously occurring token motifs. One longstanding challenge in natural language processing (NLP) is lexical ambiguity, where a single word points to multiple related (polysemous) or unrelated (homonymous) meanings (Navigli, 2009;

Haber and Poesio, 2024). Next-token completions predicated on ambiguous words require iteratively teasing apart the cued and uncued meanings that are entangled in the ambiguous static token embedding (Grindrod, 2024).

Lexical ambiguity is pervasive, with some estimates placing the rate of English-language polysemous words at approximately 80% (Rodd et al., 2004). With the majority of words in the English language proving polysemous, and a smaller (but nontrivial) fraction considered homonymous (Dautriche, 2015), researchers have examined the extent to which Transformer-based LMs *can* tease apart entangled meanings in static embeddings (Haber and Poesio, 2020; García, 2021; Trott and Bergen, 2021; Garí Soler and Apidianaki, 2021; Haber and Poesio, 2021; Rivière et al., 2025). **To date, however, little work has leveraged ambiguity with the goal of characterizing specialization in attentional patterns during contextualization—or investigated the developmental processes underlying disambiguation performance in the final model**. Understanding how attention heads differentiate meanings across contexts provides a window into the mechanisms that support contextualization and semantic composition in Transformer LMs.

Investigating the mechanisms underlying model behavior has generally taken one of two prominent forms, involving either the modification of activation patterns and recording the resulting effects on LM logit outputs (Wang et al., 2023; Conmy et al., 2023); or the modification of learned static model parameters (Nelson et al., 2021; Olsson et al., 2022; Merullo et al., 2024; Chang and Bergen, 2025). The chief virtue of the latter is the ability to directly edit the information the LM has encoded about the statistics of a linguistic corpus. Of note, the query and key matrices of an attention head dictate which tokens its attention will be allocated to in the preceding context. Specialization in these matrices should heavily contribute to successful contextualization, and indeed, targeted ablations of induction heads does demonstrably impair in-context learning in a range of LMs of differing sizes and architectures (Olsson et al., 2022).

3 Phase 1: Identification

In Phase 1, we first aimed to characterize the *developmental trajectory* of disambiguation performance over the course of pretraining in the Pythia

---

[1]All code and data required to reproduce the analyses for each model at each checkpoint is available at the following Github repository: `https://github.com/seantrott/entangled_meanings`

suite of models (from $14M$ to $12B$). This included questions about the overall shape of this trajectory and the timing of specific "milestones" (i.e., *when* marked changes occurred).

Second, with a focus on $14M$ and $410M$ specifically, we attempted to identify specific attention heads that directed attention from the target ambiguous word (e.g., "lamb") to the disambiguating cue (e.g., "marinated")—and further, whether the developmental trajectories of these heads overlapped with changes in disambiguation performance. These *candidate disambiguation heads* could then be further stress-tested (Section 4) and ablated (Section 5) to assess both their generalizability and functionality.

3.1 Dataset

We used the RAW-C dataset (Trott and Bergen, 2021) as a behavioral probe. RAW-C contains relatedness judgments of ambiguous English words across 672 minimal sentence pairs. Each word appears in four sentences, with two sentences per sense, resulting in six sentence pairs per word (corresponding to four unique Different-Sense pairs, and two unique Same-Sense pairs). Each sentence pair is associated with a mean relatedness judgment, ranging from 1 (totally unrelated) to 5 (same sense).

The full dataset contains both ambiguous nouns and ambiguous verbs: nouns are disambiguated solely by a prenominal modifier (e.g., "marinated lamb"), while verbs are disambiguated solely by a post-verbal clause (e.g., "broke the promise"). Because Pythia models are auto-regressive and only attend to previous contexts, we used only the noun stimuli (comprising 504 sentence pairs total). Of these, 318 of the ambiguous word senses were classified as polysemous; 186 were classified as homonymous (see Trott and Bergen (2021)).

The sentence *pairs* were used to characterize changes in disambiguation performance, while *individual sentences* were used as inputs for the attention head analysis.

3.2 Language Models

To characterize overall disambiguation performance at select checkpoints, we selected the Pythia suite of language models (LMs), a series of English LMs ranging in size from $14M$ to $12B$ parameters, all trained on the same data (Biderman et al., 2023; van der Wal et al., 2025). All models

were run on 20 checkpoints[2], and $14M$ specifically was run on all available 154 checkpoints.

We selected LMs $14M$ and $410M$ for analysis of attention head behavior. Pythia-$14M$ was the smallest model and thus a suitable "model organism" for developing an experimental protocol. Pythia-$410M$, although far from the biggest model, performed nearly as well as $12B$, rendering it both *performant* and *tractable* as a subject of analysis. For $14M$, we assessed all 154 checkpoints; for $410M$, we assessed 20 checkpoints. We also assessed the attention head behavior of all models at their final step.

In general, the Pythia suite was well-suited for our research questions in several ways. First, a number of pre-training checkpoints are made publicly available, facilitating analysis of the developmental trajectory of our target behavior. Second, in addition to a "main" model release, each model (at each checkpoint) was trained on nine training runs, each initialized with a different random seed. This allows us to evaluate the *robustness* of our empirical results, while controlling for model architecture and training specifications. The primary analyses in this manuscript focus on the "main" release of each model. Analyses of random seeds are discussed in Appendix Section 7.1. All models were accessed through the HuggingFace *transformers* library (Wolf et al., 2020). Models were run either on a Mac laptop (M2, 2022) or on an NVIDIA DGX-H200.

3.3 Evaluating Disambiguation Performance

Following past work (Nair et al., 2020; Trott and Bergen, 2021; Rivière et al., 2025; Bojanowski et al., 2017; Schlechtweg et al., 2020), we calculated the *cosine distance* between the contextualized embeddings of the target word across each sentence pair (e.g., "She liked the marinated *lamb*" vs. "She liked the friendly *lamb*", **Figure 1a**). In cases where the target word was tokenized into multiple tokens, we computed the average embedding of those tokens. We repeated this procedure for each layer of each model, such that a given sentence pair was associated with L distance measures for a given model (where L is the number of layers in the model). Then, for each layer ℓ of each model, we regressed:

---

[2]Complete list of checkpoints: [0, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1000, 2000, 5000, 10000, 25000, 50000, 75000, 100000, 143000].

$$Relatedness \sim Distance_\ell \qquad (1)$$

The resulting R^2 measure reflects the proportion of variance in human relatedness judgments explained by the distribution of cosine distances obtained from layer ℓ. For the primary *developmental analyses* below, we selected the R^2 at each checkpoint from the layer that performed best at the *final step*. For example, if Layer 3 in $14M$ achieved the best R^2 at the final step, we analyzed the trajectory of performance at Layer 3 specifically. Note that we obtained qualitatively similar results using the best R^2 from each checkpoint, independent of layer. For instance, when considering all 154 checkpoints for which $14M$ was assessed, the correlation in R^2 trajectories across these measures was $r = 0.96$; moreover, when considering the subset of 20 checkpoints for which we assessed all models in the Pythia suite, the correlation between these measures ranged from $r = 0.98$ (for $14M$) to $r = 1$ (for 410M, 1B, 1.4B, 2.8B, and 6.9B).

3.4 Identifying Candidate Attention Heads

To isolate candidate attention heads contributing to disambiguation performance, we calculated a *disambiguation score* for each head at each model checkpoint. Specifically, each sentence was tokenized and presented to a given model in isolation. Then, for each head, we obtained the attention score from the target token (e.g., "lamb") to the disambiguating cue (e.g., "marinated"). If either the target token or the disambiguating cue consisted of multiple tokens, we averaged across scores for those tokens. We then performed a linear regression analysis to identify which attention heads showed changes in attention over pretraining that best predicted changes in disambiguation performance.

3.5 Results

3.5.1 Disambiguation Performance

We first quantified the disambiguation performance of each model (as measured by R^2) at the *final training step*. As depicted in **Figure 1b**, larger models performed better than smaller models: $14M$ achieved an R^2 of 0.15, while $6.9B$ (the best-performing model) achieved an R^2 of 0.598. Moreover, there was a significant relationship between model size (Log Number of Parameters) and disambiguation performance (R^2) [$\beta =$

$0.16, SE = 0.02, p < 0.001$]: each order of magnitude increase in model size was associated with a 0.16 improvement in task performance. Interestingly, however, there appeared to be diminishing returns to scale on this task: $410M$ achieved an R^2 of 0.524, which was 92% of the performance of $12B$ (the largest model tested)—despite having approximately 3% as many parameters. No model obtained human-level performance (0.64), though $6.9B$ came the closest. Notably, the maximum $R2$ was not always derived from the final layer's representations, as in the case of Pythia-$14M$, whose maximum $R2$ came from Layer 3 (**Figure2a**).

We then examined the developmental trajectory of disambiguation performance. We found clear evidence of discontinuities in R^2 throughout pretraining, with marked shifts in performance at step 1000 and step 2000—corresponding to 2.1B and 4.2B tokens seen, respectively (**Figure2b**). Focusing specifically on $14M$ (for which all 154 checkpoints were assessed), the model had achieved close to its final-step performance (0.15) by step 2000 (4.2B tokens). This trajectory is consistent with past work documenting sudden "phase shifts" in model performance (Hu et al., 2023; Chen et al., 2024); moreover, it is striking that close-to-maximal performance was achieved quite early in pre-training—approximately 1.4% of the way through (**Figure2b**).[3] A linear regression predicting R^2 from the Log Training Step found a significant relationship between the two variables [$\beta = 0.04, SE = 0.001, p < .001$], with Log Training Step explaining about 85% of the variance over time in R^2.

In contrast to $14M$, larger models appeared to continue improving well past these early steps—perhaps suggesting that larger model capacity was better able to reap the benefits of the training data (Kaplan et al., 2020; Hoffmann et al., 2022).

3.5.2 Identifying Candidate Attention Heads

We then identified candidate attention heads contributing to these changes in performance. Heads displayed considerable variance in the degree to which they attended to the disambiguating cue at the final step (**Figure 1c**).

In Pythia-$14M$, the strongest final-step atten-

---

[3] Interestingly, we observed a subsequent "dip" in performance using Layer 3 representations between steps 5000 and 50000. Using the R^2 from the *best-performing layer* at each step instead resulted in a somewhat smoother (though qualitatively similar) trend.
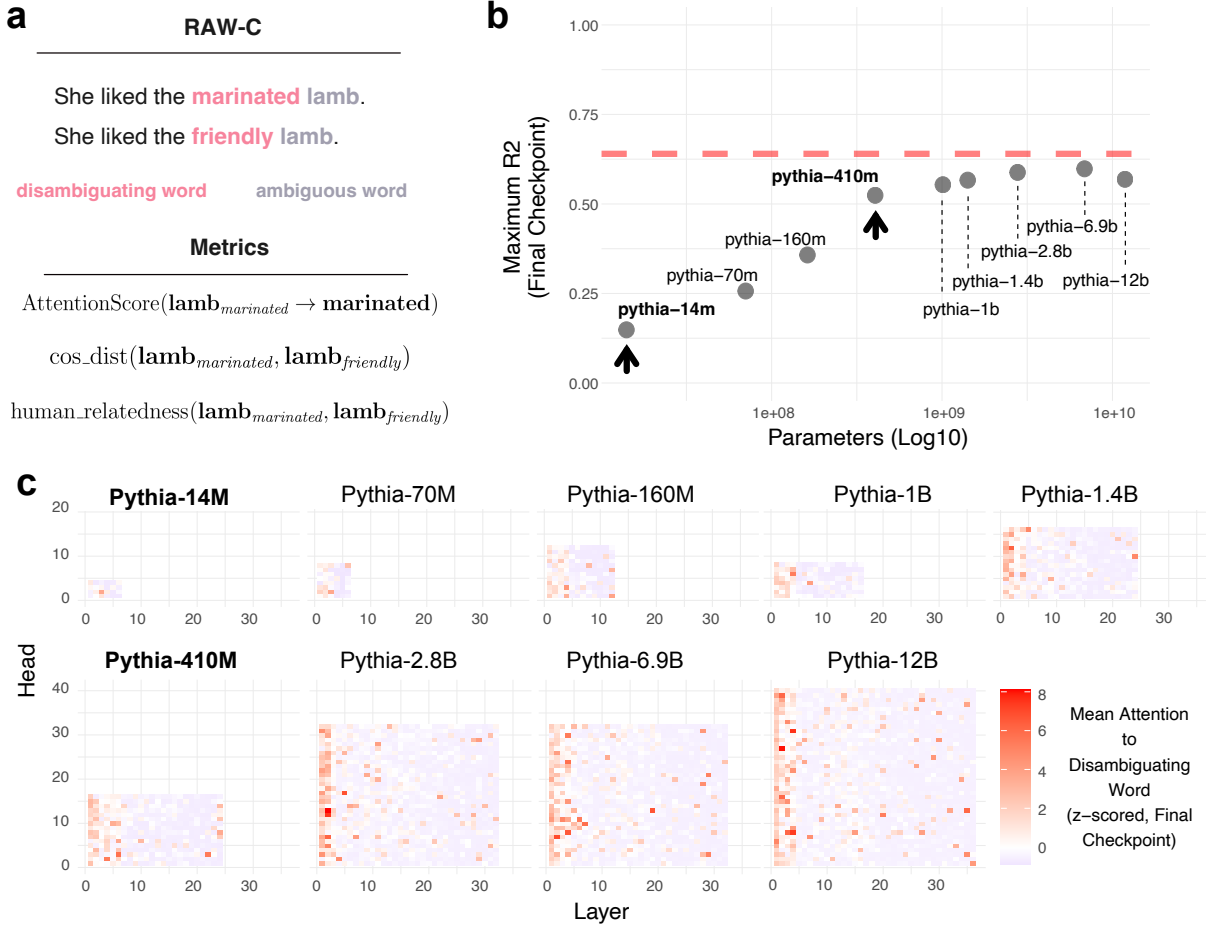
Figure 1: **Disambiguation performance at final checkpoint for Pythia language models (LMs).** **(a)** Sample RAW-C sentences, evoking different senses for the target ambiguous word (lamb) with single differing disambiguating word (marinated, friendly). We obtain: the AttentionScore from the ambiguous word to disambiguating word, per sentence; the cosine distance between contextualized representations for the target ambiguous word across sentences in a pair; and the publicly available human relatedness judgments between the target ambiguous word across sentences in a pair. **(b)** Max R^2 obtained from the final checkpoint of nine Pythia LMs, by number of parameters. Arrows mark models of interest, $-14M$ and $-410M$. Horizontal dashed line represents mean human interannotator agreement. **(c)** Each subpanel shows the head index by layer index for a given LM; warmer colors indicate higher z-scored mean attention scores to disambiguating words, for the final checkpoint of each LM.

tion to the disambiguating word came from Head 2 in Layer 3 (hereafter Head $(3, 2)$); this head's average attention score (0.76) was approximately 3.7 standard deviations over the mean attention to the disambiguating word in Pythia-$14M$ heads. Further, as seen in **Figure 2c**, two heads in layer 3—$(3, 1)$ and $(3, 2)$—exhibited patterns of attention that were (relatively) aligned temporally with changes in disambiguation performance (see Appendix **Figure 8** for a view of all layers). This was confirmed quantitatively by constructing a series of linear models regressing R^2 over training against the attention score α from a given head h over training (p-values were corrected for multiple comparisons using a false-discovery rate cor-

rection procedure). Strong positive relationships were obtained for Head $(3, 2)$ and Head $(3, 1)$.

Unsurprisingly, $410M$ had a larger number of *candidate heads*, i.e., those with an unusually high mean attention to the disambiguating word. As evident in **Figure 1c** and **Figure 2d**, these heads appeared in a variety of layers. A majority appeared in relatively *early* layers (e.g., layers $1-4$), with some appearing in the final layers (i.e., layers $22 - 24$; see Appendix **Figure 9** for a view of all layers). As with $14M$, we constructed a series of linear models regressing R^2 over training against the attention score α from a given head h over training (again correcting for multiple comparisons). The strongest positive coefficients were
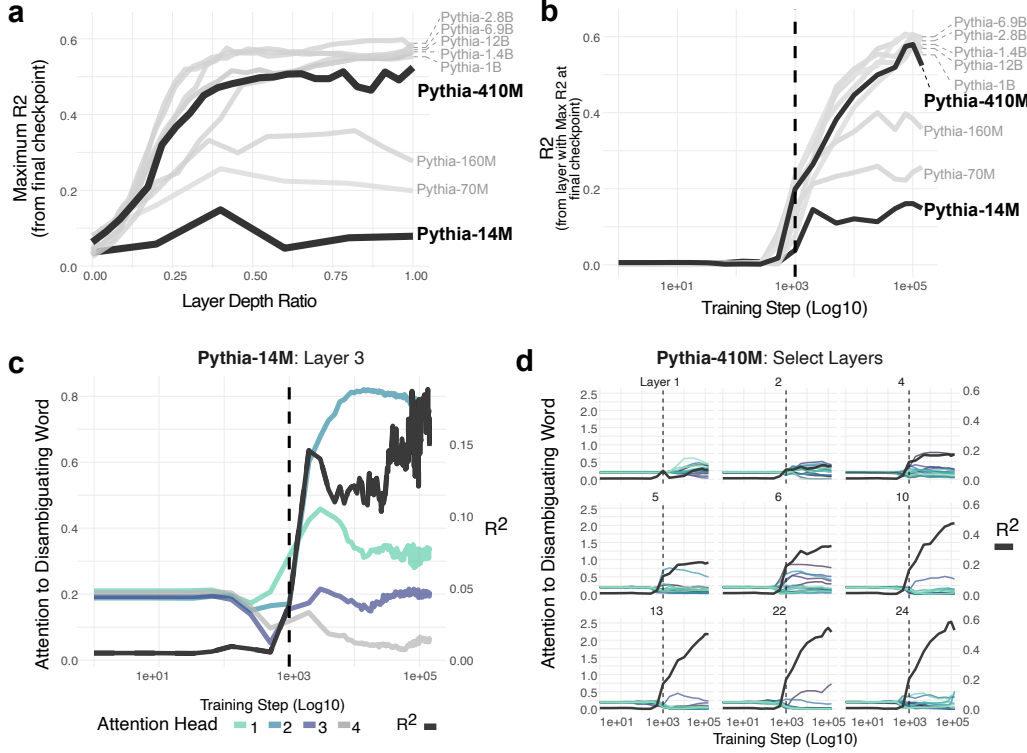
Figure 2: **Identifying candidate attention heads. (a)** Maximum R^2 from the final step, by layer depth (current layer/max number of layers), for nine Pythia LMs. **(b)** "Developmental" view of R^2, obtained from each training step for nine Pythias. Depicted R^2s are from the layer with the max R^2 at the final checkpoint (e.g. for $14M$, R^2 is from Layer 3; for $410M$, R^2 is from Layer 24). **(c)** "Developmental" view of attention to disambiguating word, for all head indices in $14M$'s Layer 3. Superimposed is the R^2 from Layer 3. Attention scores for Heads $(3, 1)$ and $(3, 2)$ covary with disambiguation performance. **(d)** Same as in **c**, but for select layers in $410M$. Layers were selected if they contained at least one head whose attention scores rose during training. Superimposed R^2 curves were drawn from each Layer depicted. Vertical dashed line in sub-panels **b-d** mark training step 1000, corresponding to $2.1M$ tokens seen cumulatively over training.

generally obtained in Layer 1, though we also observed robust relationships in later layers.

4 Phase 2: Stress Testing to Define Attentional Scope

In Phase 1, we found marked *phase shifts* in disambiguation performance that coincided with changes in the behavior of select attention heads, which systematically directed attention from the target word (e.g., "lamb") to the disambiguating cue (e.g., "marinated"). After *identifying* these candidate attention heads, we sought to assess the robustness and selectivity of their attention patterns for disambiguating words. RAW-C sentence stimuli always contain the key disambiguating modifier immediately preceding the target ambiguous word. Our goal was to evaluate the extent to which this behavior was robust to a range of stimulus perturbations (see Section 4.1), as well to compare the behavior to simpler, "lower-level"

functions such as "1-back attention".

4.1 Stimuli Manipulations

We devised three "control" tasks to *stress-test* (Shapira et al., 2024; Naik et al., 2018) the putative disambiguation circuits.

In the **1-back analysis**, we measured the average attention directed by each head from each token in a sentence to the immediately preceding token (Clark et al., 2019). We then performed a *subtraction analysis*, which allowed us to determine whether attention to the immediately preceding disambiguating cue was particularly strong, i.e., over and above 1-back tokens in general.

With **Positional modification**, we modified the *position* of the disambiguating cue relative to the target noun by adding a semantically bleached phrase between the disambiguating cue and the target noun (e.g., "tense/gaseous *kind of* atmosphere"). Then, as in Section 3), we calculated

the attention for each head from the target noun to the disambiguating cue. This allowed us to determine whether the same heads still attended to the nominal modifier even when it was separated from the target noun ($N = 310$ sentences).

Finally, with **Part-of-speech modification**, we modified the RAW-C sentences (Trott and Bergen, 2021) such that ambiguous nouns were now disambiguated by a *verb* (e.g., "He *polished/filed* the case"); we also included sentences with ambiguous verbs, which were disambiguated by a noun phrase in the subject position (e.g., "The *glass/promise* was broken"). Then, we calculated the attention for each head from the target word to the disambiguating cue. This allowed us to determine whether the same attention heads directed attention to disambiguating cues regardless of their part-of-speech ($N = 360$ sentences).

The analyses described above were carried out across all checkpoints of the Pythia-$14M$ model, as well as the 20 checkpoints of the Pythia-$410M$ model investigated in Section 3.

4.2 Results

4.2.1 1-back attention

Multiple heads in both models emerged as candidates for "1-back heads", some of which overlapped with the candidate heads identified in Phase 1. For example, in $14M$, Head $(3, 2)$ underwent a developmental trajectory that strikingly resembled the patterns described in Section 3, i.e., changes in attention emerged at roughly 2000 steps. We then asked whether the heads identified in Section 3 showed particularly strong attention to the disambiguating word, *above and beyond* their 1-back attention more generally. For each sentence and each head at each checkpoint, we subtracted the 1-back attention from that head's attention to the disambiguating word. Finally, we performed a one-tailed paired t-test at each checkpoint for each head (correcting for multiple comparisons using a false-discovery rate (FDR) procedure across all checkpoints and heads). This procedure revealed several heads with significantly higher attention to disambiguating words.

In $14M$, these heads included $(3, 1)$ and Head $(3, 2)$ (see **Figure 3a**). In $410M$, 11 heads survived the subtraction analysis, including $(1, 11), (1, 14), (6, 5),$ and $(23, 8)$. These heads may be specialized for attention to prenominal modifiers above and beyond 1-back attention.

4.2.2 Positional modification

We then asked whether attentional patterns were robust to the relative position of the prenominal modifier (e.g., "friendly sort of lamb").

In $14M$, only a single attention head—Head $(3, 1)$—appeared to be robust to this modification, showing a very similar developmental trajectory as it did for the original stimuli (**Figure 3b, *top***). In contrast, Head $(3, 2)$ (previously identified as attending strongly to the prenominal modifier) directed attention primarily to the token immediately preceding the target (e.g., "friendly sort **of** lamb"; **Figure 3b, *bottom***).[4]

In $410M$, a number of heads consistently directed attention to the disambiguating cue at the final step—and also showed strikingly similar developmental trajectories as they did for the original stimuli. This included a number of previously identified from Layer 1, such as $(1, 7)$ and $(1, 14)$, but also included heads from later layers, such as $(6, 4)$ and $(24, 8)$.

4.2.3 Part-of-speech modification

Finally, we asked which attention heads, if any, attended to disambiguating words regardless of their part-of-speech. In $14M$, we found virtually orthogonal sets of attention heads (mostly in layer 6) that directed attention to disambiguating *verbs* (e.g., "He **swung** the *bat*") vs. disambiguating *nouns* (e.g., "The **pond** was *drained*") (**Figure 3c**). Notably, neither set of heads overlapped substantively with the original heads identified in Section 3, with the possible exception of Head $(3, 1)$ attending to disambiguating nouns at approximately step 1000.

In $410M$, on the other hand, there were select heads that consistently directed attention to the disambiguating cue, regardless of its part-of-speech. Although these heads appeared to show some preference for certain parts-of-speech (e.g., nouns over verbs), the *minimum* attention directed to disambiguating cues was systematically higher than other heads independent of the cue's part-of-speech. Candidate heads identified in this process

---

[4]Interestingly, this stimulus modification brought to the fore multiple heads in Layers 1, 2, and 4 that attended substantially to the last token of the inserted string, although they were previously inattentive to the disambiguating word when it immediately preceded the target ambiguous word in original stimuli. Given that this token was typically a high-frequency preposition (e.g., "of"), it is possible these heads are sensitive to lexical properties, e.g., unigram frequency.
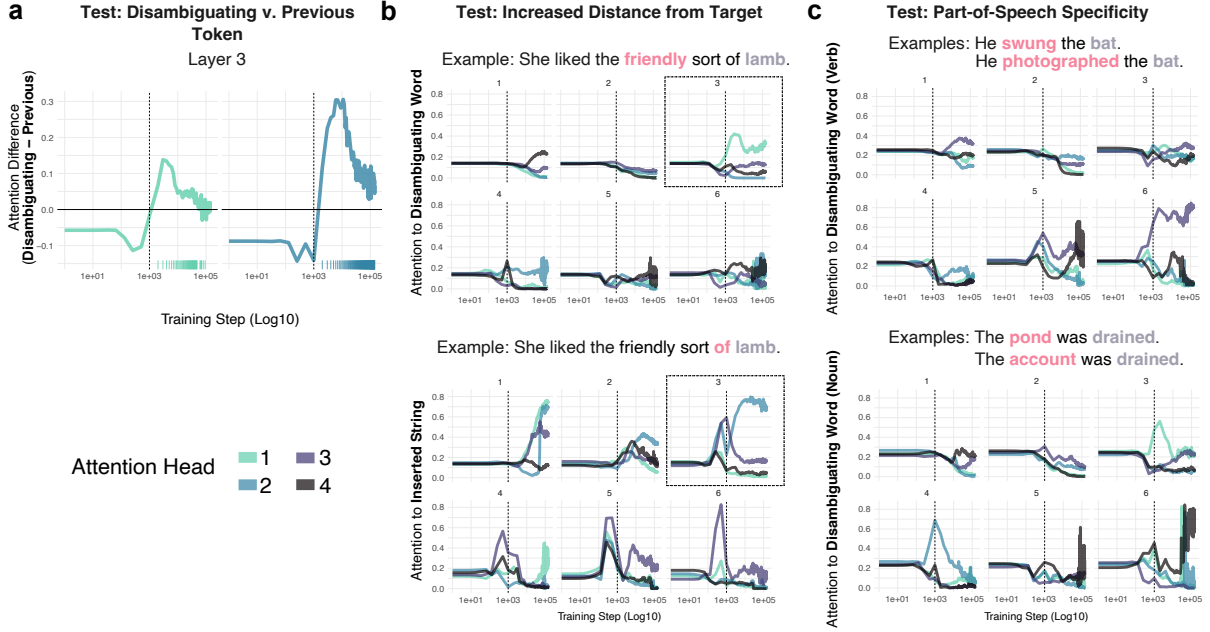
Figure 3: **Stress testing Pythia-14M candidate heads' attention to disambiguating word.** Attention Head color coding scheme applies to all panels. **(a)** Layer 3 difference in average attention scores for disambiguating word against that of all 1-back tokens, over pre-training. Ticks indicate training steps with significant difference in attention scores ($p < 0.05$), adjusted for multiple comparisons. Only Head $(3,2)$ remains significant at the final training checkpoint. **(b)** Each sub-panel corresponds to a different layer. (*top*) Attention to disambiguating word when it is separated from the target ambiguous word via inserted string. (*bottom*) Attention to last token of inserted string. The square surrounding Layer 3 highlights the attentional robustness of at least one of the two candidate heads. **(c)** Attention to disambiguating word when its part-of-speech changes to a verb (*top*) or a noun (*bottom*).

included $(1, 7), (1, 14), (4, 7)$, and $(6, 10)$.[5]

4.2.4 Constructing a "Disambiguation Index" for $410M$ Heads

To facilitate interpretation for stress-test results in a model containing a much larger number of attention heads, we constructed a composite score for each head in $410M$, reflecting a given head's robustness to stimulus perturbations and its attentional covariance with performance over pre-training. Specifically, this score is constructed by taking the average of five z-scored variables: the coefficient relating changes in attention to changes in R^2; the average final-step attention to disambiguating nouns; the average final-step attention to disambiguating verbs; the t-statistic resulting from the 1-back subtraction analysis; and the average final-step attention to disambiguating modifiers in the "sort of" analysis. A higher score on this index (e.g., > 2) indicated the extent to which

a given head consistently (i.e., *robustly*) attended to disambiguating cues across various stimulus perturbations and the extent to which this covaried with changes in disambiguation performance. This composite score is depicted for each head in **Figure 4**. Most heads received low scores, while only a handful received very high scores. The six heads with the largest score on this index included $(1, 14), (4, 7), (1, 3), (2, 3), (6, 10)$, and $(1, 13)$.

4.2.5 Summary of Stress-Testing Results

Our primary goal in Phase 2 was *stress-testing* the heads identified in Phase 1: namely, how *specialized* are their putative functions, and how *robust* is their behavior to stimulus perturbations?

The results for $14M$ were mixed: while select candidate heads may be specialized for more than relatively simple mechanisms like 1-back attention, their behavior did not generalize across sentence frames or parts of speech, suggesting that they are not "generalized disambiguation heads" per se—rather, they may *participate* in disambiguation for nouns specifically. In contrast, $410M$ contained select heads whose attentional patterns were surprisingly robust to stress-testing,
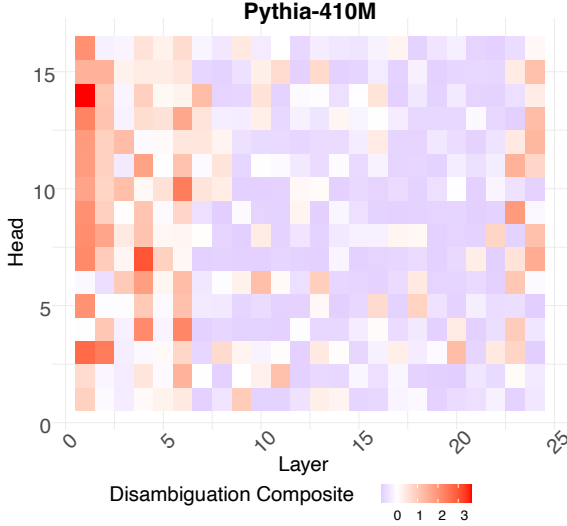
---

[5]There were also select heads with a strong *part-of-speech bias*, as measured by the difference in average attention directed towards disambiguating nouns vs. verbs. Noun-preferring heads included $(6, 5)$ and $(23, 11)$, and verb-preferring heads included $(4, 4)$ and $(11, 2)$.

Pythia-410M

Disambiguation Composite
0 1 2 3

Figure 4: **Stress testing Pythia-**$410M$**.** Warmer colors indicate larger Disambiguation Composite scores for Pythia-410M heads and layers. Larger scores reflect greater attentional robustness to the disambiguating word despite stimulus perturbations, and greater degree of attention covariance with disambiguation performance throughout pre-training.

as revealed by particularly high scores on the *disambiguation index* (Section 4.2.4; **Figure 4**), suggesting their *functional scope* was considerably broader—or at least more abstract—than the heads identified in $14M$.

Crucially, however, Phase 2 investigated only the *attentional patterns* of these heads; in Phase 3, we examine their putative *causal contributions* to disambiguation.

5 Phase 3: Causal Analysis

To determine the extent to which candidate heads' attention patterns are *necessary* for Pythia-$14M$'s and Pythia-$410m$'s disambiguation performance at different pre-training stages, we carried out a series of targeted ablations at each model checkpoint. We then asked how ablating the target heads identified in Phase 1 and Phase 2 affected the model's performance on the disambiguation task (relative to ablating a set of "control" heads).

5.1 QK Matrix Manipulations

Attention heads in the Transformer architecture consist of matrices of learnable parameters. A given head's Query ($\mathbf{W}_Q$) and Key ($\mathbf{W}_K$) matrices specifically direct attention to select tokens in the input sequence. To directly intervene on our candidate head's attention patterns, we manipulated

the values of these two matrices.

In $14M$, we selected Head $(3, 1)$, Head $(3, 2)$, and the combination of both heads. To intervene specifically on our candidate attention heads, Heads $(3, 1)$ and $(3, 2)$, we directly modified the weights in their query-key (QK) matrices. In the **Zero-Ablation** condition, we set the target head's $\mathbf{W}_Q$ and $\mathbf{W}_K$ to equally-sized matrices of zeros. In the **Step1-Copy-Ablation** condition, we set the target head's $\mathbf{W}_Q$ and $\mathbf{W}_K$ matrices to the parameter values they held at the first pre-training checkpoint (i.e., step 1); this latter ablation ensured that the target head(s) is(are) still participating in transforming embeddings, but with suboptimal parameters. We implemented both ablation types for Heads $(3, 1)$ and $(3, 2)$ on their own as well as concurrently, at each training step beginning with step 1. To control for the possibility that *any* modifications at this layer would result in a reduction in performance, we implemented a set of *baseline* conditions, in which we ablated two other heads from the same layer: $(3, 3)$ and $(3, 4)$—again, either on their own or together.

We followed the same set of procedures for select heads in $410M$. In this case, we selected the six heads with the top *disambiguation index*, as identified in Section 4.2.4; heads were shown to be relatively robust to stimulus perturbations.[6] These heads included: $(1, 14), (4, 7), (1, 3), (2, 3), (6, 10), (1, 13)$. As baseline heads, we selected a *matched control* for each target head from the same layer but with a lower disambiguation index. This ensured that the controls were matched for layer (as ablating heads in distinct layers could affect measured R^2 using representations from, say, the final layer).

5.2 Procedure

We followed the same procedure described in Phase 1 (Section 3) for measuring the disambiguation performance (i.e., R^2) of each ablated model at each checkpoint. We then calculated the difference between each ablated model's R^2 and the intact model's R^2 at an equivalent checkpoint (ΔR^2), as well as the ratio between the ablated and intact models' R^2 values ("fraction of R^2_{Intact}").

---

[6]Note that in theory, the heads with the most robust behavior need not be the heads most involved in a *particular* stimulus configuration, i.e., the one assessed in the original RAW-C stimuli; nevertheless, we selected these heads because they were the most likely candidates for serving the purpose of "disambiguation" across a number of contexts.
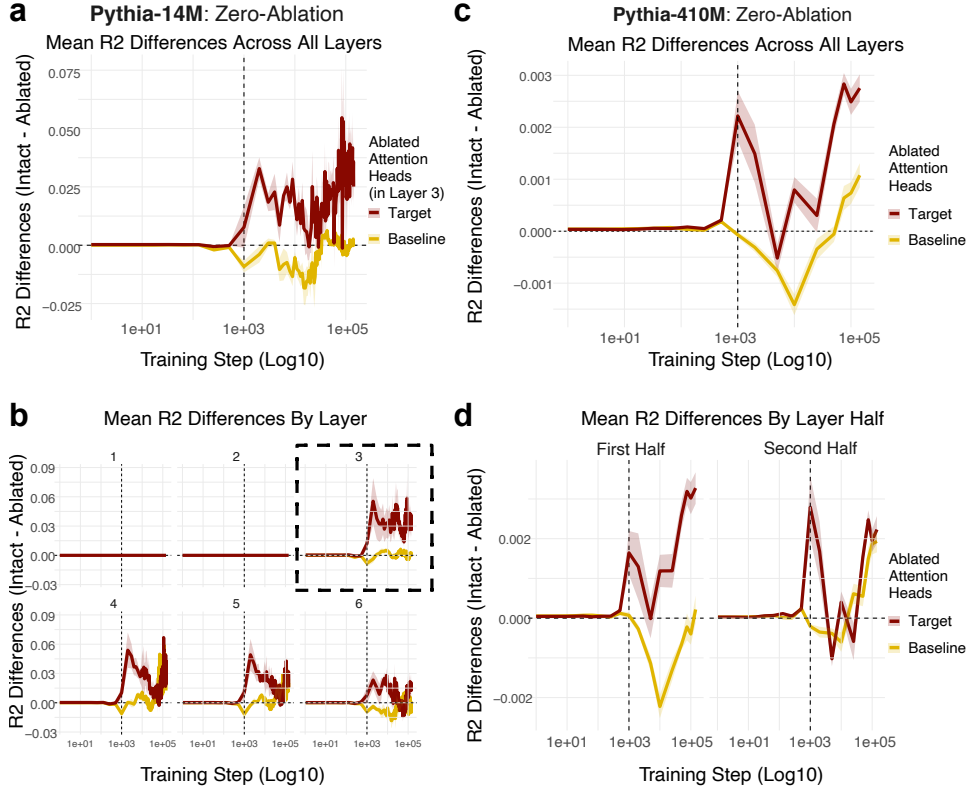
Figure 5: **Target head ablations decrease disambiguation performance relative to intact models.** **(a)** Mean difference in R^2 across all layers and all combinations of Pythia-$14M$'s target head ablations, for all training steps. Values > 0 indicate that the intact model's R^2 exceeded that of the ablated model's R^2, reflecting causal effect of ablation. Target manipulations refer to zero-ablations of previously-identified Layer 3 heads. Baseline manipulations refer to zero-ablations of Layer 3 heads whose attention to disambiguating words fail to increase with disambiguation performance. **(b)** Same as in **a**, but parcelled out by layer, to illustrate localization of ablation effects, which remain robust throughout training in Layer 3. Dashed square marks the only layer (Layer 3) to suffer head ablations. **(c)** Same as in **a**, but for Pythia-$410M$. Target head ablations causally decrease model performance. **(d)** Same as in **c**, but parcelled out by early versus late layers, to illustrate the selectivity of target-head ablation to earlier layer representations. By the end of training, the effects of target-head ablations are indistinguishable from those of baseline-head ablations.

5.3 Results

As depicted in **Figure 5a&c**, ablating the target heads resulted in impaired performance in both $14M$ and $410M$ on average across layers. This effect was systematically larger when ablating target heads relative to ablating the baseline heads, and the *developmental trajectory* of performance reductions was reflective of the trajectories reported in Phase 1 and Phase 2: i.e., clear divergences first emerge at approximately 1000 steps.

Notably, the effect of ablating target heads depended to some extent on which layers were used to extract contextualized representations. In $14M$, the strongest effects were observed in Layer 3 (**Figure 5b**), which was also the maximally-performing layer (as identified in **Figure 2a**); in $410M$, the impact of target-head ablation was

most pronounced in the first twelve layers, and was not distinguishable from baseline-head ablations in the final layers (**Figure 5d**) other than training step 1000. Similarly, ablating the different target heads in $14M$ produced different effects at this model's different layers and checkpoints: ablating Head $(3, 1)$ impaired performance earlier in pretraining, while ablating Head $(3, 2)$ hurt performance in later stages. (Ablating *both* resulted in more consistent reductions in performance throughout pre-training; see Appendix **Figure 11**.) These differences mirror the different developmental trajectories of the two heads identified in Section 3.5 (see also **Figure 2c** and Appendix **Figure 8**). This question of whether different heads matter to different degrees at different timepoints is explored in more detail in the Gen-

eral Discussion (Section 6.1).

To quantify the impact of each ablation type, we built a series of linear regression models for both $14M$ and $410M$ predicting ΔR^2 or fraction of R^2_{Intact} as a function of CONDITION (Baseline vs. Target) and LOG TRAINING STEP. We conducted this analysis both for representations from the optimal layer (Layer 3 in $14M$, Layer 24 in $410M$) and also averaging performance reductions across layers (as in **Figure 5a, b**).[7]

In $14M$, a significant coefficient was found for CONDITION for each ablation type and each dependent variable, suggesting that variability in both indices was attributable to ablating the target heads. Specifically, zero-ablating the Target heads in $14M$ was associated with a larger ΔR^2 $[\beta = 0.03, SE = 0.001, p < .001]$ and smaller fraction of R^2_{Intact} $[\beta = -0.21, SE = 0.007, p < .001]$ (**Figure 5a,c**). The step1-copy ablations were also associated with a larger ΔR^2 $[\beta = 0.05, SE = 0.002, p < .001]$ and smaller fraction of R^2_{Intact} $[\beta = -0.33, SE = 0.01, p < .001]$ (Appendix **Figure 10a,c**). Qualitatively identical results were found when considering average performance across layers. Put another way, ablating the target heads resulted in measurably larger *reductions* in disambiguation performance than did ablating the baseline heads.

In $410M$, we found a systematic (though small) effect of ablating Target heads on *average* performance across layers, though not on the best-performing layer (see also Figure **5d**). When considering average performance across layers, zero ablations were associated with a larger ΔR^2 $[\beta = 0.001, SE = 0.0002, p = 0.001]$; the effect on R^2_{Intact} was only marginally significant $[\beta = -0.007, SE = 0.004, p = 0.05]$. The step1-copy ablations were also associated with a larger ΔR^2 $[\beta = 0.001, SE = 0.0002, p = 0.003]$ and a smaller fraction of R^2_{Intact} $[\beta = -0.004, SE = 0.001, p < .001]$. These effects were both very small in absolute terms and also absent in the best-performing layer (layer 24). Together, this suggests that the function of each individual head may overlap, consistent with past work on redundancy in attention heads (Michel et al., 2019); and further, that the impact of ablating target vs. baseline heads in earlier layers was indistinguishable by the

---

[7]Qualitatively similar results were obtained using the raw training step number. Further, allowing an interaction between CONDITION and LOG TRAINING STEP suggested that the effect of the Target ablations increased for later steps.

final layers.

6 General Discussion

In this work, we report a critical *inflection point* in disambiguation performance of both Pythia-$14M$ and Pythia-$410M$ over the course of pre-training, which coincides with increased attention to the disambiguating word from a subset of attention heads (Section 3); changes in the behavior of certain attention heads account for as much as 77% of the variance in changes in disambiguation performance throughout pre-training. In both models, we observed heads that attended preferentially to the disambiguating cue above and beyond a general preference for 1-back tokens (Section 4.2.1). Further, in $410M$, we identified select heads that were also robust to positional and part-of-speech manipulations (Section 4.2.4), though $14M$ contained no such heads (Section 4). Finally, the ablation analyses point to a clear (and relatively large) *causal role* in disambiguation for the candidate heads in $14M$. The effect of ablating individual heads in $410M$ was weaker (and more pronounced in early layers), though this is less surprising given that $410M$ had 16x as many heads as $14M$ overall (Section 5).

6.1 On Development and "Passing the Baton"

These results join a growing body of work adopting a "developmental" perspective on the mechanisms underlying LM behaviors (Chen et al., 2024; Olsson et al., 2022; van der Wal et al., 2025). As in the study of humans (De Barbaro et al., 2013), an ontogenetic approach offers unique benefits: in this case, it enables us to identify inflection points in the emergence of capabilities, yielding insights into prerequisite behaviors the LM must achieve to display the capability, and allowing us to link the capability to fine-grained mechanisms.

Here, the results indicate that the representations and mechanisms crucial to contextualizing ambiguous nouns with their modifiers develop relatively early in pre-training. Further, the checkpoints identified align to some degree with previous work on the development of other mechanisms across model scales and random initializations (Tigges et al., 2024; van der Wal et al., 2025; Olsson et al., 2022; Trott, 2025). Future work could investigate the precise changes in weight

matrices that subserve these developments and potentially identify the biases in initial parameterization that lead to different patterns of head "specialization" across random seeds (see Appendix **Figure 6**). This work could also investigate the role of other model components in disambiguation, such as the value matrices or the residual stream.

We also identified intriguing qualitative changes in attention head behavior (and possible interactions) over the course of pre-training. For instance, some heads played a stronger role earlier in pre-training, then appeared to lose influence as training proceeded. We preliminarily term this phenomenon "passing the baton", as other heads (in some cases, heads in the same layer) come to encode the relevant information over the course of pre-training. Similarly, ablations of heads in early layers affected the quality of downstream representations to different degrees at different points in pre-training, potentially reflecting the emergence of alternative routes to transmitting the relevant information. These phenomena are themselves ripe for future investigation, and are only discoverable by adopting a developmental perspective.

6.2 On Assessing Functional Scope

Identifying the "function" of a circuit (biological or artificial) is notoriously challenging (Haklay et al., 2025). Here, the results of our stress-testing point to qualitative differences in the robustness (and redundancy) of attention head behaviors across $14M$ vs. $410M$.

Specifically, target heads in $410M$ were considerably more *robust* to manipulations of position or part-of-speech than heads in $14M$. One speculative explanation for this difference is that the larger parameter count of $410M$ grants more opportunities for *functional abstraction*. While heads in $14M$ perform operations that ultimately subserve the high-level task we call "disambiguation", they may in fact be specialized for lower-level functions—in contrast, heads in $410M$ may be better candidates for generalized disambiguation heads. Notably, this conclusion depends on rigorous stress-testing, i.e., assessing the robustness of a model component's behavior to different perturbations.

At the same time, ablating heads in $14M$ led to larger reductions in R^2 (about 30x) than in $410M$. Again, one explanation for this is the difference in model size: $410M$ has 16x as many heads

as $14M$, raising the possibility of some overlap across those heads' functions. From this perspective, the fact that ablating a single head (out of 384) results in any reduction in performance could be seen as surprising. Indeed, there is considerable evidence for *redundancy* in attention head functions in transformer language models (Michel et al., 2019; Bian et al., 2021; He et al., 2024; Kovaleva et al., 2019).

We also note that while we did assess the behavioral robustness of attention heads in both models (see Section 4), we only assessed their functional involvement in disambiguating the original RAW-C stimuli (i.e., the sentence pairs for which we actually had relatedness judgments). Future work could build on the RAW-C dataset and collect human judgments for the modified stimuli as well. Another open question is whether the heads identified in each model are actually *selective* for ambiguous target words in particular, or whether they participate in contextualizing any target word (ambiguous or otherwise). Future research could ask whether the behavior of these heads is robust to the status of the target word, e.g., whether it is ambiguous ("marinated *lamb*") or unambiguous ("marinated *pork*").

6.3 The Search for Generalization

The degree to which we are licensed to draw inferences about a wider class of language models from any one pattern of results is an important question to consider.

Following Trott (2025) and Fehlauer et al. (2025), the current work does *not* take for granted that the disambiguation performance and/or the corresponding attention specialization patterns obtained in a single model instance (e.g. the default Pythia-$14M$) should necessarily replicate even within other instances of the same model architecture, pretrained on an identical sequence of token batches, for the same number of pretraining steps, but initialized with a different set of randomly selected parameters.

Our work engages with these generalizability considerations directly by (a) replicating experiments with the available Pythia-$14M$'s random seeds (**Figure 6; see Appendix Section 7.1**), and (b) replicating the full set of interpretability analyses in a model an order of magnitude larger, Pythia-$410M$, after having assessed disambiguation performance across the entirety of the Pythia

suite (from $14M$ to $12B$ parameters). Future work could pursue additional questions of generalizability by carrying out the full set of interpretability analyses on other large LMs in the Pythia suite beyond $14M$ and $410M$, or expand this work to other model families with publicly available checkpoints, e.g., OLMo 2 (OLMo et al., 2025). Moreover, questions regarding the generalizability of attentional specialization patterns in the context of *different languages* could leverage the growing body of ambiguity datasets in languages other than English (Baldissin et al., 2022; Garí Soler and Apidianaki, 2021; Schlechtweg et al., 2024; Abuín and Garcia, 2025; Rivière et al., 2025).

Our attempts to explore generalization across random seeds of Pythia-$14M$ revealed an intriguing pattern of results: while we observed remarkable robustness in terms of *developmental* trajectories of *disambiguation performance* across seeds, there was marked (and systematic) variance in which *layer* hosted the key attention heads (see Appendix 7.1). This suggests that any given capability may be subserved by heterogenous mechanistic solutions—even in the same LM architecture, trained on the same sequence of tokens and following the same training objectives and hyperparameter specifications. *Mechanistic heterogeneity* underlying what is otherwise "virtually indistinguishable network activity" (Prinz et al., 2004) is a hallmark of even relatively "simple" biological neural networks. Observing this phenomenon in a small LM suggests similar challenges to the field of mechanistic interpretability at large. Future research could work to identify the *axes of correspondence* for behaviors that are "conserved" across model instances (Tigges et al., 2024; Trott, 2025; Fehlauer et al., 2025), with the goal of linking properties of individual model instances (e.g., size, training data, initial parameters) to the mechanisms and behaviors they exhibit.

Acknowledgments

References

Marta Vázquez Abuín and Marcos Garcia. 2025. Assessing lexical ambiguity resolution in language models with new wic datasets in galician and spanish. *Procesamiento del Lenguaje Natural*, 74:305–319.

Gioia Baldissin, Dominik Schlechtweg, and Sabine Schulte im Walde. 2022. DiaWUG: A dataset for diatopic lexical semantic variation in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2601–2609, Marseille, France. European Language Resources Association.

Yuchen Bian, Jiaji Huang, Xingyu Cai, Jiahong Yuan, and Kenneth Church. 2021. On attention redundancy: A comprehensive study. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 930–945, Online. Association for Computational Linguistics.

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, Lintana Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Tyler A. Chang and Benjamin K. Bergen. 2025. Bigram subnetworks: Mapping to next tokens in transformer language models. *arXiv preprint arXiv:2504.15471*.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in

mlms. In *The Twelfth International Conference on Learning Representations*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. 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.

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

Isabelle Dautriche. 2015. *Weaving an ambiguous lexicon*. Ph.D. thesis, Université Sorbonne Paris Cité.

Kaya De Barbaro, Christine M. Johnson, and Gedeon O. Deák. 2013. Twelve-month "social revolution"emerges from mother-infant sensorimotor coordination: A longitudinal investigation. *Human Development*, 56(4):223–248.

Finlay Fehlauer, Kyle Mahowald, and Tiago Pimentel. 2025. Convergence and divergence of language models under different random seeds. *arXiv preprint arXiv:2509.26643*.

Marcos García. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162

of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

Jumbly Grindrod. 2024. Transformers, contextualism, and polysemy. *arXiv preprint arXiv:2404.09577*.

Janosch Haber and Massimo Poesio. 2020. Word sense distance in human similarity judgements and contextualised word embeddings. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 128–145.

Janosch Haber and Massimo Poesio. 2021. Patterns of lexical ambiguity in contextualised language models. *arXiv preprint arXiv:2109.13032*.

Janosch Haber and Massimo Poesio. 2024. Polysemy—evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, and Yonatan Belinkov. 2025. Position-aware automatic circuit discovery. *arXiv preprint arXiv:2502.04577*.

Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. 2024. What matters in transformers? not all attention is needed. *arXiv preprint arXiv:2406.15786*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Michael Y Hu, Angelica Chen, Naomi Saphra, and Kyunghyun Cho. 2023. Latent state models of training dynamics. *arXiv preprint arXiv:2308.09543*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and

Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Talking heads: Understanding inter-layer communication in transformer language models. *Advances in Neural Information Processing Systems*, 37:61372–61418.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2025. The quest for the right mediator: Surveying mechanistic interpretability through the lens of causal mediation analysis.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Elhage Nelson, Nanda Neel, Olsson Catherine, Henighan Tom, Joseph Nicholas, Mann Ben, Askell Amanda, Bai Yuntao, Chen Anna, Conerly Tom, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

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. 2025. 2 olmo 2 furious.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

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

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

Pamela D. Rivière, Anne L. Beatty-Martínez, and Sean Trott. 2025. Evaluating contextualized representations of (Spanish) ambiguous words:

A new lexical resource and empirical analysis. In *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.

Jennifer M. Rodd, M. Gareth Gaskell, and William D. Marslen-Wilson. 2004. Modelling the effects of semantic ambiguity in word recognition. *Cognitive science*, 28(1):89–104.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23.

Dominik Schlechtweg, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte Im Walde. 2024. The DURel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 137–149, St. Julians, Malta. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

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

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

Sean Trott. 2025. Toward a theory of generalizability in LLM mechanistic interpretability research. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Sean Trott and Benjamin Bergen. 2021. RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7077–7087.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Oskar van der Wal, Pietro Lesci, Max Muller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. 2025. Polypythias: Stability and outliers across fifty language model pre-training runs. *arXiv preprint arXiv:2503.09543*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*.

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. 2020. 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. Association for Computational Linguistics.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.

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

7 Appendix

7.1 Generalizability of Attention Patterns Across Pythia-$14M$ Random Seeds

The analyses above were primarily conducted on two model instances: Pythia-$14M$ and Pythia-$410M$. While the results are encouraging in that both model instances displayed similar developmental trajectories in their disambiguation performance and onset of candidate disambiguation heads, it remains an open question whether these results would generalize *within* a given model architecture but *across random seeds* (Zhao et al., 2025; Trott, 2025; van der Wal et al., 2025). Here, we ask whether the same model (Pythia-$14M$) trained on the same data but with different random random seeds displays regularities across certain *axes of correspondence* (Trott, 2025), e.g., whether candidate attention heads arise at similar *timepoints* or *locations* across models.

7.1.1 Methods

We used the nine random seeds released for Pythia-$14M$ (van der Wal et al., 2025). Each model was assessed at all 154 training checkpoints and accessed through the HuggingFace *transformers* library (Wolf et al., 2020). For tracking changes in disambiguation performance and attention head patterns, we implemented exactly the same procedure used in Phase 1 (see Section 3) for each of the nine random seeds at each training checkpoint. We also introduced a novel behavioral task to assess sensitivity to modifier-noun constructions, which we report on in Appendix Section 7.2.

7.1.2 Results

The developmental trajectories of both disambiguation performance and attention to the disambiguating word were strikingly similar across random seeds. Despite variance in final step performance, each random seed showed sharp changes in R^2 at similar checkpoints (**Figure 6a**). Similarly, regardless of *where* the "maximally attentive" head was located, the largest changes in attention again occurred between steps 1000 and 2000 (**Figure 6c**). Finally, the random seeds exhibited *bimodality* in where these attention heads developed: in roughly half the seeds, the head with maximal attention to the disambiguating word emerged in Layer 3, while in the other half, it emerged in Layer 4 (**Figure 6b**). This is consistent with other work revealing apparent "bimodality" across random initializations (Zhao et al., 2025).

7.2 Sensitivity to Modifier-Noun Constructions

We introduced a novel behavioral task to assess changes in each model's knowledge of modifier-noun constructions. For each sentence, we compared the probability assigned by a model to the original modifier-noun construction (i.e., "wooden beam") and a swapped version (i.e., "beam wooden"). We then calculated the log ratio of these probabilities $Log(\frac{p(S_{original})}{p(S_{reversed})})$; a positive log ratio indicated that a higher probability was assigned to the Original ordering, while a negative log ratio indicating a higher probability was assigned to the Reversed ordering. This procedure was repeated across each random seed for Pythia-14M.

We observed clear "phase transitions" in performance in the modifier-noun task as well, i.e., at steps 512, 1000, and 2000 (**Figure 7**). Moreover, these changes were highly predictive of changes in disambiguation performance. In a linear mixed effects model with R^2 as a dependent variable and Log Ratio and Log Training Step as fixed effects (and Seed as a random intercept), we found that Log Ratio exhibited a positive relationship with R^2 [$\beta = 0.04, SE = 0.001, p < .001$]. Moreover, using Akaike Information Criterion (AIC) as a measure of model fit—where a lower AIC value corresponds to a better fit—we found that a regression model with only Log Ratio outperformed a model with only Log Training step ($\Delta AIC > 700$). Together, this suggests that across seeds,
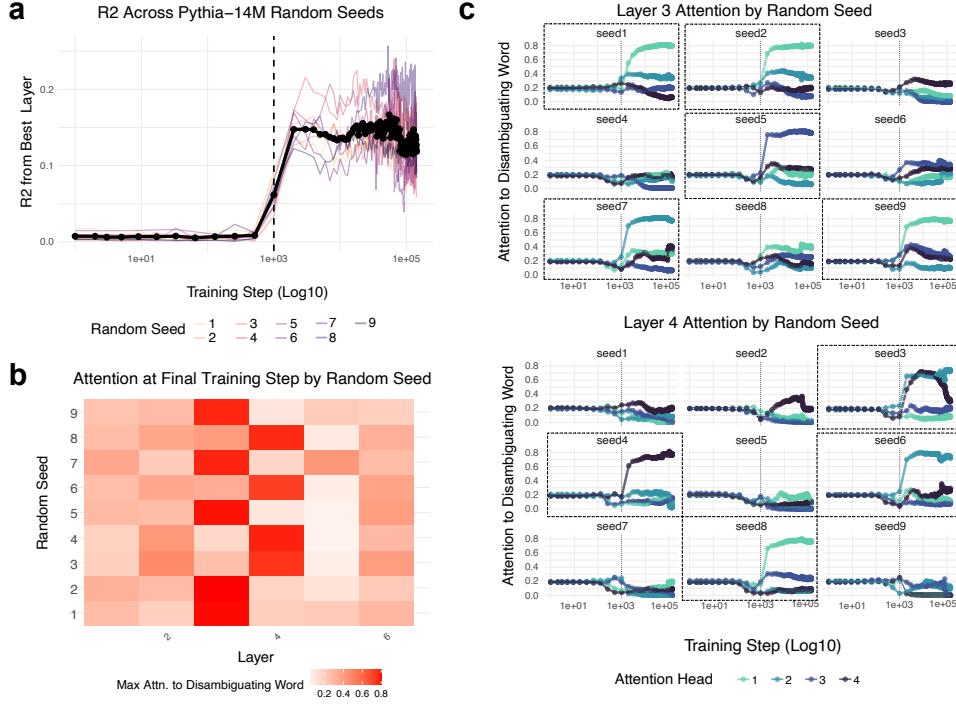
Figure 6: **Generalizability of Attention Patterns Across Random Seeds (a)** R^2 from layer with the maximum R^2 by training checkpoint, across all Pythia-$14M$ random seeds. Black curve represents average max R^2 across seeds. **(b)** Maximum attention to disambiguating word in original RAW-C sentences at final training step by random seed (rows) and layer (columns). Warmer colors indicate larger maximum attention scores. The layer that contains critical attention heads varies according to random initialization, but particularly likely to be Layers 3 and 4. **(c)** *(top)* Layer 3 heads' attention to disambiguating word over the course of pre-training, by random seed. Boxed sub-panels mark seeds containing heads with largest attention to disambiguating word in this layer. *(bottom)* Same as in *top*, but for Layer 4.

there is an inflection point—relatively early in training—involving weight changes to specific attention heads query-key matrices, which results in improved disambiguation performance and enhanced sensitivity to appropriate modifier-noun ordering.

7.3 Supplementary Figures

We include multiple supplementary figures below, which are meant to serve as companions to select main text figures. Each of these is marked as such, and offers a fuller picture of attention head results for Pythia-$14M$ (**Figure 8**) and $410M$ (**Figure 9**) relative to $R2$, over the course of pre-training; ablation results for the Step-1-Copy manipulation for both models (**Figure 10**), which is fully described methodologically in Section 5; and results for zero-ablations of individual target heads and their combination in Pythia-$14M$ (**Figure 11**).
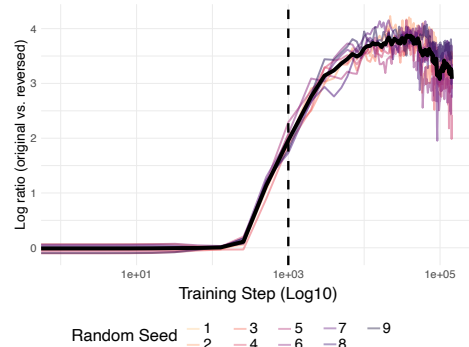

Figure 7: **Trajectory of sensitivity to modifier-noun constructions is consistent across Pythia-14M random seeds.** Log ratios of the probability assigned to original modifier-noun constructions against that assigned to reversed versions of the constructions. Black curve reflects the average across color-coded random seeds.
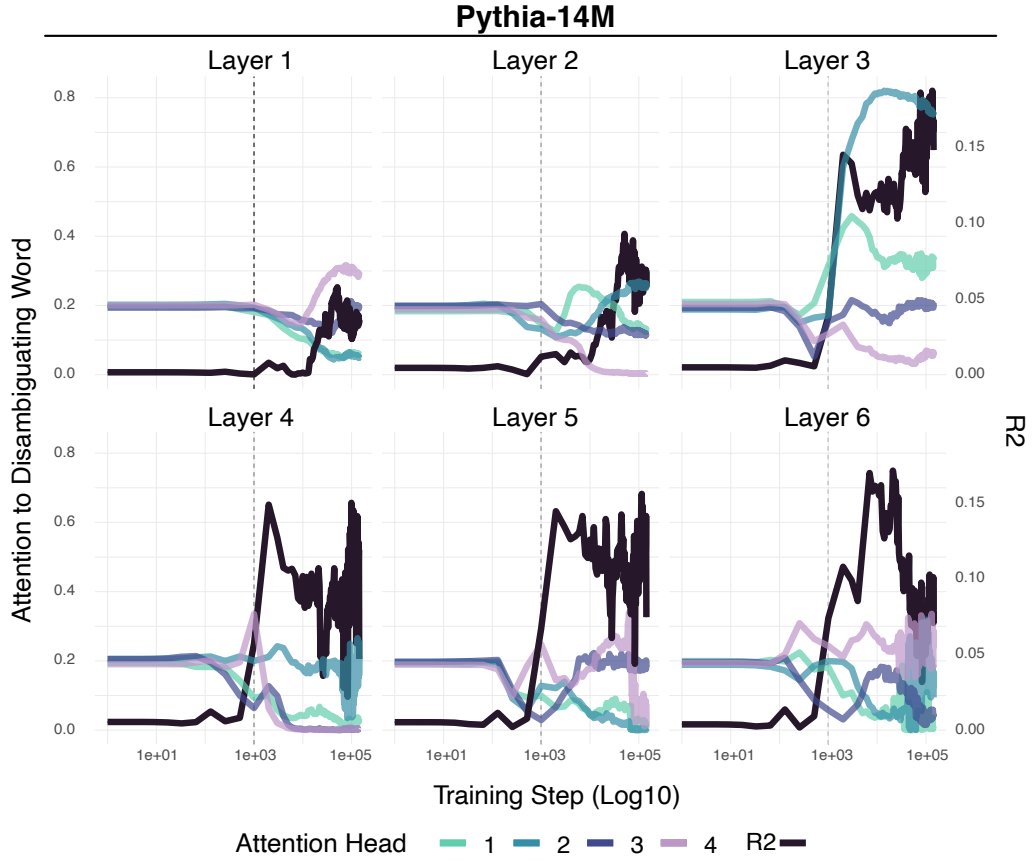
Figure 8: **Companion to Main Text Figure 2c: Identifying candidate attention heads**. Pythia-$14M$'s attention to the disambiguating word over the course of pre-training, for all color-coded attention heads, for all layers. Superimposed in each sub-panel is the $R2$ (black curve) from each corresponding layer's representations. Layer 3 emerges as the layer containing heads with most pronounced increases in attention time-locked to disambiguation performance.
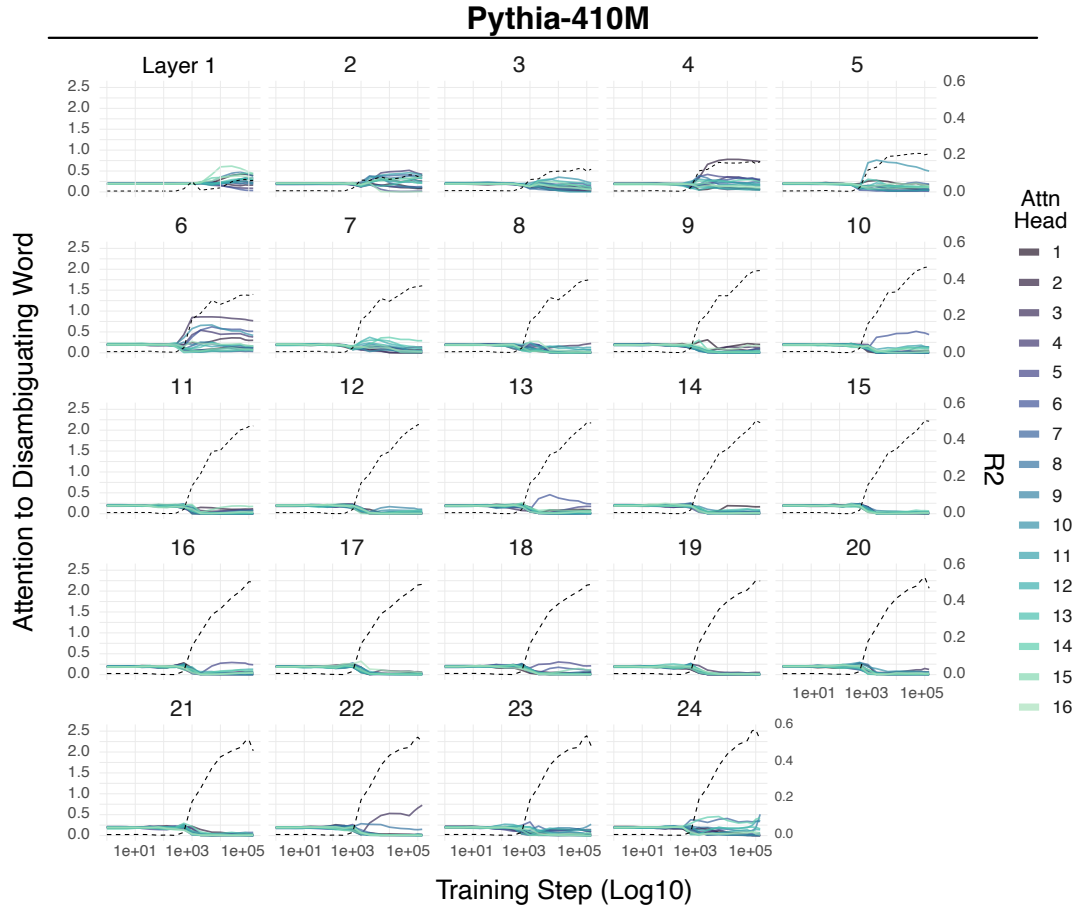
Figure 9: **Companion to Main Text Figure 2d: Identifying candidate attention heads.** All Pythia-$410M$ layers' attention head trajectories for disambiguating word, over pre-training. Superimposed $R2$ (dashed lines) from each corresponding layer's representations.
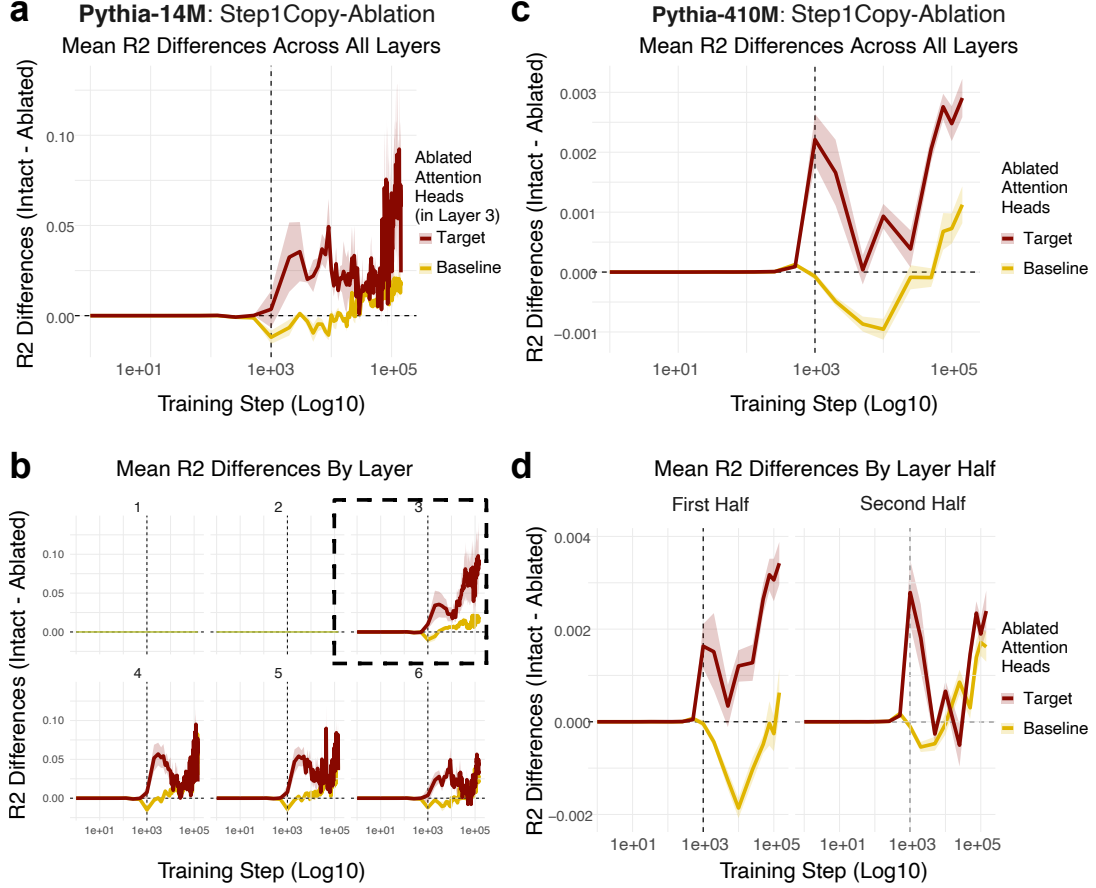
Figure 10: **Companion to Main Text Figure 5: Step-1-Copy Ablations for Pythia-**$14M$ **and** $410M$. **(a)** Mean difference in $R2$ across all layers and all combinations of Pythia-$14M$'s target head ablations, for all training steps. Values > 0 indicate that the intact model's $R2$ exceeded that of the ablated model's $R2$, reflecting causal effect of ablation. Target manipulations refer to ablations of previously-identified Layer 3 heads. Baseline manipulations refer to ablations of Layer 3 heads whose attention to disambiguating words fail to increase with disambiguation performance. **(b)** Same as in **a**, but parcelled out by layer, to illustrate localization of ablation effects, which remain robust throughout training in Layer 3. Dashed square marks the only layer (Layer 3) to suffer head ablations. **(c)** Same as in **a**, but for Pythia-$410M$. Target head ablations causally decrease model performance. **(d)** Same as in **c**, but parcelled out by early versus late layers, to illustrate the selectivity of target-head ablation to earlier layer representations. By the end of training, the effects of target-head ablations do not differ from those of baseline-head ablations.
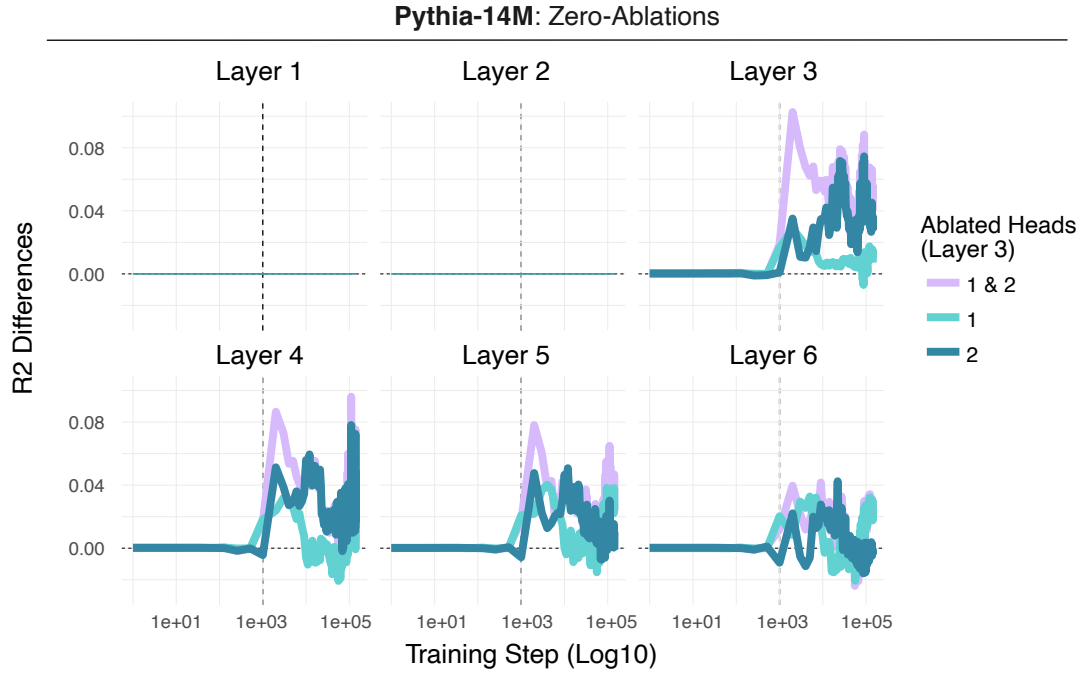
Figure 11: **Companion to Main Text Figure 5: Individual Target Head Zero-Ablation Results, for Pythia-**$14M$ **(a)** Difference in $R2$ (Intact - Ablated) resulting from different color-coded combinations of attention head ablations in Layer 3, across all training checkpoints. Dashed vertical line marks step 1000. At step 2000, ablating *both* Heads $(3, 1)$ and $(3, 2)$ yields much larger performance deficits in Layer 3 representations relative to those of the intact model than ablating each head in isolation. Yet, by the last checkpoints, ablating only Head $(3, 2)$ yields nearly the same effect in Layer 3 representations as having ablated both heads. Representations in later model layers, particularly Layer 6, are less affected than those of intermediate layer representations.