

The Emergence and Sparsification of a Syntactic Circuit

Qing Yao¹ Sasha Boguraev¹ Tiago Pimentel² Kyle Mahowald¹

¹The University of Texas at Austin ²ETH Zurich

{qyao, sasha.boguraev, kyle}@utexas.edu tiago.pimentel@inf.ethz.ch

Abstract

To understand the emergence of grammatical behavior in language models, we track circuits for subject–verb number agreement in pythia-1b and in ten seeds of pythia-410m across training. We find that subject–verb agreement is learned early, and is initially handled by a diffuse collection of neurons. However, after agreement *behavior* has stabilized, the *circuit* responsible for it continues to reorganize. The attention heads which mediate agreement often shift during training: in the case of pythia-1b, behavior is controlled by two heads early in training but by a third (different) head later. Furthermore, the circuit shrinks over training, with pythia-1b’s final circuit being strikingly sparse. We take these results to be informative about how linguistic knowledge emerges and is represented in language models.

 [sparsification-of-sva-circuits](#)

1 Introduction

Our goal is to tell the story, in as much detail as possible, of how subject–verb number agreement emerges across training in language models (LMs). Number agreement has long served as a probe for syntactic competence in language models (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018; Futrell et al., 2019; Mueller et al., 2020; Warstadt et al., 2020; Wei et al., 2021). It has also been used to probe how these models encode grammatical information internally (Finlayson et al., 2021; Lasri et al., 2022; Arora et al., 2024). Notably, this task’s historical interest results from its symbolic specification, which, despite being simple, resists classic statistical methods.¹ A minimal circuit for handling it has to (1) recognize and store the grammatical number of the subject,

¹E.g., it requires tracking structural information — grammatical subjecthood — over potentially unbounded distances, and can thus not be solved by simple Markov models.

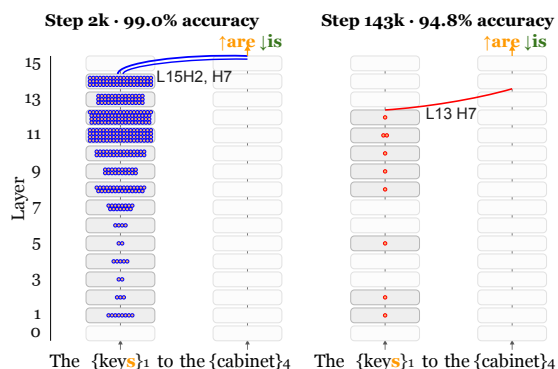


Figure 1: The subject–verb agreement circuit for pythia-1b at step 2k (left) and step 143k (right). While agreement is acquired early (within 2% of training) and the model remains proficient thereafter, the underlying circuit continues to reorganize and sparsify. Circles represent the smallest subset of top-ranked neurons whose sufficiency and necessity exceed 50% (see Section 4.2).

(2) transport this information to the prediction position, and (3) license the corresponding output form.

In this work, we first (Experiment 1) trace the emergence of such a circuit over the course of language model training, studying ten seeds of pythia-410m and one seed of pythia-1b. We find that agreement is learned quickly and is initially controlled by a diffuse circuit. Over the course of training, however, we find that while the model’s performance remains mostly unchanged, the underlying circuit gradually evolves: it both sparsifies (concentrating in a smaller set of neurons) and is rerouted within the model (being implemented by different attention heads). By the end of training, a surprisingly sparse circuit captures most of the agreement signal in pythia-1b (Figure 1).

Such a gradual transition is consistent with recent reports of emergent syntactic behavior (Chen et al., 2024), and with work in developmental LM interpretability more broadly (Wang et al., 2025; Chang et al., 2024; Fehlaue et al., 2025; Kangaslahti et al., 2026), including recent work track-

ing the emergence of circuits longitudinally in open-checkpoint models (Tigges et al., 2024; Ou et al., 2025; Hakimi et al., 2025; Aljaafari et al., 2026). This circuit’s trajectory also parallels accounts of grokking (Power et al., 2022; Nanda et al., 2023; Varma et al., 2023; Merrill et al., 2023), where a task is first learned by memorization, but that solution is replaced by a sparser and more general circuit after training accuracy has saturated.

Later in our paper (Experiment 2), we zoom in on pythia-1b. We first track how the circuit evolves in this model, tracing how a subset of neurons remains relevant for agreement throughout training, but changes “allegiance” at some point, shifting the attention head they write to. Second, we further sparsify this model’s final circuit while preserving its performance. We find that just 18 neurons and a single attention head can capture the primary channel for agreement in this model. We show this sparse circuit generalizes across different forms of subject–verb number agreement and to in-the-wild examples from WikiText (Merity et al., 2017). Finally, we demonstrate that this circuit’s information flows through a one-dimensional channel of the relevant transfer head. Overall, we show that mechanistic reorganization and sparsification continue beneath a stable agreement behavior. In the case of pythia-1b, this culminates in a sparse and generalizable circuit for number agreement.²

2 Finding agreement in LMs

How can we understand and characterize the mechanisms responsible for a behavior in LMs? The field of (mechanistic) interpretability focuses on answering this question. Historically, researchers relied on diagnostic probes to find signals of the task within LMs’ representations (Alain and Bengio, 2016; Hewitt and Manning, 2019; Hewitt and Liang, 2019; Pimentel et al., 2020; Kulmizev et al., 2020, *i.a.*). In the years since, though, the field has grown into a substantial literature, with causality becoming one of its central aspects (Olah et al., 2020; Elhage et al., 2021; Elazar et al., 2021), and circuit-discovery a major thread within it (Wang et al., 2023; Hanna et al., 2023; Conmy et al., 2023; Stolfo et al., 2023; Hanna et al., 2024, *i.a.*).

Notably, some of the earliest work on this topic tried to identify the mechanisms behind number agreement in LMs. Studying LSTMs, Giulianelli

et al. (2018) and Lakretz et al. (2019) identified components or circuits in these models which implemented number agreement. Studying transformer models, Clark et al. (2019) identified syntax-aware attention heads in BERT, and Ferrando and Costa-jussà (2024) identified a bilingual number-transport attention head for English and Spanish in gemma-2b. More recently, there has been a shift towards trying to identify more fine-grained circuits within LMs. Marks et al. (2025), for instance, used sparse autoencoders (SAEs) to identify sparse circuits (with fewer than 250 nodes) responsible for subject–verb agreement in both pythia-70m and gemma-2-2b. Building on such work, Arora et al. (2026) showed that circuits over MLP activations can be as sparse as SAE-based ones.

Most relevantly, Tigges et al. (2024) trace Pythia circuits across four tasks, one of which is subject–verb agreement. They find that the underlying algorithm implemented by the circuit remains stable across training, but that the circuit’s components and size can change. While they focus on a variety of tasks and localize their circuit at a coarse level of MLP blocks and attention heads, we zoom in on a single task, studying its behavior across many random seeds and focusing our analyses at the neuron level. We track how many and which neurons are needed to causally control agreement across training, we characterize how they reroute between number-transport heads, and we describe a complete subject-to-prediction circuit in the model.

Overall, our work provides three primary contributions. First, building on prior longitudinal circuit analyses across Pythia checkpoints and scales (e.g., Tigges et al., 2024), we study subject–verb agreement at finer granularity, combining head- and neuron-level interventions to track the agreement circuit across checkpoints and random seeds. Second, we characterize not only which components change in this circuit, but how agreement-relevant neurons sparsify and reroute as the dominant transport head changes. Finally, we show that most of the agreement behavior in pythia-1b can be captured by a sparse circuit.

3 Methods and materials

3.1 Models

We study models of the pythia family (Biderman et al., 2023). We focus our analysis on pythia-1b — a 16-layer autoregressive transformer with 8 attention heads per layer, a hidden size of 2048 neu-

²Throughout this paper, we focus on subject–verb number agreement, but often term it more compactly as ‘subject–verb agreement’ or simply ‘agreement’.

rons, and MLPs with intermediate dimension 8192 — and on the ten seeds of pythia-410m, nine of which are from PolyPythias (van der Wal et al., 2025), each with 24 layers, 16 attention heads per layer, hidden size 1024, and MLP dimension 4096. By performing our analysis across model sizes and seeds, we study the behavior and internal representations of the models more generally.

3.2 Subject–verb agreement task

We study subject–verb agreement, focusing on sentences with an intervening prepositional phrase, a task we refer to as PP for convenience. An exemplar minimal pair can be seen in (1).

- (1) **b.** The₀ **key**₁ to₂ the₃ cabinet₄ → **is**
s. The₀ **keys**₁ to₂ the₃ cabinet₄ → **are**

We denote token roles with subscripts: pos₁ for the subject position, pos₂ for the last token position of the preposition, and pos₄ for the final token/prediction position. In order to generate a large number of unique minimal pairs, we adapt the templates from Arora et al. (2024). Specifically, we restrict the lexical inventory to single-token nouns and verbs (in both singular and plural forms) and only generate minimal pairs differing in subject number.

In Section 5, we evaluate how the circuits we discover generalize to other agreement examples. Beyond PP sentences, in that section, we also investigate sentences with a subject relative clause distractor (SUBJECTRC), an object relative clause distractor (OBJECTRC), or no distractor (SIMPLE).

For the PP task, we create a training set of 400 minimal pairs, balanced in the number of subjects and distractors. We use this set to identify candidate circuit components. We also create disjoint evaluation sets for PP and the other templates, each consisting of 192 balanced minimal pairs (with the exception of SIMPLE, which has 32 minimal pairs). For both behavioral validation (Section 3.3) and circuit evaluation (Section 3.4), we use the evaluation sets. For more details on all templates, along with associated exemplars, see App. C.

3.3 Behavioral validation

To measure the behavioral performance of the studied models on subject–verb agreement, we calculate their accuracy on the evaluation set using the following paradigm. A model is deemed to correctly process a minimal pair if for each item, the probability for the correct continuation is higher than the alternative. That is, given the minimal-pair

sentences in (1), a model is accurate if it correctly predicts for *both* sentences that

$$p(\text{is}|\text{The key ...}) > p(\text{are}|\text{The key ...}) \quad \text{and} \quad (2a)$$

$$p(\text{is}|\text{The keys ...}) < p(\text{are}|\text{The keys ...}). \quad (2b)$$

3.4 Circuit evaluation

Let $\mathcal{C}$ be a model component, or a collection of components, proposed to be a part of a circuit of interest and thus responsible for a certain task. To evaluate the causal effect of $\mathcal{C}$ on that task, we intervene on its activations to calculate its **necessity** and **sufficiency** for downstream behavioral control. This section details these metrics.

Both of these metrics rely on a metric we refer to as the **agreement gap**, G . Let $\mathbf{b}$ and $\mathbf{s}$ be base and source minimal-pair sentences, with $\ell_{\mathbf{b}}$ and $\ell_{\mathbf{s}}$ their respective continuations. In ex. (1), for instance, $\ell_{\mathbf{b}} = \text{is}$ and $\ell_{\mathbf{s}} = \text{are}$. Given this pair, G is the expected difference in log probability between the base and source continuations:

$$G^{\mathbf{b}} = \mathbb{E}_{(\mathbf{b}, \mathbf{s}) \in D} \log \left(\frac{p_{\theta}(\ell_{\mathbf{b}}|\mathbf{b})}{p_{\theta}(\ell_{\mathbf{s}}|\mathbf{b})} \right), \quad (3)$$

$$G^{\mathbf{s}} = \mathbb{E}_{(\mathbf{b}, \mathbf{s}) \in D} \log \left(\frac{p_{\theta}(\ell_{\mathbf{b}}|\mathbf{s})}{p_{\theta}(\ell_{\mathbf{s}}|\mathbf{s})} \right)$$

where p_{θ} stands for the evaluated model run under a specific configuration. $G^{\mathbf{b}}$ and $G^{\mathbf{s}}$ therefore represent agreement gaps of the model when processing the base sentence or source sentence, respectively. A higher preference for the base continuation corresponds to a more positive gap score. For a behaviorally correct model, $G^{\mathbf{b}}$ should be positive, while $G^{\mathbf{s}}$ should be negative. Our notions of sufficiency and necessity then measure how much this gap changes under manipulations of activations.

A model component $\mathcal{C}$ is **necessary** if performing a mean ablation (Wang et al., 2023) causes the model’s agreement gap to drop to near-zero:

$$\text{Nec}(\mathcal{C}) = \frac{G^{\mathbf{b}} - G^{\mathcal{C} \leftarrow \text{avg}}}{G^{\mathbf{b}}} \quad (4)$$

where $G^{\mathcal{C} \leftarrow \text{avg}}$ is the agreement gap on input $\mathbf{b}$ when we replace $\mathcal{C}$ ’s activations with their average value over a training set.³

A model component $\mathcal{C}$ is **sufficient** if exchanging its activations across minimal pairs $\mathbf{b} \leftarrow \mathbf{s}$ shifts the verb predictions for $\mathbf{b}$ towards those of $\mathbf{s}$. This

³Mean activations are computed on training sets for the PP template and on the evaluation set for other templates.

is measured by the following metric:

$$\text{Suff}(\mathcal{C}) = \frac{G^{\mathbf{b}} - G^{\mathcal{C} \leftarrow \mathbf{s}}}{G^{\mathbf{b}} - G^{\mathbf{s}}}, \quad (5)$$

where $G^{\mathcal{C} \leftarrow \mathbf{s}}$ is the agreement gap when the model is run with input $\mathbf{b}$ with $\mathcal{C}$'s activations replaced by their values under input $\mathbf{s}$. Intuitively, a component is sufficient if intervening on $\mathcal{C}$'s activations has roughly the same causal effect as directly swapping the subject token for its singular/plural counterpart.

Relation to other metrics. Existing circuit work evaluates **completeness** and **faithfulness** (Wang et al., 2023; Marks et al., 2025; Arora et al., 2026). Within a predefined scope (global set of relevant components), a circuit is **complete** if ablating the circuit degrades performance as much as ablating the entire scope, and **faithful** if ablating its complement preserves performance. Our notion of necessity is conceptually similar to completeness, as ablating the entire scope should typically bring the model close to chance performance — and thus near zero agreement gap. We choose to evaluate sufficiency rather than faithfulness as defining a circuit's complement (to ablate) is difficult, especially when evaluating single components.

3.5 Circuit discovery methods

This section details the methods we use to localize our circuit. Throughout our experiments, we discover and evaluate our circuits on the PP template; we later show that the discovered circuit generalizes to the other forms of number agreement.

Head identification. We identify attention heads responsible for transporting a subject's number by measuring each head's necessity and sufficiency at the prediction position, pos_4 , across the training set. We intervene on the full head to compute these metrics, not only on its subject–prediction values.⁴

Neuron identification. We follow Arora et al. (2026) in using Relevance Patching (RelP; Jafari et al., 2025), a gradient-attribution method, to localize neurons responsible for LM behavior. We utilize the unpaired variant of RelP which scores components by the product of their activation on the base input and the gradient of a task metric with respect to that component. It differs from attribution patching methods (Nanda, 2023; Kramár et al., 2024; Syed et al., 2024) by computing the gradient

⁴After a head is selected as a circuit component, it is then evaluated for necessity and sufficiency on the evaluation sets.

on a locally linear replacement model — where in the backward pass each nonlinearity is frozen at its forward value. This results in a less noisy and more faithful gradient (Jafari et al., 2025). Details of our RelP implementation, including the linearization rules, an integrated-gradient robustness check, and task-specific modifications, are in App. D.1.

We score each MLP neuron, $\mathcal{C}_n$, as in (6), where $v_n(\mathbf{b})$ is $\mathcal{C}_n$'s activation on the base input, m is the per-example base agreement gap, and the gradient is taken relative to the replacement model:

$$\text{RelP}(\mathcal{C}_n) = \mathbb{E}_{(\mathbf{b}, \mathbf{s}) \in D} \left[v_n(\mathbf{b}) \frac{\partial m(\mathbf{b}, \mathbf{s})}{\partial v_n(\mathbf{b})} \right], \quad (6)$$

where $m(\mathbf{b}, \mathbf{s}) = \log \frac{p_{\theta}(\ell_{\mathbf{b}} | \mathbf{b})}{p_{\theta}(\ell_{\mathbf{s}} | \mathbf{b})}$.

A positive RelP score indicates that $\mathcal{C}_n$ pushes the model toward verbs licensed by their subject. We apply this method to score every post-GELU MLP neuron in each layer at the relevant token positions for a given analysis. As RelP is an attribution method, we treat it as a hypothesis generator for causally-relevant neurons. To ensure their importance, we then causally validate such candidate neurons through interventions detailed in Section 3.4.

Neuron and head routing. Having separately identified neurons and heads relevant to agreement, we want to establish routing relations between them. For instance, given two heads that could both transport number information, we want to determine which route a candidate neuron takes. To this end, we attribute each neuron to heads (possibly to both or neither) using a path-patching-like method (Wang et al., 2023; Goldowsky-Dill et al., 2023). For neurons before the prediction position, we (1) intervene on the neuron and measure the downstream causal effect and (2) intervene on the same neuron while holding a head $\mathcal{C}_h$'s corresponding value edge fixed at the base activation. Any effect that drops out under the second intervention relative to the first is attributed to the route via $\mathcal{C}_h$. Prediction-side neurons are likewise classified by which head interventions they read out from. We describe our methodology in detail in App. D.2.

4 Experiment 1: Behavior–mechanism decoupling in training

In our first experiment, we track the internal organization of pythia-1b and the ten seeds of pythia-410m as they develop the capacity to perform subject–verb agreement. At each checkpoint,

we measure their agreement capabilities, identify the neurons associated with agreement using RelP, and determine which heads are sufficient and necessary for agreement. Our goal here is not yet to produce complete agreement circuits, but to provide a description of the models’ behavior and identify agreement-relevant components through training.

4.1 Agreement behavior is learned within 2% of total training steps

First, we measure the behavioral performance of our studied models across training. The y-axis in Figure 2 shows our results. At steps 256 and 512, the models have not learned agreement yet, always preferring singular continuations over plural ones, regardless of subject number. Since agreement accuracy requires predicting both singular and plural verbs correctly, through step 512, no model is even 2% accurate on agreement. By step 2k, though, agreement is high for all ten pythia-410m seeds, with a median accuracy of 96.9% (range 88.5–100%), and a median agreement gap of 3.35. At the end of training, the models’ median accuracy and gap are 91.7% and 3.39, respectively. This shows agreement capabilities are learned early in training and remain relatively stable throughout. The larger pythia-1b shows the same pattern — its accuracy and gap change from 99.0% and 3.21 at step 2k to 94.8% and 3.48 at the final checkpoint.⁵

Overall, our results are consistent with [van der Wal et al. \(2025\)](#)’s finding that linguistic knowledge is largely acquired by $10^3 - 10^4$ steps. This raises a question, though: once agreement behavior is saturated, is the mechanism behind this syntactic competence effectively learned and frozen in place, or does it continue to evolve in ways that behavioral measurements do not reveal? To answer this question, we first track the neurons that process subject number at the subject, and then the attention heads that transport it to the prediction position.

Takeaway. The models learn agreement after only around 2k training steps, with behavior remaining generally consistent thereafter.

⁵The better performance at step 2k compared to the final checkpoint is somewhat mysterious. The final checkpoint errors are cases with opposite-number distractors, suggesting that models are perhaps falling for agreement attraction effects. Broadly, though, this result is consistent with the “unruly” BLiMP curves shown in [Bunzeck and Zarri  \(2024\)](#).

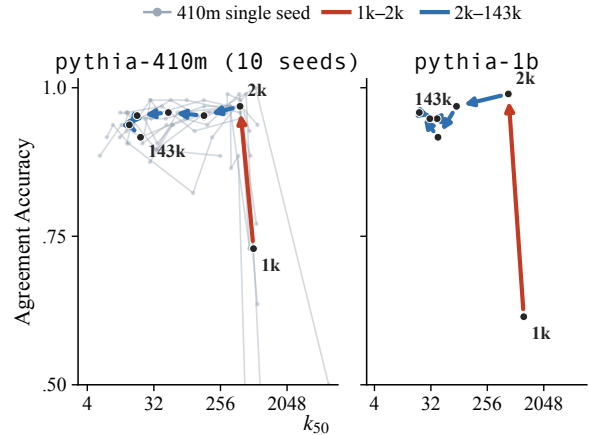


Figure 2: Agreement accuracy vs. # of subject-neurons in the circuit through training across our models. Each point is a checkpoint. Typically, steps 1k–2k form the agreement acquisition phase. In the remaining steps, the circuit sparsifies before reaching some equilibrium.

4.2 Circuit neurons sparsify through training

We now consider the neurons implicated in the agreement circuits at the subject position, and compare how they change across training checkpoints. In particular, if the agreement circuit is effectively “finalized” after proficiency is achieved in this task, the number of circuit neurons should remain stable.

For each model and checkpoint, we identify the neurons most responsible for agreement by (1) computing a RelP score for all MLP neurons across all layers at the subject position on the training set, (2) ranking them in descending order, since a larger positive RelP score indicates stronger support for agreement, and (3) evaluating their causal effect on the model. For (3), we define k_{50} as the smallest k such that a circuit consisting of the top- k ranked neurons, $\mathcal{C}_k$, is at least 50% necessary and sufficient under intervention.⁶

$$k_{50} = \min \{k \mid \min(\text{Nec}(\mathcal{C}_k), \text{Suff}(\mathcal{C}_k)) \geq 50\%\}. \quad (7)$$

Our results are in Figure 2. We find that the agreement circuit tends to sparsify throughout early training — even as performance remains constant — leveling off around halfway through training. Further, the same neurons are generally deemed important across checkpoints. For example, in pythia-1b, we find that all 21 neurons in step 18k’s k_{50} are present in the larger 42-neuron k_{50} at step 8k, with 20 also appearing in the 41-neuron k_{50} at step 50k, and 18 remaining in the final checkpoint’s k_{50} . The same pattern holds across most

⁶Finding the precise value of k is prohibitively expensive. In practice, we compute the score in (7) at intervals of powers of two, and interpolate linearly in $\log_2$ space for an estimate.

runs of pythia-410m: taking the intersection of the k_{50} neuron sets of all post-8k checkpoints, we find a median of 85% overlap with each seed’s smallest k_{50} neuron set. For more details on neuron overlap across checkpoints and on whether results hold across thresholds of k , see App. E.

Notably, we observe a substantial increase in k_{50} during training for two pythia-410m seeds (3 and 4). Interestingly, these seeds are exactly the training outliers identified by van der Wal et al. (2025), which underperformed on benchmarks after suffering loss spikes during training. These loss spikes occur precisely where we see sharp rebounds in individual k_{50} counts (at steps 83k and 123k for seeds 3 and 4, respectively; more details in App. E.3).

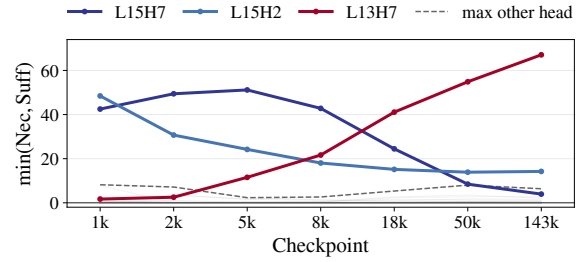
This analysis tells us that agreement circuits become sparser at the subject position during training, involving fewer neurons. But the complete circuit also involves transporting that signal to the prediction position. Next, we ask whether the number-transporting heads remain stable after agreement is learned, or if they also undergo reorganization.

Takeaway. After models learn agreement, their circuits still sparsify, with progressively fewer neurons being required to capture the same level of causal control over agreement across training.

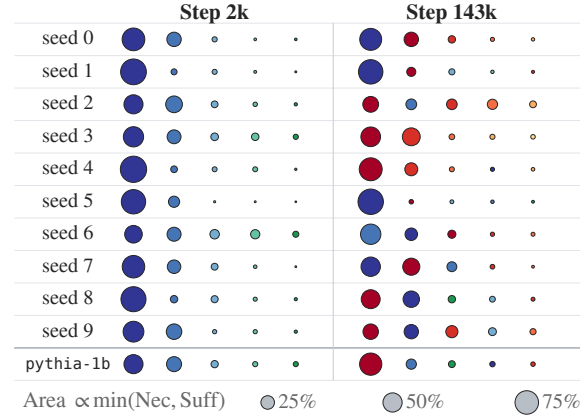
4.3 Number is carried by a small set of heads

We now identify the attention heads in the agreement circuits at the prediction position by similarly measuring necessity and sufficiency across training. Figure 3a demonstrates that in pythia-1b, early agreement circuits are almost entirely mediated by two heads (L15 H7 and L15 H2). However, by step 18k, head L13 H7 becomes the most important, and remains so through the final checkpoint (although L15 H2 persists as a weaker secondary channel). In comparison, all other heads contribute negligibly.

Figure 3b tells a similar story for pythia-410m. Across all models, a small set of heads is responsible for carrying agreement information, often with a clear leading head individually achieving over 50% in necessity and sufficiency. Despite consistently small numbers of relevant heads, none of the examined models have the same top-3 heads at steps 2k and 143k. (Detailed results in App. E.4.)



(a) Head scores across training for pythia-1b.



(b) Top-five heads across model runs.

Figure 3: Attention-head reorganization. Top: In pythia-1b, attention heads reorganize across training checkpoints, with the heads important early, L15 H7 and L15 H2, ceding importance to L13 H7. All other heads are negligible. **Bottom:** The identities of the leading heads change across training. Circle area is proportional to each head’s individual $\min\{\text{Nec}, \text{Suff}\}$ score, with each individual head a different color.

Takeaway. We find that agreement is generally mediated through a small set of late-layer attention heads, and the leading transport heads often shift even after behavior stabilizes.

5 Experiment 2: The sparsification of the pythia-1b circuit

Having shown the general developmental trajectory of agreement circuits across seeds and models, we now zoom in on the final circuit for one model: pythia-1b. First, we show that its neuron- and head-level reorganizations happen in tandem. Next, we prune the circuit down to 18 neurons and one attention head. We then demonstrate this hand-constructed sparse circuit generalizes to novel templates and in-the-wild text. Finally, we show the head’s transport can be further localized to just a one-dimensional subspace of its output channel.

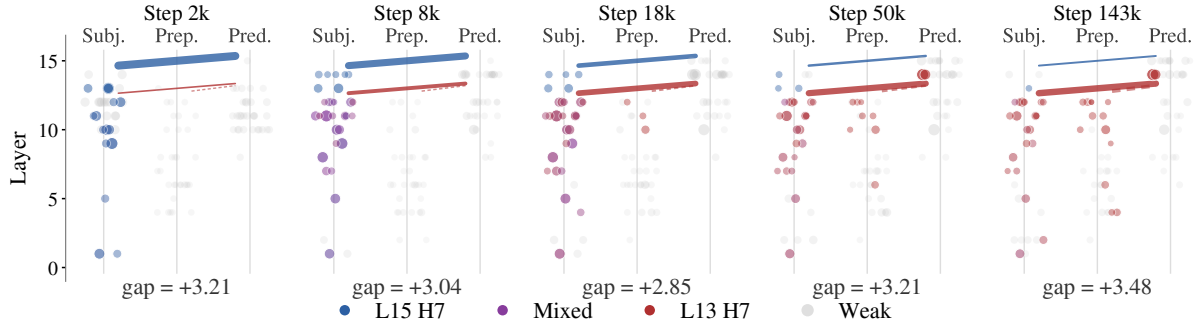


Figure 4: Top-32 ranked ReLP neurons across positions and attention heads L13 H7 and L15 H7 across checkpoints as identified on PP training set ($N = 400$). Neurons are colored by their correspondence to the respective heads, on a gradient between blue (L15 H7) and red (L13 H7), determined by path patching. Gray neurons are not attributed to either route. Line width indicates each head’s sufficiency. We see that throughout training, there is a gradual transition from an L15 H7-routed mechanism to an L13 H7-routed one.

5.1 Joint reorganization

In Section 4, we showed that while agreement is learned early, its circuit reorganizes during training. Two changes occur together, namely that the number of subject-neurons in the circuit decreases, and the heads transporting the number information shift from a L15 route to a L13 H7 route. In this section, we link these observations and show that circuit neurons change the pathway they feed into or read out from as the dominant head changes.

To this end, we first collect the top-32 ranked neurons according to ReLP at each checkpoint across three token positions: subject pos_1 , preposition pos_2 , and prediction pos_4 . We then apply the routing attribution method from Section 3.5 to measure the extent to which a neuron reads from or writes to L15 H7 (the head primarily responsible for information transport early in training) or L13 H7 (the head responsible late in training).

Figure 4 shows our results across checkpoints. We see that as the relative head importance changes, the neurons follow. That is, early in training, we see neurons route through L15 H7, the early primary transport channel. However, as training progresses and L13 H7’s channel emerges, pre-L13 neurons begin to show mixed routing. Then, as L13 H7 becomes increasingly prominent, most neurons reorganize to start writing to or reading from that channel. By the final checkpoint, both upstream and downstream neurons are primarily aligned with the L13 H7 channel.

Takeaway. The agreement circuit in pythia-1b shifts from a route mediated predominantly by L15 H7 to one mediated by L13 H7 throughout training. As the heads shift, the neurons follow.

5.2 A final sparse circuit

Having shown the trajectory of the joint neuron and head circuit through training, we now zoom in on the final checkpoint of pythia-1b to produce a sparse circuit for agreement.

Circuit construction. Our goal here is to construct a sparse circuit for subject–verb agreement. To do so, we made a series of choices that we believe would represent the model’s internal agreement mechanism while minimizing the circuit size.

We started with two qualitative observations based on the results of the earlier analyses and further manual inspection of the results. First, number agreement at the final checkpoint is largely carried by one head, L13 H7. Second, MLPs propagate agreement information across layers, while self-attention is not strongly involved in this process.⁷

Our goal was then to construct a circuit consisting of the one relevant head and a set of relevant neurons that captures the role of the MLPs in processing agreement. To this end, we again used ReLP to identify a set of potentially important neurons for the circuit. We further augmented this set with a few extra hand-picked neurons, identified as important via other methodologies.⁸ This resulted

⁷We show in App. F.2 that ablating self-attention has negligible effect on agreement.

⁸We performed an initial analysis, prior to applying ReLP, using Distributed Alignment Search (DAS; Geiger et al., 2024), a supervised interpretability method which localizes LM behavior to causally sufficient subspaces in the higher-dimensional activation space. Consistent with previous works (e.g. Arora et al., 2024), we found that the number feature could be localized to a single dimension of the activation space, but this subspace appeared to change across layers due to MLPs writing additional number information into the residual stream. This motivated us to identify important neurons which write to these DAS subspaces. For each neuron, we

Components C	G	Measure
Necessity	$G^{C \leftarrow \text{avg}}$	Nec.
C_{full}	+0.76	78%
C_{scope}	+0.58	83%
Sufficiency	$G^{C \leftarrow s}$	Suff.
C_{full}	-2.29	83%
C_{scope}	-2.64	88%

Table 1: Necessity and sufficiency of the full circuit, C_{full} , relative to the scope, C_{scope} , tested on an evaluation set of 192 PP items. The circuit is nearly as necessary and sufficient as the full scope, despite containing around 1/6000 as many MLP neurons.

in a small set of 21 candidate neurons. We then greedily pruned neurons whose removal had negligible effects on circuit necessity and sufficiency. This procedure removed three redundant neurons, leaving us with our final circuit.

Notably, this circuit’s construction process involved several judgment calls. Our goal was not to provide a systematic solution for its creation, but to try to qualitatively understand and improve the circuit iteratively. There are no guarantees that this circuit is maximally performant or sparse. At the end of the procedure, we had a sparse circuit, C_{full} , which contains 18 neurons and one attention head. Ten neurons are at the subject position pos_1 and upstream of the head L13 H7 (in L1–L12 MLPs); eight neurons are at the prediction position pos_4 and downstream of the head (in L14–L15 MLPs).

Importantly, we do not see the circuit as the *complete* agreement mechanism for pythia-1b — Figure 3a and Figure 4 show that L15 H2 and preposition neurons also carry agreement signal — nor is it constructed to maximize any metric. We interpret the circuit to be a sparse representation of how the model transports the number feature from the subject to the prediction position via a specific L13 H7 route. The scope of the circuit C_{scope} thus consists of L1–L12 MLPs at pos_1 , L13 H7, and L14–15 MLPs at pos_4 .⁹

computed its number-dependent contribution to the next-layer as the product of its paired activation difference between singular and plural inputs and the neuron’s output vector. The cosine similarity of the resulting contribution with the next-layer DAS subspace was then used to rank neurons. DAS settings and the neuron selection and pruning procedures are in App. F.1.

⁹We omit the subject L0 MLP from the scope as MLP ablation there is globally destructive. We take this as evidence that the L0 MLP may be responsible for *constructing* the number feature with our circuit responsible for *transporting* the feature to readout. Full details regarding scope selection and L0 MLP ablation results are in App. F.2.

Circuit components. The circuit seems to carry out the minimal mechanism laid out in Section 1. (1) *Amplifier* neurons C_A at the subject process and propagate the number feature, (2) an attention head C_H transports this information to the prediction position, and (3) the head outputs and readout neurons C_P and C_N license the corresponding output form.

Our characterization of the neurons is as follows: neurons in C_A have high activations on plural subjects, while their activations on singular inputs are comparable to those on non-nouns — we thus think of C_A as *plural detectors*. Readout neurons come in two flavors: *pro-agreement* neurons C_P promote the correct verb continuations, while *anti-agreement* neurons C_N seem to oppose agreement. For the complete set of circuit components and more details on this characterization, see App. F.3.

Evaluating the circuit. We now measure the necessity and sufficiency of our final circuit, providing the same for the entire scope as a baseline in Table 1. We find our circuit is 78% necessary, with its mean ablation dropping the agreement gap nearly as much as mean ablating the full scope. We also find our circuit is 83% sufficient, which is comparable to the 88% sufficiency of the entire scope. Further, the circuit’s complement has little effect on agreement (App. F.4).

Takeaway. A sparse circuit of 18 neurons and a single attention head captures nearly all of the agreement information flow within the scope.

5.3 Characterizing the final circuit

Having produced and validated a sparse circuit on PP templates, we now show that it mediates agreement more generally. Then, we demonstrate that we can compress the sparse circuit even further, localizing the information flow within C_H to a one-dimensional subspace.

Generalization. First, we show our circuit is necessary and sufficient for agreement in further settings. In particular, we measure our circuit on three synthetic templates with novel structure — agreement across a subject relative-clause (SUBJECTRC), an object relative-clause (OBJECTRC), or no distractors (SIMPLE) — as well as on mined natural examples from WikiText (Merity et al., 2017) for which we manually construct the minimal pair. Templates and exemplars for each evaluation are in App. C.

In Table 2, we broadly see that our circuit’s

Template	G^b	G^s	Nec.	Suff.
PP	+3.48	-3.48	78%	83%
SUBJECTRC	+4.50	-4.50	86%	89%
OBJECTRC	+4.32	-4.32	76%	80%
SIMPLE	+6.47	-6.47	66%	77%
WikiText	+4.29	-3.75	73%	82%

Table 2: Circuit necessity and sufficiency across templates ($N = 32$ for SIMPLE, $N = 192$ for other templates, and $N = 80$ for WikiText). Mean activations are computed over the larger training set for PP, and directly on the evaluation set otherwise.

necessity and sufficiency on these templates are comparable to those for PP. The circuit performs slightly worse on the SIMPLE template, which likely reflects the structural difference of the stimuli. In sum, these results suggest that our circuit is general, and not overfit to just the PP templates from which it was discovered.

We also perform additional generalization tests, showing that circuit activations drawn from one template can control agreement in another, and that the circuit holds under most perturbations to the templates and on novel subjects and verbs. These results, along with failure cases (the circuit failing on templates with an appended adverb, and SIMPLE being a weak donor for cross-template intervention) are found in App. F.5.

C_H can be localized to a 1D channel. We can further localize the head transport to a subspace of L13 H7’s output-value circuits (Elhage et al., 2021) using DAS. To this end, we train three seeds of DAS on the PP training set to successfully localize behavior to a 1D subspace of the head’s value output. (See App. F.6 for DAS training details.)

Table 3 shows the results. We find a 1D subspace which essentially captures the head’s total causal effect, and both intervening on and mean ablating the subspace complement barely affect agreement. These results suggest that the head’s contribution to the agreement task can be surgically localized to a one-dimensional channel.

Takeaway. Our circuit is general despite being discovered on the PP template, mediating agreement robustly across synthetic and naturalistic sentences. Further, the information is routed through a single dimension within the attention head.

Components C	G	Measure
Necessity	$G^{C \leftarrow \text{avg}}$	Nec.
C_H (Full)	+1.15	67%
C_H (1D subspace)	+1.02	$70.58 \pm 0.03\%$
C_H (255D Complement)	+3.63	$-4.13 \pm 0.02\%$
Sufficiency	$G^{C \leftarrow s}$	Suff.
C_H (Full)	-1.25	68%
C_H (1D subspace)	-1.47	$71.05 \pm 0.02\%$
C_H (255D Complement)	+3.68	$-2.82 \pm 0.02\%$

Table 3: The 1D channel identified by DAS has a comparable effect to the entire 256D C_H . We report the mean necessity and sufficiency $\pm$ standard error across three independent seeds of DAS.

6 Conclusion

We believe that the circuit discovered makes good on the mechanistic account of number agreement that we started with. Within the scoped L13 H7-mediated pathway in pythia-1b, ten subject-position neurons propagate number information, L13 H7 carries it to the prediction position through a one-dimensional causal channel, and eight downstream neurons control the verb read-out. Agreement circuits in pythia-1b and the ten pythia-410m seeds were diffuse early in training but generally became sparser over time.

We leave it to future work to analyze how and why this transition occurs, although we hypothesize the following: the early circuit emerges stochastically out of a set of features that proved useful for doing the subject–verb agreement task. Then, this diffuse circuit serves as a scaffold for a sparser and more specialized circuit to emerge. Our hope is that, by granularly studying a single syntactic phenomenon, we can complement wider investigations into LMs’ syntactic abilities and thereby better characterize linguistic knowledge in LMs (as called for by Futrell and Mahowald, 2026).

Limitations

The goal of this paper was to focus on just one linguistic task of subject–verb agreement, but this is also, of course, a limitation. We also focus only on English. Future work could generalize these methods to other constructions, models, and languages. For present purposes, we felt that a wider investigation would have undermined the particular motivation of this project.

Further, we find neuron sparsification and head rerouting to be common across model training runs, but have not provided an explanation for it. Another

limitation is that the method for identifying the circuit in Experiment 2 was iterative and involved manual analyses. We were able to proceed in this way because of the highly focused nature of the work. But more work would be needed to develop a general procedure for applying it to other tasks and settings.

Acknowledgments

We acknowledge funding from NSF CAREER grant 2339729 to Kyle Mahowald and funding from the Harrington Fellowship to Qing Yao. The work of Tiago Pimentel was funded by the Data Analytics Lab at ETH Zürich. We thank Jed Guzelka-baagac, Felix Sarnthein, Rodrigo Casado, Souvik Banerjee, Jan Tempus, and Philip Whittington for comments on this manuscript.

References

- Guillaume Alain and Yoshua Bengio. 2016. [Understanding intermediate layers using linear classifier probes](#). *Preprint*, arXiv:1610.01644.
- Nura Aljaafari, Danilo Carvalho, and Andre Freitas. 2026. [Emergence and localisation of semantic role circuits in LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 39402–39433, San Diego, California, United States. Association for Computational Linguistics.
- Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. [CausalGym: Benchmarking causal interpretability methods on linguistic tasks](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663, Bangkok, Thailand. Association for Computational Linguistics.
- Aryaman Arora, Zhengxuan Wu, Jacob Steinhardt, and Sarah Schwettmann. 2026. [Language model circuits are sparse in the neuron basis](#). In *Forty-third International Conference on Machine Learning*.
- 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. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.
- Bastian Bunzeck and Sina Zarriß. 2024. [Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly](#). In *Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning*, pages 39–55, Gothenburg, Sweden. Association for Computational Linguistics.
- Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2024. [Characterizing learning curves during language model pre-training: Learning, forgetting, and stability](#). *Transactions of the Association for Computational Linguistics*, 12:1346–1362.
- 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, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- 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*, pages 276–286, Florence, Italy. 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](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. [Amnesic probing: Behavioral explanation with amnesic counterfactuals](#). *Transactions of the Association for Computational Linguistics*, 9:160–175.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. [A mathematical framework for transformer circuits](#). *Transformer Circuits Thread*.
- Finlay Fehlaue, Kyle Mahowald, and Tiago Pimentel. 2025. [Convergence and divergence of language models under different random seeds](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32982–32991, Suzhou, China. Association for Computational Linguistics.
- Javier Ferrando and Marta R. Costa-jussà. 2024. [On the similarity of circuits across languages: a case study on the subject-verb agreement task](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.
- Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. [Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models](#).

- 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 1828–1843, Online. Association for Computational Linguistics.
- Richard Futrell and Kyle Mahowald. 2026. [How linguistics learned to stop worrying and love the language models](#). *Behavioral and Brain Sciences*, 49:e198.
- Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. [Neural language models as psycholinguistic subjects: Representations of syntactic state](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.
- Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. [Finding alignments between interpretable causal variables and distributed neural representations](#). In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.
- Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. [Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.
- Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. [Localizing model behavior with path patching](#). *Preprint*, arXiv:2304.05969.
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. [Colorless green recurrent networks dream hierarchically](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.
- Ahmad Dawar Hakimi, Ali Modarressi, Philipp Wicke, and Hinrich Schuetze. 2025. [Time course MechInterp: Analyzing the evolution of components and knowledge in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12633–12653, Vienna, Austria. Association for Computational Linguistics.
- Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. [How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 76033–76060.
- Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. [Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms](#). In *First Conference on Language Modeling*.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Farnoush Rezaei Jafari, Oliver Eberle, Ashkan Khakzar, and Neel Nanda. 2025. [ReLP: Faithful and efficient circuit discovery in language models via relevance patching](#). *Preprint*, arXiv:2508.21258.
- Sara Kangaslahti, Elan Rosenfeld, and Naomi Saphra. 2026. [Hidden breakthroughs in language model training](#). In *The Fourteenth International Conference on Learning Representations*.
- János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. [AtP*: An efficient and scalable method for localizing LLM behaviour to components](#). *Preprint*, arXiv:2403.00745.
- Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. [Do neural language models show preferences for syntactic formalisms?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091, Online. Association for Computational Linguistics.
- Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. [The emergence of number and syntax units in LSTM language models](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.
- Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. [Probing for the usage of grammatical number](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.

- Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. [Sparse feature circuits: Discovering and editing interpretable causal graphs in language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted syntactic evaluation of language models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. [Pointer sentinel mixture models](#). In *International Conference on Learning Representations*.
- William Merrill, Nikolaos Tsilivis, and Aman Shukla. 2023. [A tale of two circuits: Grokking as competition of sparse and dense subnetworks](#). In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.
- Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. [Cross-linguistic syntactic evaluation of word prediction models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.
- Neel Nanda. 2023. [Attribution patching: Activation patching at industrial scale](#).
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. [Progress measures for grokking via mechanistic interpretability](#). In *The Eleventh International Conference on Learning Representations*.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. [Zoom in: An introduction to circuits](#). *Distill*.
- Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. 2025. [How do LLMs acquire new knowledge? A knowledge circuits perspective on continual pre-training](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19889–19913, Vienna, Austria. Association for Computational Linguistics.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. [Information-theoretic probing for linguistic structure](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.
- Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. [Grokking: Generalization beyond overfitting on small algorithmic datasets](#). *Preprint*, arXiv:2201.02177.
- Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. [A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. [Axiomatic attribution for deep networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.
- Aaquib Syed, Can Rager, and Arthur Conmy. 2024. [Attribution patching outperforms automated circuit discovery](#). In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.
- Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. [LLM circuit analyses are consistent across training and scale](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 40699–40731. Curran Associates, Inc.
- Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem H. Zuidema, and Stella Biderman. 2025. [PolyPythias: Stability and outliers across fifty language model pre-training runs](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.
- Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. 2023. [Explaining grokking through circuit efficiency](#). *Preprint*, arXiv:2309.02390.
- George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. 2025. [Differentiation and specialization of attention heads via the refined local learning coefficient](#). In *The Thirteenth International Conference on Learning Representations*.
- 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*.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. [Frequency effects on syntactic rule learning in transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Computation cost

We use one A40 GPU to run all experiments. For pythia-1b, a ReLP attribution pass over all MLP neurons at one token position, using 400 examples with batch size 4, takes around 40 seconds of GPU time after model loading. Integrated-gradient variants are slower, about 2 minutes per checkpoint-position.

DAS training uses 400 training examples, batch size 4, 5 epochs, and 1-dimensional interventions. In our runs, this takes around 2 minutes per layer. Forward-pass based validation is comparatively cheap: a single necessity or sufficiency evaluation over an evaluation set takes about 1–2 seconds.

B Use of AI assistants

We used AI assistants to aid in developing code for experiments and analyses. We also used AI for summarizing results and suggesting edits.

C Task templates and data

We first introduce our synthetic templates: PP, SUBJECTRC, OBJECTRC, and SIMPLE. Then we discuss extensions of these synthetic templates, both in structure and in lexical items, that are used to further stress-test circuit generalization. Lastly, we show the list of all WikiText (Merity et al., 2017) agreement sentences.

All templates share 16 subject pairs (singular/plural forms of the same lemma). The singular forms are the following: *guard, doctor, farmer, author, officer, athlete, manager, senator, consultant, teacher, customer, pilot, executive, minister, actor, architect*. All subjects, objects, and RC verbs are single tokens under the pythia-1b tokenizer. Base and source sentences always use the singular/plural pair of the same lemma, so they differ only in subject number. We measure model predictions on the verb tokens {*is, was, has*} for singular and {*are, were, have*} for plural; each is a single token.

C.1 Synthetic templates

Below, we detail all synthetic templates used to find and test the circuit. Exact templates and exemplars can be seen in Table 4.

PP. This is the primary template used to find and evaluate our circuit. We draw distractor noun-phrases (NPs) from 16 human-role lemmas, using both singular and plural forms of these distractors.

We select from 6 prepositions (*near, behind, in front of, next to, across from, to the side of*).

OBJECTRC. We use this template to measure generalization. We again draw distractor noun-phrases (NPs) from 16 human-role lemmas, but use only the singular form of these distractors. Within the distractors, we utilize 8 past-tense verbs (*hated, injured, liked, etc.*).

SUBJECTRC. We use this template to measure generalization. Like OBJECTRC, distractor noun-phrases (NPs) are drawn from 16 human-role lemmas with only the singular form, and utilize the same set of 8 past-tense verbs.

SIMPLE. We use this template to measure generalization. We do not sample any distractor NPs for this template as it possesses no distractor, with the position after the subject position being the prediction position.

Adjective insertion. **Adjective-on-subject** inserts a single adjective between the determiner and the subject noun, e.g., *The tall guard near the doctors*. The **adjective-on-object** perturbation instead inserts a single adjective inside the distractor NP, e.g., *The guard near the tall doctors*. There are 16 adjectives total, all applicable to human nouns. They are the following: *tall, old, quiet, busy, new, kind, rich, brave, calm, smart, happy, nice, shy, bold, wise, young*.

Adverb insertion. The **adverb** perturbation places a frequency adverb after the distractor object and before the predicted verb, e.g., *The guard near the doctors often*. Unlike the adjective perturbations, this changes the lexical category of the token immediately before the prediction position. It is therefore a stronger test of whether the downstream readout module is tied to the final distractor NP token used in the original PP template. There are 6 adverbs: *always, often, usually, sometimes, rarely, never*.

Prefix insertion. The prefix perturbations add material before the subject. The **single-prefix** condition uses one of the six neutral discourse markers: *Apparently, Clearly, Surprisingly, Obviously, In fact, For context*; the **locative-prefix** condition uses one of six neutral locative phrases: *In the park, At the office, At the hospital, At the theater, In the garden, On the street*. We avoid tensed prefixes such as *Yesterday*, because they can independently

affect the distribution over verb forms. These perturbations shift the absolute position of the subject and all downstream tokens while leaving the dependency structure unchanged.

Longer distractor. This perturbation inserts two hand-designed adjectives in the object NP, e.g., *The guard near the respected senior doctors*. This lengthens the intervening attractor phrase and increases the distance between subject and prediction. There are 6 such adjective pairs: *experienced local*, *respected senior*, *careful young*, *skilled retired*, *friendly new*, *talented junior*.

C.2 Templates for lexical generalizations

The lexical-generalization experiments keep the PP dependency structure fixed and vary the lexical items used either in the subject slot or in the next-token verb label. In all cases, the base/source pair is still a minimal subject-number pair. Concrete examples are in Table 5.

Held-out-subject. This condition replaces the original subject lexicon with 13 single-token human-role lemmas: *lawyer*, *engineer*, *nurse*, *scientist*, *driver*, *artist*, *musician*, *dancer*, *priest*, *cook*, *singer*, *worker*, *painter*. Both singular and plural forms are single tokens under the pythia-1b tokenizer. The rest of the template is untouched.

Held-out-verb. This condition uses the original PP subject, preposition, and distractor pools, but changes the next-token prediction labels. We evaluate 36 single-token third-person singular/plural verb pairs: *runs/run*, *walks/walk*, *eats/eat*, *goes/go*, *sees/see*, *knows/know*, *says/say*, *gets/get*, *takes/take*, *makes/make*, *comes/come*, *gives/give*, *finds/find*, *tells/tell*, *thinks/think*, *feels/feel*, *writes/write*, *plays/play*, *works/work*, *buys/buy*, *sells/sell*, *hears/hear*, *looks/look*, *seems/seem*, *calls/call*, *means/mean*, *wants/want*, *needs/need*, *loves/love*, *hates/hate*, *lives/live*, *opens/open*, *closes/close*, *starts/start*, *stops/stop*, *waits/wait*.

C.3 WikiText agreement cases

We mined 80 agreement sentences from WikiText, balanced across 40 plural-subject and 40 singular-subject cases. For each sentence we manually created a counterfactual that differs in subject number. The subject, counterfactual subject, and verb continuations tokenize as single tokens. When the attested WikiText verb is past tense, we manually

convert the verb to its present tense and score the singular/plural pair.

For each item below, **b** (base) is the attested WikiText prefix plus its attested verb (converted to present-tense, if necessary), and **s** (source) is the manually constructed opposite-number subject version plus the opposite-number present-tense verb.

Plural-subject cases.

1. **b:** The origins of Imagism are
s: The origin of Imagism is
2. **b:** The characteristics of the sculpture have
s: The characteristic of the sculpture has
3. **b:** The eyes of the monument are
s: The eye of the monument is
4. **b:** The affairs of the CWGC are
s: The affair of the CWGC is
5. **b:** The walls of the nave are
s: The wall of the nave is
6. **b:** The leaders of the Coalition are
s: The leader of the Coalition is
7. **b:** The details of the murder are
s: The detail of the murder is
8. **b:** The reasons behind it are
s: The reason behind it is
9. **b:** The origins of Olaf are
s: The origin of Olaf is
10. **b:** The dimensions of the Tetrarch are
s: The dimension of the Tetrarch is
11. **b:** The objectives of the force are
s: The objective of the force is
12. **b:** The stones of the dolmen are
s: The stone of the dolmen is
13. **b:** The symptoms of AML are
s: The symptom of AML is
14. **b:** The actions of the governor are
s: The action of the governor is
15. **b:** The highlights of the show are
s: The highlight of the show is
16. **b:** The commands of the starter are
s: The command of the starter is

Condition	Template prefix	Example prefix
SIMPLE	The {subject}	<i>The guard</i>
SUBJECTRC	The {subject} that {verbed} the {object}	<i>The guard that liked the doctor</i>
OBJECTRC	The {subject} that the {object} {verbed}	<i>The guard that the doctor liked</i>
PP	The {subject} {prep} the {object}	<i>The guard near the doctors</i>
PP + adjective-on-subject	The {adj} {subject} {prep} the {object}	<i>The tall guard near the doctors</i>
PP + adjective-on-object	The {subject} {prep} the {adj} {object}	<i>The guard near the tall doctors</i>
PP + adverb	The {subject} {prep} the {object} {adverb}	<i>The guard near the doctors often</i>
PP + single-prefix	{prefix} The {subject} {prep} the {object}	<i>Apparently, the guard near the doctors</i>
PP + locative-prefix	{locative-prefix} the {subject} {prep} the {object}	<i>At the office, the guard near the doctors</i>
PP + longer-distractor	The {subject} {prep} the {adj1} {adj2} {object}	<i>The guard near the respected senior doctors</i>

Table 4: Agreement templates used in cross-template and perturbation generalization experiments. The object/distractor pool contains both singular and plural forms for PP-based templates, and base and source examples differ only in the subject number. We evaluate $N = 192$ pairs ($N = 32$ for SIMPLE).

Condition	Template / readout	Example
Held-out subjects	The {held-out-subject} {prep} the {object}; readout is <i>is/are, was/were, has/have</i>	<i>The lawyer near the doctors</i> → <i>is/are</i>
Held-out verbs	The {subject} {prep} the {object}; readout is the held-out verb-pair set	<i>The guard near the doctors</i> → <i>runs/run</i>

Table 5: Lexical-generalization templates. We test a wider range of subject and verb tokens in the PP template to ensure circuit robustness.

17. **b:** The roles of each deity are
s: The role of each deity is
 18. **b:** The effects of antimony are
s: The effect of antimony is
 19. **b:** The plots of such works are
s: The plot of such works is
 20. **b:** The roles of these area commands are
s: The role of these area commands is
 21. **b:** The reasons for going to war are
s: The reason for going to war is
 22. **b:** The chances of rapid intensification are
s: The chance of rapid intensification is
 23. **b:** The origins of the Knights Hospitaller are
s: The origin of the Knights Hospitaller is
 24. **b:** The arrangements for conscientious objectors are
s: The arrangement for conscientious objectors is
 25. **b:** The regions of north County Galway and east County Mayo have
s: The region of north County Galway and east County Mayo has
 26. **b:** The origins of Hurricane Lorenzo are
s: The origin of Hurricane Lorenzo is
 27. **b:** The shows in Japan are
s: The show in Japan is
 28. **b:** The heads in Lisbon are
s: The head in Lisbon is
 29. **b:** The characteristics of its body, head, fins, and teeth are
s: The characteristic of its body, head, fins, and teeth is
 30. **b:** The levels of ionizing radiation on Io, Europa, and Ganymede are
s: The level of ionizing radiation on Io, Europa, and Ganymede is
 31. **b:** The events of the battle are
s: The event of the battle is
 32. **b:** The districts of Somerset are
s: The district of Somerset is
 33. **b:** The corners of the base have
s: The corner of the base has
 34. **b:** The identities of the perpetrator or the instigator are
s: The identity of the perpetrator or the instigator is
 35. **b:** The statues in the west tower are
s: The statue in the west tower is
 36. **b:** The levels in Perfect Dark have
s: The level in Perfect Dark has
 37. **b:** The songs on the album are
s: The song on the album is
 38. **b:** The objectives of the German, Austrian and Ottoman advance are
s: The objective of the German, Austrian and Ottoman advance is
 39. **b:** The tactics employed by the Anzac Mounted Division are
s: The tactic employed by the Anzac Mounted Division is
 40. **b:** The lines about the corners of the eyes and mouth are
s: The line about the corners of the eyes and mouth is
- Singular-subject cases.**
1. **b:** The game of checkers is
s: The games of checkers are
 2. **b:** The influence of the churches is
s: The influences of the churches are
 3. **b:** The strength of the rearguards is
s: The strengths of the rearguards are
 4. **b:** The tail behind the stings has
s: The tails behind the stings have
 5. **b:** The family of four is
s: The families of four are
 6. **b:** The head of a given divine image is
s: The heads of a given divine image are
 7. **b:** The number of pilots licensed by the CAA to fly powered aircraft in 2005 is
s: The numbers of pilots licensed by the CAA to fly powered aircraft in 2005 are
 8. **b:** The objective of regulation is
s: The objectives of regulation are

9. **b:** The goal of the blockade is
s: The goals of the blockade are
10. **b:** The attack on Ancona is
s: The attacks on Ancona are
11. **b:** The objective of the bombardment of Ancona is
s: The objectives of the bombardment of Ancona are
12. **b:** The application of the Johnson-Corey-Chaykovsky reaction in organic synthesis is
s: The applications of the Johnson-Corey-Chaykovsky reaction in organic synthesis are
13. **b:** The tone of these chapters is
s: The tones of these chapters are
14. **b:** The structure of Dominican society is
s: The structures of Dominican society are
15. **b:** The panel to the east of the north portico is
s: The panels to the east of the north portico are
16. **b:** The door of the central shrine has
s: The doors of the central shrine have
17. **b:** The state of Polish primary schools is
s: The states of Polish primary schools are
18. **b:** The inclusion of occupational debris over the bones is
s: The inclusions of occupational debris over the bones are
19. **b:** The use of Sb as the standard chemical symbol for antimony is
s: The uses of Sb as the standard chemical symbol for antimony are
20. **b:** The origin of this is
s: The origins of this are
21. **b:** The formation of halogenated antimony compounds is
s: The formations of halogenated antimony compounds are
22. **b:** The budget of the film is
s: The budgets of the film are
23. **b:** The cost of construction is
s: The costs of construction are
24. **b:** The peak of the roof is
s: The peaks of the roof are
25. **b:** The method and timing of mowing is
s: The methods and timing of mowing are
26. **b:** The influence of weather on chick survival is
s: The influences of weather on chick survival are
27. **b:** The goal of induction therapy is
s: The goals of induction therapy are
28. **b:** The island of Hispaniola, shared by Haiti and the Dominican Republic, is
s: The islands of Hispaniola, shared by Haiti and the Dominican Republic, are
29. **b:** The consequence of failing a puzzle of this sort is
s: The consequences of failing a puzzle of this sort are
30. **b:** The relationship between the tolling points and interchanges along SR 878 and SR 874 is
s: The relationships between the tolling points and interchanges along SR 878 and SR 874 are
31. **b:** The purpose of this plan is
s: The purposes of this plan are
32. **b:** The amount of water in a toilet or sink is
s: The amounts of water in a toilet or sink are
33. **b:** The form of the poem is
s: The forms of the poem are
34. **b:** The rate of protein synthesis is
s: The rates of protein synthesis are
35. **b:** The study of proteins in vivo is
s: The studies of proteins in vivo are
36. **b:** The field of bioinformatics is
s: The fields of bioinformatics are
37. **b:** The task allocated to the British Commonwealth force is
s: The tasks allocated to the British Commonwealth force are
38. **b:** The definition of leg before wicket is
s: The definitions of leg before wicket are
39. **b:** The object on the right seated on legs alongside a box is
s: The objects on the right seated on legs alongside a box are

40. **b**: The song, “a cheery little tune about death, destruction and the end of the world” is
s: The songs, “a cheery little tune about death, destruction and the end of the world” are

D Additional methodological details

D.1 ReLP neuron attribution

We apply ReLP (Jafari et al., 2025; Arora et al., 2026) to form a broad candidate set of MLP neurons before running causal validation. The attribution score is computed over the PP training set ($N = 400$).

Replacement-gradient attribution. For each layer L and neuron n , we capture the post-nonlinearity MLP activation $v_{L,n}$ at the relevant token position and score

$$\text{score}_{L,n} = \mathbb{E}_{(\mathbf{b}, \mathbf{s})} \left[v_{L,n}(\mathbf{b}) \frac{\partial m(\mathbf{b}, \mathbf{s})}{\partial v_{L,n}(\mathbf{b})} \right],$$

where $m(\mathbf{b}, \mathbf{s}) = \log(p(\ell_{\mathbf{b}} | \mathbf{b})) - \log(p(\ell_{\mathbf{s}} | \mathbf{b}))$ is the per-example agreement gap.

Linearizations are architecture specific. For Pythia, we use the following locally linear replacement model: LayerNorm centering and scaling factors are frozen, GELUs are replaced by their local linear gates, and attention softmax probabilities are detached before the value-weighted sum.

There are two implementation choices which deviate from Arora et al. (2026) — both of which we show do not qualitatively affect our results. First, the score in Equation (6) uses the zero-baseline form $v_n \cdot \nabla$, which Arora et al. (2026) adopt in the unpaired setting, rather than the activation-difference form $\Delta v_n \cdot \nabla$ they use with paired data. The former estimates a neuron’s contribution on the base input, while the latter estimates the effect of activation replacement from the counterfactual. Zero-baseline is a computationally simpler method since it does not require a source-activation pass and storage. This choice has little impact on the candidates: comparing both the zero-baseline form $v_n \cdot \nabla$ and the activation-difference form $\Delta v_n \cdot \nabla$, we find that they rank the top 50 neurons identically and overlap on 944 of the top 1024 neurons pooled across positions. This suggests our findings are largely unaffected by the choice of zero-baseline.

Second, we also detach the outputs of all other MLPs when scoring a neuron within a particular layer L , so the gradient reaching L flows only through the residual stream and attention. Arora

Step	Scoring	Percent Recovered			
		Full pool		Top-16	
		Nec	Suff	Nec	Suff
2k	Direct (used)	89.4	92.7	12.2	17.8
	Total effect	88.8	92.5	10.8	16.2
143k	Direct (used)	97.6	98.7	51.7	64.6
	Total effect	94.9	94.9	54.9	66.3

Table 6: Sparsification is robust to the stop gradient implementation. All values are percentages on the PP evaluation set. “Full pool” is the necessity/sufficiency of the 1024-neuron circuit with respective attention heads, and “Top-16” is the fraction of the gain from the head-only baseline to the full pool captured by the 16 highest-ranked neurons.

et al. (2026) introduce this for edge attribution, but we also apply it to node scoring here. There is no clear unbiased choice here: our approach will underweight neurons whose contributions depend on downstream MLP processing, while the alternative can give credit to both upstream and downstream neurons that mediate the same effect.

We compared the two approaches and found that the broader ranking changes, but the conclusion does not. We recomputed the top neuron candidates for pythia-1b with the stop-gradient removed (i.e., total-effect scoring, where the gradient can flow through downstream MLPs). The two methods overlap in 42 of the top 50 ranking neurons, though this overlap falls to 356 at top-1024. However, the final circuit’s 18 neurons are all within the top-35 in both rankings, and the sparsification claim of Section 4.2 does not change.

Table 6 shows that the portion of necessity or sufficiency recovered from adding the top-16 neurons to the head-baseline remains low early in training and high later in training, regardless of direct- or total-effect scoring.

Integrated-gradient check. As a robustness check during candidate generation, we also compute integrated gradients on MLP activations (Sundararajan et al., 2017; Marks et al., 2025). This method interpolates the captured MLP activation from the counterfactual source activation to the base activation, keeps all other activations on the base run, and averages gradients along a k -step path. The resulting score is

$$(v_{L,n}^{\text{base}} - v_{L,n}^{\text{src}}) \cdot \frac{1}{k} \sum_{i=1}^k \frac{\partial m}{\partial v_{L,n}^{(i)}}.$$

The integrated-gradient scores are used to check that high-ranked candidates are not biased by the local linearizations.

Using $k = 8$, we find that at the final checkpoint of pythia-1b, all 18 neurons in the final circuit fall within the top 35 neurons under both pooled RelP rankings and integrated gradients, and the two agree on 44 of the top 50 neurons.

D.2 Neuron-head routing attribution

At each checkpoint, we consider the top-32 positive RelP neurons at the subject, preposition, and prediction positions. In Section 5.1, we ask whether these neurons route through the early L15 H7 pathway, the later L13 H7 pathway, a mixture of the two, or neither.

Our detailed methodology is as follows. For an upstream neuron n , we first compute its sufficiency in isolation. Next, we repeat the intervention, but this time hold one downstream head h at its base value (for the edge between the neuron’s position and the prediction position). This manipulation blocks n ’s effect from propagating through h . We denote the neuron’s *routed sufficiency* via h as:

$$\text{RSuff}_{n \rightarrow h} = \text{Suff}(n) - \text{Suff}(n; \text{block}(h)).$$

The routed sufficiency value denotes the portion of the neuron’s effect that disappears once h is blocked.

Similarly, for a downstream neuron, n , and an upstream head, h , we first intervene on the head, then repeat the intervention whilst restoring the neuron n ’s activation.

$$\text{RSuff}_{h \rightarrow n} = \text{Suff}(h) - \text{Suff}(h; \text{restore}(n)).$$

This measures how much the head intervention effect is mediated through n .

Visualization. We now explain how we colored neurons in Figure 4. These computations and thresholds are purely for visualization purposes.

We first identify the neurons on either head’s route. This takes different forms for upstream and downstream neurons. For an upstream neuron to qualify, we jointly block both heads, and require the neuron to lose at least 5% of the standalone sufficiency. For a downstream neuron, we require restoring the activation to remove at least 5% of the joint sufficiency from an intervention on both heads. If a neuron does not meet these thresholds, it is considered to not be on either head’s route.

For the neurons on either head’s routes, we attribute their colors as follows. We first define

$$r_{13}(n) = \begin{cases} \text{RSuff}_{n \rightarrow 13}, & n \text{ is upstream,} \\ \text{RSuff}_{13 \rightarrow n}, & n \text{ is downstream,} \end{cases}$$

and define $r_{15}(n)$ likewise. This allows us to define positive routing shares

$$s_{13}(n) = \frac{[r_{13}(n)]_+}{[r_{13}(n)]_+ + [r_{15}(n)]_+},$$

$$s_{15}(n) = 1 - s_{13}(n),$$

where $[x]_+$ denotes $\max(x, 0)$.

This metric gives us a gradient between how much a neuron is L15 H7-routed or L13 H7-routed. Neurons are duly colored by this gradient: blue when $s_{15} = 1$, purple when the two shares are equal, and red when $s_{13} = 1$.

E Experiment 1: Additional results

E.1 Different levels of k

We also produce Figure 2 at different threshold levels. Define k_{30} and k_{70} similarly to eq. 7, we reproduce Figure 2 in Figure 5 and Figure 6, respectively.

We see that broadly speaking, the number of neurons in these sets decrease over training. For k_{70} in pythia-1b, we see a rebound at step 143k. This is likely due to the final circuit of pythia-1b relying on downstream prediction-position readout neurons, which we find in Experiment 2.

We take these results to be consistent with the trend that causal effects of subject neurons becomes more concentrated over training. This notion of sparsification may not always be reflected in our figures, and building a complete circuit requires examining model- and checkpoint-specific mechanisms beyond the subject position.

E.2 Overlap of neuron sets across training

For each checkpoint, we report each model’s k_{50} set size, and report the overlap between consecutive evaluated checkpoints’ k_{50} sets. Since this set does not monotonically decrease, we “normalize” by also showing the smaller size of two consecutive sets.

Our results are in Table 7. We see that generally, latter checkpoint’s neuron sets are largely subsets of the previous checkpoint’s neuron sets. In the rare case that k_{50} counts expand (possibly due to instabilities in how they are estimated), the

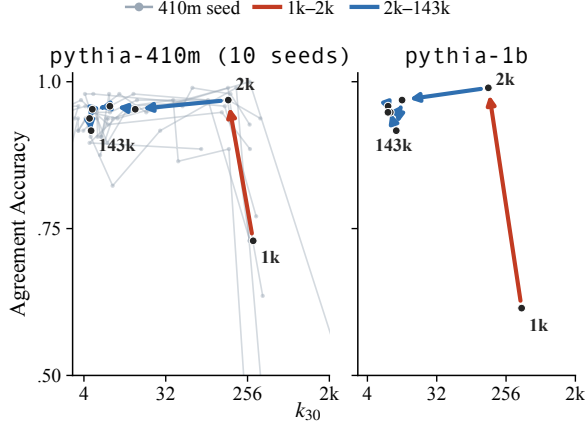


Figure 5: Agreement accuracy vs. k_{30} over training checkpoints.

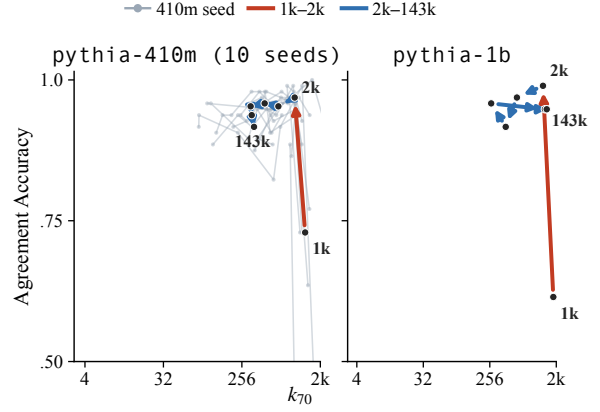


Figure 6: Agreement accuracy vs. k_{70} over training checkpoints.

Run	2k	5k	8k	18k	50k	143k
Seed 0	766	152 [108/152]	55 [53/55]	15 [15/15]	14 [11/14]	12 [12/12]
Seed 1	462	170 [102/170]	85 [64/85]	22 [22/22]	12 [11/12]	6 [6/6]
Seed 2	345	153 [81/153]	39 [36/39]	29 [28/29]	23 [21/23]	48 [23/23]
Seed 5	477	319 [136/319]	311 [182/311]	178 [126/178]	58 [50/58]	29 [29/29]
Seed 6	569	198 [118/198]	109 [95/109]	15 [15/15]	9 [9/9]	8 [8/8]
Seed 7	473	90 [66/90]	29 [28/29]	17 [17/17]	16 [14/16]	14 [13/14]
Seed 8	467	114 [73/114]	42 [40/42]	14 [12/14]	15 [12/14]	14 [14/14]
Seed 9	472	154 [91/154]	47 [45/47]	24 [24/24]	36 [22/24]	58 [34/36]
Seed 3 (outlier)	491	68 [52/68]	35 [35/35]	14 [14/14]	12 [10/12]	30 [0/12]
Seed 4 (outlier)	446	357 [119/357]	349 [183/349]	160 [108/160]	35 [35/35]	255 [1/35]
pythia-1b	555	83 [70/83]	42 [40/42]	21 [21/21]	41 [20/21]	31 [29/31]

Table 7: Size and overlap of the k_{50} neuron sets across checkpoints. The first number is the k_{50} size, the numerator of the fraction is the overlap between the checkpoint’s and the previous checkpoint’s k_{50} set, and the denominator of the fraction is the size of the smaller k_{50} set of the two.

neurons belonging to the smaller set are largely retained. Outliers show a qualitatively different pattern entirely, with abrupt changes in their neuron set memberships.

E.3 Late-training outlier trajectories

The pattern of models first learning behavior before the internal mechanism sparsifies seems to hold across non-outlier seeds. However, the outlier seeds are interestingly different — Figure 7 shows that at exactly the training setbacks reported by [van der Wal et al. \(2025\)](#), the outlier models’ k_{50} scores rebound. For seed 3, the setback is not on agreement behavior but only on the mechanism, so subsequent training simply re-sparsifies its neuron mechanism. However, seed 4 loses a significant portion of its agreement abilities. As such, in the remaining training steps, it relearns agreement, but runs out of time for re-sparsification.

Taken together, these results suggest that there

is a general pressure to sparsify after agreement is learned, even for outlier runs.

E.4 Attention-head localization across checkpoints and seeds

To identify candidate transport heads, we run a scan on all heads at the prediction position at each checkpoint, testing for both necessity and sufficiency. We rank heads by $\min(\text{Nec}, \text{Suff})$.

For pythia-1b, we show the top-5 heads at each checkpoint in Table 8, bolding the three heads in the emergence analysis (L13 H7, L15 H7, and L15 H2). We omit early checkpoints because normalizing against a near-zero agreement gap for necessity/sufficiency is unstable. Across the ten seeds, the dominant channel at the end of training is between layers 16–23 (the last layer has index 23). On average, the head settles at a depth of 83% relative to the total number of layers, comparable to $13/16 \approx 0.81$ in pythia-1b. This suggests that

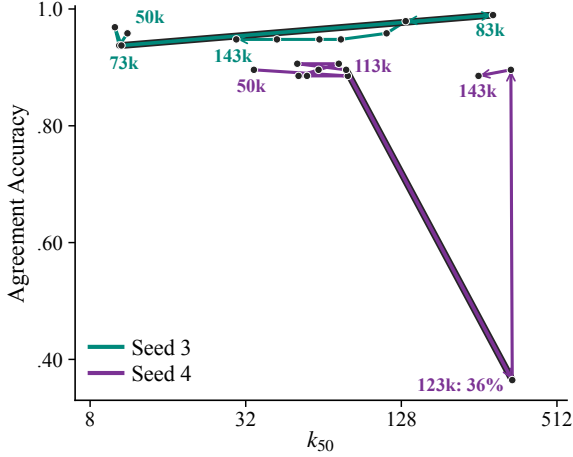


Figure 7: k_{50} rebound co-occurs with training regressions. Seed 3 rebound happens early and the model recovers a sparse neuron mechanism by end of training. Seed 4 suffered setbacks late and its k_{50} ended high.

the late-layer placement of the transport head may not be incidental.

We similarly track the top heads across seeds of pythia-410m by $\min(\text{Nec}, \text{Suff})$, with the tracked heads and results in Table 9. The composition of the top-three head set changes in every run, suggesting that the primary transport channel is generally unstable across training. The single most dominant head changes in both pythia-1b and six seeds of pythia-410m. Among such runs, in three seeds the transport head moves from a later layer toward a middle layer, echoing the pythia-1b pattern. Two seeds, 4 and 6, shift toward a later layer.

We leave for future studies to understand the learning pressures that drive these dynamics.

F Experiment 2: Additional results

F.1 Initial circuit discovery details

The neuron candidates for our final circuit were those which write toward the next-layer number representation and those with high ReLP scores. Here we provide some details on how we identified number representations in the model, and how we pruned to obtain the final circuit.

Distributed alignment search for number. We use Distributed Alignment Search (DAS; Geiger et al., 2024) to test whether the agreement information carried by L13 H7 is concentrated in a low-dimensional subspace. To this end, we train DAS, with the default orthogonal alignment map set at one dimension, on the PP training set with learning rate $5e-3$ and batch size 4, for 5 epochs. We

Checkpoint	Head	$\min\{\text{Nec}, \text{Suff}\}$
1k	L15 H2	48.4%
	L15 H7	42.5%
	L15 H0	8.2%
	L15 H1	7.8%
	L9 H6	1.7%
2k	L15 H7	49.4%
	L15 H2	30.7%
	L15 H0	7.2%
	L9 H6	3.6%
	L7 H6	3.5%
5k	L15 H7	51.2%
	L15 H2	24.2%
	L13 H7	11.5%
	L15 H0	2.3%
	L7 H6	1.7%
8k	L15 H7	42.8%
	L13 H7	21.6%
	L15 H2	18.0%
	L7 H6	2.6%
	L15 H0	1.3%
18k	L13 H7	41.1%
	L15 H7	24.5%
	L15 H2	15.1%
	L7 H6	5.3%
	L13 H0	2.7%
50k	L13 H7	54.9%
	L15 H2	13.9%
	L15 H7	8.5%
	L7 H6	8.0%
	L13 H0	3.5%
143k	L13 H7	67.1%
	L15 H2	14.2%
	L7 H6	6.4%
	L15 H7	3.9%
	L13 H4	2.7%

Table 8: Top five attention heads at each checkpoint scored by $\min(\text{Nec}, \text{Suff})$. The three main heads are bolded.

Run	Step 2k		Step 143k		Movement
	Leader	Score	Leader	Score	
Seed 0	L22 H11	69.9%	L22 H11	65.2%	Stable
Seed 1	L19 H15	89.9%	L19 H15	80.7%	Stable
Seed 2	L23 H10	49.4%	L21 H9	33.9%	Earlier
Seed 5	L20 H15	80.4%	L20 H15	87.6%	Stable
Seed 6	L21 H6	42.1%	L22 H4	56.2%	Later
Seed 7	L17 H3	68.6%	L17 H3	50.9%	Stable
Seed 8	L22 H7	85.0%	L16 H13	48.6%	Earlier
Seed 9	L23 H1	58.6%	L16 H14	32.4%	Earlier
Seed 3 (outlier)	L23 H11	55.6%	L23 H0	52.5%	Same-layer
Seed 4 (outlier)	L17 H8	93.7%	L22 H6	71.3%	Later

Table 9: Leading head across ten seeds of pythia-410m at step 2k and step 143k. Score denotes $\min(\text{Nec}, \text{Suff})$. There is no predictable pattern in how the models reorganize the transport channels.

train at every layer in the subject and prediction positions with five initialization seeds. To score a neuron’s output alignment with one representative number-feature subspace, we use the average of the five subspaces after normalization.

Pruning. Neuron candidates surfaced by output-alignment or RelP are not necessarily causally important, and it is possible that two candidates can exhibit interaction effects that lower their joint effects. For these reasons, after constructing an initial circuit with promising neurons, we apply leave-one-out validation to prune the circuit.

On the 21-neuron circuit, we remove each remaining neuron and re-evaluate the circuit on the PP training set. We remove the neuron whose deletion had the least effect on circuit necessity and sufficiency. This applied to both neurons with positive and negative agreement effects, as we want to keep strong anti-agreement neurons to better represent the model’s mechanism. We repeated this process three times to remove three neurons.

F.2 Circuit scope

Our final circuit is presented as a sparse representation of how pythia-1b carries the number feature from the subject position (from L1 onwards) to the prediction position via L13 H7. To this end, we ignore pos_0 and pos_3 , exclude the L0 MLP from the scope, and ignore the effects of self-attention. We provide justification for these decisions below.

Position selection. We first show that we can safely ignore neurons at pos_0 and pos_3 (which are both the token “the”) by confirming that no relevant heads transfer information from those positions, and that their RelP attributions are low.

To this end, we first compute each head’s sufficiency on the PP training set at each checkpoint.

Then, we compute the sufficiency from intervening on the head, but restricted only to the attention between a source token position and the prediction position. The fraction of the edge-restricted sufficiency relative to the full-head sufficiency serves as a measure for the degree of importance for this position.

We show this metric across heads and checkpoints in Table 10. Position-specific edge patching shows that the early L15 transport heads are almost entirely subject-to-prediction channels. By contrast, L13 H7 starts as a weak subject edge but becomes the dominant late channel, reading both from the subject edge and from the preposition edge at 143k. The determiner positions are not accessed by any of these heads.

Table 11 further shows the maximum absolute RelP score over all MLP neurons in all layers, demonstrating that the estimated effect of neurons at determiner positions pos_0 and pos_3 are negligible. This result provides further confirmation that excluding these positions has negligible impact on the final circuit.

The L0 MLP. We find ablating the complement of the circuit is unusually destructive at L0. That is, we compare full MLP mean ablation at the subject position for L0 and L5, using the PP subject-position mean as the replacement and measure cross-entropy of the agreement target at the prediction position. Our results can be seen in Table 12. On both PP templates and hand-curated WikiText agreement prefixes, L0 mean ablation raises CE by about 0.82 nats, while L5 mean ablation leaves CE essentially unchanged. For this reason, we only present our circuit relative to a subject-to-prediction path that excludes subject L0. Further, these results suggest that L0 may be responsible

Checkpoint	Head	subject $\rightarrow$ pred.	prep. $\rightarrow$ pred.	det. $\rightarrow$ pred.	object $\rightarrow$ pred.
256	L15 H7	0.006	0.001	0.000	0.001
256	L15 H2	-0.012	-0.002	0.000	-0.002
256	L13 H7	0.009	0.002	0.000	0.000
1k	L15 H7	0.391	-0.004	0.000	-0.004
1k	L15 H2	0.446	0.000	0.000	0.000
1k	L13 H7	0.006	0.000	0.000	0.000
2k	L15 H7	0.497	0.000	0.000	-0.001
2k	L15 H2	0.312	0.000	0.000	0.000
2k	L13 H7	0.017	0.001	0.000	0.000
5k	L15 H7	0.505	0.002	0.000	0.000
5k	L15 H2	0.235	0.000	0.000	0.000
5k	L13 H7	0.101	0.005	0.000	0.000
8k	L15 H7	0.426	0.001	0.000	0.002
8k	L15 H2	0.175	0.000	0.000	0.001
8k	L13 H7	0.186	0.009	0.000	0.002
18k	L15 H7	0.242	0.001	0.000	0.002
18k	L15 H2	0.141	0.002	0.000	0.001
18k	L13 H7	0.324	0.041	0.000	0.005
50k	L15 H7	0.080	0.001	0.000	0.001
50k	L15 H2	0.120	0.002	0.000	0.005
50k	L13 H7	0.360	0.087	0.001	0.020
70k	L15 H7	0.059	0.001	0.000	0.001
70k	L15 H2	0.134	0.004	0.000	0.009
70k	L13 H7	0.363	0.130	0.003	0.016
143k	L15 H7	0.039	0.001	0.000	0.002
143k	L15 H2	0.110	0.005	0.000	0.019
143k	L13 H7	0.380	0.184	0.002	0.016

Table 10: Position-specific edge interventions into the prediction position. Entries are fractions of the full head sufficiency recovered by the single-edge intervention. L15 H7 and L15 H2 are concentrated on subject-to-prediction edges, while L13 H7 is a strong subject edge that increasingly uses the preposition-to-prediction edge late in training.

Checkpoint	pos ₀ initial <i>The</i>	pos ₁ subject	pos ₂ prep.	pos ₃ distractor <i>the</i>	pos ₄ object/pred.
256	0.00004	0.00025	0.00004	0.00004	0.00031
1k	0.00069	0.00598	0.00038	0.00014	0.00135
2k	0.00064	0.05164	0.00150	0.00051	0.01050
5k	0.00093	0.22402	0.00680	0.00200	0.07530
8k	0.00114	0.28259	0.01666	0.00528	0.15537
18k	0.00392	0.27067	0.06123	0.01214	0.25152
50k	0.00274	0.33274	0.11713	0.01787	0.47838
70k	0.00252	0.31816	0.13967	0.02424	0.47499
143k	0.00353	0.39645	0.23543	0.03269	0.68888

Table 11: Maximum absolute ReIP score by PP template token position across checkpoints. Largest ReIP scores occur at the subject and prediction-side positions after agreement emerges, while the sentence-initial determiner and the distractor determiner remain small throughout.

Dataset	Ablated	clean CE	ablated CE	Δ CE
PP	L0	4.05	4.86	+0.82
PP	L5	4.05	4.03	−0.01
WikiText	L0	2.08	2.90	+0.83
WikiText	L5	2.08	2.07	−0.01

Table 12: Cross-entropy impact of full-MLP mean ablation at the subject position. CE is measured on the correct agreement target at the prediction position. PP evaluation set ($N = 192$), WikiText ($N = 80$).

Module	Components
$\mathcal{C}_A$ (pos_1)	L1 n7988, L5 n3793, L8 n6308, L9 n5060, L9 n3101, L10 n747, L10 n6267, L11 n6624, L11 n239, L12 n1535
$\mathcal{C}_H$	L13 H7
$\mathcal{C}_P$ (pos_4)	L14 n1160 (pl); L14 n4146, L14 n6888 (sg)
$\mathcal{C}_N$ (pos_4)	L14 n6221, L15 n2511, L15 n3382, L15 n6710 (pl); L15 n2305 (sg)

Table 13: Components of the subject–verb agreement circuit $\mathcal{C}_{\text{full}}$ in pythia-1b. $\mathcal{C}_H$ transports the number feature, and $\mathcal{C}_P$ and $\mathcal{C}_N$ neurons promote or suppress the labeled verb forms.

for the *construction* of the number feature, with our circuit merely responsible for its *transport* to prediction.

Subject-side self-attention makes little contribution to agreement. We find that mean ablating of each layer and head demonstrates that the self-attention contribution at the subject position, $\alpha_h[\text{pos}_1, \text{pos}_1]W_O^h V_h(x_{\text{pos}_1})$ is negligible. In particular, we replace such values with the corresponding dataset mean while leaving non-self attention edges untouched. Across the PP dataset this procedure changes the agreement gap from +3.48 to +3.43, a drop of only 1.5%. This supports ignoring attention heads at the subject position. At the prediction position, no single self-attention head under ablation causes a drop of more than 2.5% of the baseline (L9 H7), and heads L14 or later have near-zero mean ablation effects.

F.3 Circuit components and functional roles

A full list of circuit components is in Table 13, consisting of four modules: $\mathcal{C}_A$, $\mathcal{C}_H$, $\mathcal{C}_P$, and $\mathcal{C}_N$. Here we describe our interpretation of each module’s contribution to the final circuit, as well as the relevant metrics supporting such interpretations.

Neuron	sg mean	pl mean	non-noun mean
L1 n7988	−0.117	1.796	−0.102
L5 n3793	−0.102	3.307	−0.105
L8 n6308	−0.091	2.831	−0.056
L9 n5060	−0.108	3.552	−0.090
L9 n3101	−0.085	1.190	−0.068
L10 n747	−0.106	1.466	−0.049
L10 n6267	−0.117	2.562	−0.018
L11 n6624	−0.121	1.637	−0.016
L11 n239	−0.103	1.008	−0.077
L12 n1535	−0.082	2.873	0.006

Table 14: Activations for $\mathcal{C}_A$ at the subject position. Means are post-activation MLP neuron values. The selected neurons distinguish plural nouns from singular nouns, while singular nouns remain close to non-nouns.

Amplifiers. The upstream neurons in $\mathcal{C}_A$ behave primarily as plural detectors. To demonstrate this, we first hand-wrote 95 singular/plural noun pairs, and then used AI assistance to extract words ending in -s in the first 5000 rows of WikiText, giving us a total of 447 singular/plural noun pairs. We also manually generate 144 non-nouns which are single-token verbs, adjectives, adverbs, and function words. We measure all MLP neurons’ activations at the subject position for these words. To mirror the template, activations are extracted in the context $\langle | \text{endof text} | \rangle \text{The } \{\text{word}\}$.

The $\mathcal{C}_A$ neurons activate similarly, near 0 on singular nouns and non-nouns, but activate strongly on plural nouns. Most of the selected $\mathcal{C}_A$ circuit neurons are among the top one or two plural-selective neurons (i.e., largest mean-activation difference between plural and singular inputs) in their layer: L1 n7988, L5 n3793, L8 n6308, L9 n5060, L11 n6624, and L12 n1535 are ranked first by plural-selectivity in their respective layers; L9 n3101, L10 n6267, L10 n747, and L11 n239 are also highly ranked. See Table 14.

Transport. Results from Experiment 1 at the final checkpoint tell us that L13 H7 is the most important head for agreement. Further, App. F.6 shows that we can localize the number channel to a 1D subspace within the head. Taken together, these results characterize $\mathcal{C}_H$ as the transport channel for the circuit, moving relevant information from the subject position to the prediction position.

Readout: pro- and anti-agreement neurons. The prediction-side neurons are readout components. They are gated by subject number and write directly toward or away from the correct verb-number direction. The two strongest pro-readouts,

Neuron	module	sg mean	pl mean
L14 n1160	C_P plural	-0.06	2.88
L14 n4146	C_P singular	1.67	-0.05
L14 n6888	C_P singular	0.95	-0.11
L14 n6221	C_N plural-gated	-0.05	0.69
L15 n2511	C_N plural-gated	-0.05	1.36
L15 n3382	C_N plural-gated	-0.08	1.22
L15 n6710	C_N plural-gated	-0.07	0.77
L15 n2305	C_N singular-gated	1.46	-0.10

Table 15: Readout neuron activations at the prediction position. The pro- and anti-readouts are gated by the number state reaching the prediction position: plural-side neurons activate on plural-subject examples, while singular-side neurons activate on singular-subject examples. Each mean is computed over 96 singular-subject and 96 plural-subject base sentences in the PP evaluation set ($N = 192$).

L14 n1160 and L14 n4146, have opposite output alignments: L14 n1160 promotes plural verbs such as *are*, *were*, and *have*, while suppressing singulars such as *is*, *was*, and *has*; L14 n4146 has the opposite alignment.

Their activations also appear to be gated by the upstream subject number: L14 n1160 is active on plural-subject examples and nearly inactive on singular-subject examples, while L14 n4146 is active on singular-subject examples and nearly inactive on plural-subject examples. L14 n6888 is an additional singular-side readout with the same gating pattern. See Table 15.

The anti-readouts are active components of the same computation but write in the opposite direction from the pro-readouts. For a baseline agreement gap of 3.48, individually mean ablating anti-agreement neurons increases the gap by 2.7% – 7.0%. Mean ablating all five increases the gap 24.7% relative to baseline ($N = 192$).

E.4 Off-circuit interventions

As an additional robustness test, we intervene on non-circuit MLP activations within the declared scope. Here, we intervene on all MLP blocks within the scope (L1–L12 MLPs at pos_1 and L14–L15 MLPs at pos_4) except the circuit neurons.

For each base sentence b in the PP evaluation set, we intervene on b with 20 randomly sampled donor PP sentences and measure the agreement gap. If the complement does not participate in agreement, the post-intervention agreement gap should also show little to no sensitivity to the donor’s number. Indeed, Table 16 shows that the circuit-complement

Donor	all	sg	pl
Baseline	+3.48	+3.61	+3.35
Same-number donor	+3.68	+3.71	+3.66
Cross-number donor	+3.54	+3.59	+3.48

Table 16: Agreement gap under off-circuit intervention ($N = 192$, $N_{\text{donors}} = 20$ each; 1924 same-class and 1916 cross-class trials). With the circuit untouched, replacing relevant non-circuit activation with a donor’s activation leaves agreement intact regardless of donor class.

Base	Donor			
	PP	SUBJRC	OBJRC	SIMPLE
PP	83%	98%	95%	60%
SUBJRC	79%	90%	88%	58%
OBJRC	72%	81%	80%	53%
SIMPLE	87%	94%	93%	78%

Table 17: Sufficiency from cross-template intervention. Rows are the base template; columns are the donor template (each pair drawn at random, with the donor’s subject opposite-number from base’s). Donors are drawn to have opposite number to each base sentence. $N = 32$ base examples for SIMPLE and $N = 192$ otherwise, balanced across sg/pl-base.

has little effect on agreement under intervention.

E.5 Generalization datasets and additional results.

In the main text, we have shown that the circuit generalizes to other synthetic templates and naturalistic sentences. We further show that the final circuit is a robust agreement mechanism by showing that the circuit (1) can be used to control agreement across templates, (2) applies to most perturbations to the PP template, with the exception of adding an adverb before predicting the next verb, and (3) applies to subjects and verbs in general.

Circuit activations are transferable. The circuit controls agreement across templates, so we expect that activations on the modules are transferable across templates. We test this by intervening on the circuit C_{full} across templates, and measuring sufficiency.

For each pair of templates, we create $N = 192$ pairs ($N = 32$ for SIMPLE) differing in subject number drawn from their evaluation sets. We intervene on circuit components, and measure sufficiency. Table 17 contains our results.

We see that interventions across templates achieve comparable sufficiency to interventions

Perturbation	G^b	$G^{C \leftarrow s}$	G^s	Suff.
None	+3.48	-2.30	-3.48	83%
Adj on subj	+3.26	-2.18	-3.29	83%
Adj on obj	+3.08	-2.34	-3.19	86%
Adv after obj	+3.70	+0.56	-3.87	42%
Single prefix	+3.00	-1.88	-3.00	81%
Loc. prefix	+4.33	-3.02	-4.31	85%
Long distract	+3.11	-2.43	-3.17	88%

Table 18: On-circuit intervention on the perturbed PP template ($N = 192$). The circuit applies to all settings apart from adding an adverb after the distractor object.

Setting	G^b	$G^{C \leftarrow s}$	G^s	Suff.
Baseline	+3.48	-2.32	-3.48	83%
Held-out subjs	+3.69	-2.26	-3.58	82%
Held-out verbs	+2.19	-1.51	-2.16	85%

Table 19: The circuit generalizes on the PP template with held-out subjects or with held-out verbs. $N = 192$ pairs per condition. Held-out subjects experiment samples from 13 novel subjects. Held-out verb experiment averages continuation probabilities across 36 novel verb singular/plural pairs.

within templates. The SIMPLE template is the odd one out: circuit activations on other templates transfer into the SIMPLE template effectively, but not vice versa.

Generalization to perturbations. We test whether the circuit is sensitive to absolute positions of the subject and prediction tokens by considering 6 perturbations of the PP template described in App. C.1: adjective on subject, adjective on object, adverb insertion, clause-initial prefix, locative prefix, and longer distractor. For each perturbation, we apply interventions on C_{full} across pairs differing in subject number and report sufficiency. Results can be seen in Table 18. We find that the circuit is sufficient for most perturbations, staying within 81% – 88% range, compared to the PP template’s 83%. Adding an adverb after the object is the exception, which suggests that the circuit readout components may be specific to predictions after NPs. Follow-up work could test this by holding constant the circuit components before the readout position and searching for relevant readout neurons.

Generalization to novel words. So far, we have evaluated our circuit on a limited set of 16 subjects and objects, and measured probabilities on only six verb forms. Here, we show that the circuit

is not only general across templates and perturbations, but robust to the choice of lexical items in the templates.

For this test, we create 13 single-token singular/plural subject pairs not in the original 16-pair lexicon, and 36 single-token singular/plural verb pairs. Examples of these are shown in App. C.2.

We then intervene on C_{full} , and find that it indeed applies to novel subjects and verbs (Table 19). The lower baseline gap in held-out verbs reflects the model’s weaker preference for less common verb completions, but the intervention still brings the gap 85% of the way to the natural target.

F.6 DAS analysis of the transport channel

Training details. We use a rotation alignment map with 1D intervention. The DAS intervention is applied to L13 H7’s value-output channel at the prediction position. In pythia-1b, this is the 256-dimensional head slice inside the attention value output for L13 H7. Training uses the PP training set with batch size 4, 5 epochs, Adam optimizer with learning rate $5 \cdot 10^{-3}$, and three random seeds.

Ablation evaluation. In Section 5, we demonstrate the sufficiency of the 1D channel for controlling agreement behavior. In this section, we further evaluate mean ablation variants on the PP evaluation set. We first compute the mean L13 H7 pre-output-projection activation at the prediction position over the base examples. Let $x \in \mathbb{R}^{256}$ be the natural L13 H7 slice for an example, let $\bar{x}$ be the mean slice, and let v be the normalized 1D DAS direction.

For necessity, we compare three ablations:

$$\text{full: } x' = \bar{x}, \quad (8)$$

$$\text{1D channel: } x' = x + (\langle \bar{x}, v \rangle - \langle x, v \rangle)v, \quad (9)$$

$$\text{complement: } x' = \bar{x} + (\langle x, v \rangle - \langle \bar{x}, v \rangle)v. \quad (10)$$

The full ablation replaces the entire 256-dimensional head slice with its mean. The 1D-channel ablation replaces only the DAS direction component with its mean value and preserves the orthogonal complement. The complement ablation does the reverse, mean ablating the orthogonal complement while preserving the natural DAS-direction component.