
Circuit-tuning: A Mechanistic Approach for Identifying Parameter Redundancy and Fine-tuning Neural Networks

Anonymous Author(s)

Affiliation

Address

email

Abstract

The study of mechanistic interpretability aims to reverse-engineer a model to explain its behaviors. While recent studies have focused on the static mechanism of a certain behavior, the learning dynamics inside a model remain to be explored. In this work, we develop an interpretable fine-tuning method for analyzing the mechanism behind learning. We first introduce the concept of node-level intrinsic dimensionality to describe the learning process of a model in a computational graph. Based on our theory, we propose circuit-tuning, a two-stage algorithm that iteratively builds the minimal subgraph for a specific task and updates the key parameters in a heuristic way. Experimental results confirm the existence of the intrinsic dimensionality at the node level and demonstrate the effectiveness of our method for transparent and interpretable fine-tuning. We visualize and analyze the circuits before, during, and after fine-tuning, providing new insights into the self-organization mechanism of a neural network in the learning process.

1 Introduction

Transformer-based large language models (LLMs) have demonstrated outstanding performance in a wide range of tasks [61]. However, an LLM is often treated as a "black box" because of its complex inner mechanisms, which brings a lot of issues about AI safety and reliability, highlighting the need for interpretability [24]. Mechanistic interpretability aims to discover the underlying mechanisms inside a model so as to provide better control and improved design of it [53], showing the potential of reverse-engineering a model. Recent studies in this field include analyzing the circuit responsible for a single behavior [41, 62], extracting features via sparse dictionary learning [5, 57], applying a steering vector [60] to modify the model behaviors, etc.

Despite the success of the above methods, they are limited to post-hoc analyses of a trained model. For example, Wang et al. [62] and Hanna et al. [18] studied the interactions of the attention heads / MLPs and discovered the circuits for indirect object identification and mathematics in GPT-2-small. This kind of static analysis of model behaviors during inference fails to explain the learning dynamics of a model, i.e., how a model acquire an ability, or generalize to various scenarios, which is of great importance in mechanistic interpretability [53]. Recently, Nanda et al. [38], Olsson et al. [42] studied the phase transition during training. Prakash et al. [46], Lee et al. [28], Jain et al. [23], Wang et al. [63] focused on narrow or synthetic tasks for fine-tuning. Bricken et al. [4] studied the change in sleeper agent features via stage-wise model diffing. While these works focus on specific scenarios, there remain **two limitations**: (1) there lacks a general and unified way or pipeline for interpreting the learning process that is applicable to various scenarios from the view of mechanistic interpretability; (2) current studies mostly focus on post-hoc interpretations of fine-tuning, without daring to provide guidance for a more precise and effective fine-tuning process with interpretability. These limitations

36 make it difficult for interpretability works to be well applied in practice to more general settings, or
37 to provide more inspiration for traceable and steerable training.

38 To solve the above two concerns, we introduce a general and interpretable fine-tuning method, making
39 it possible for transparent training. Specifically, inspired by prior works on intrinsic dimensionality,
40 we view the learning process inside a model as *finding and optimizing the minimum subgraph*
41 *corresponding to the intrinsic dimensions in a computational graph*. Previous estimation of intrinsic
42 dimensionality was conducted directly in the parameter space, while the research on mechanistic
43 interpretability usually adopts a representational perspective. To bridge these two paradigms, we
44 introduce the *node-level intrinsic dimensionality* for the first time, enabling the study of learning
45 dynamics with intrinsic dimensionality in the representation space.

46 Based on our theory, we propose *circuit-tuning*, a method that performs interpretable fine-tuning in a
47 computational graph. Our method is able to: (1) precisely and dynamically localize the key parts in a
48 computational graph during fine-tuning on any tasks and (2) construct and optimize the circuit to
49 fit a new data distribution, thereby actively providing guidance for the fine-tuning process from an
50 interpretable perspective. Experiments confirm the validity of our theory and the effectiveness of our
51 method on various scenarios. We visualize and analyze a transparent, interpretable, and traceable
52 training process, provide new findings about the self-organization mechanism inside a model during
53 fine-tuning, and open new horizons for controllable and steerable training.

54 In summary, our contributions are mainly as follows:

- 55 • We redefine intrinsic dimensionality and extend the estimation of it from the parameter space
56 to the representation space, thereby introducing that concept into mechanistic interpretability
57 and making it possible to study the learning dynamics in a computational graph.
- 58 • We propose circuit-tuning, a general way of interpretable fine-tuning. Our method integrates
59 heuristic and dynamic circuit discovery into fine-tuning for the study of learning dynamics
60 for the first time, serving as a feasible way of interpretability-driven training.
- 61 • We confirm the effectiveness of our method through experiments and gain new insights into
62 the mechanisms behind learning through causal analysis and visualization, providing new
63 insights and improvements for the study of learning dynamics.

64 2 Related work

65 2.1 Intrinsic dimensionality

66 Intrinsic dimension was studied in early works to estimate the dimensionality of a dataset [15, 58,
67 6, 29]. Inspired by the redundancy in parameters, Li et al. [31] defined intrinsic dimensionality as
68 the minimum number of parameters needed to reach satisfactory solutions for an objective function.
69 Aghajanyan et al. [1] focused on fine-tuning and defined it as the minimal description length in terms
70 of a specific task, providing new insights into the generalization during pre-training. Prior works
71 directly looked at the weights in a model and measure the intrinsic dimensionality in the parameter
72 space, while the decomposition of a model and the following research on mechanistic interpretability
73 focused on the activations in the representation space. The divergence of research perspectives has
74 brought difficulties to the study of intrinsic dimensionality in the field of mechanistic interpretability,

75 2.2 Mechanistic interpretability for learning

76 Mechanistic interpretability aims to understand the computational mechanisms of a model [53].
77 Existing works can be divided into *static research* on the trained model and *dynamic research*
78 during training according to whether the model is in an inference state or a training state. In the
79 static research, circuit discovery is a widely used technique that aims to look for the subgraph in a
80 computational graph that acts as an algorithm implemented in a model for a certain behavior. Meng
81 et al. [35] proposed activation patching to identify activations relevant to the output, while Nanda [37]
82 proposed attribution patching to accelerate it. Wang et al. [62], Conmy et al. [9] focused on edges
83 and proposed path patching. Others optimized this technique from various aspects [56, 27, 34, 2, 32].
84 Despite their success, current studies are limited to discover the circuit of an existing model behavior
85 of a trained model in a static style, while our method is able to form a new circuit of a *non-existent*
86 capability through iterations in a dynamic mode during training.

87 Compared to static research, there is fewer research on learning dynamics. For phase transition,
88 Nanda et al. [38] delved into the study of grokking, while Olsson et al. [42] analyzed the emergency
89 of induction heads for in-context learning. For fine-tuning mechanisms, Prakash et al. [46] focused on
90 entity linking, Jain et al. [23] focused on compiled models and probabilistic context-free grammars,
91 Wang et al. [63] focused on math, while Lee et al. [28] focused on DPO for toxicity reduction.
92 Kotha et al. [26] explored the catastrophic forgetting. Wu et al. [64] proposed ReFT that learns an
93 intervention on the representations for efficient fine-tuning. Parallel to our work, Ren and Sutherland
94 [50] proposed a framework for learning dynamics by decomposing the change of a model’s prediction.
95 However, current studies are limited to a single scenario with various analytic methods, while there
96 lacks a unified method for the study of learning dynamics on common tasks, and no attempt was
97 made to actively introduce interpretability into fine-tuning for mechanistic study.

98 Unlike prior works, we successfully transfer the concept of intrinsic dimensionality from the parameter
99 space to the representation space, and dynamically integrate circuit discovery into fine-tuning as a
100 heuristic method for the study of learning dynamics for the first time, providing new insights for the
101 study of learning dynamics from the mechanistic view.

102 3 The theory of node-level intrinsic dimensionality

103 3.1 Background

104 As discussed in Section 2.1, intrinsic dimensionality implies the redundancy in parameters. Measuring
105 the exact intrinsic dimensionality is computational intractable. Therefore heuristic methods are
106 adopted to measure the upperbound of it. Considering a set of parameters $\theta^D = [\theta_0, \dots, \theta_m]$ and an
107 objective function $f(\cdot, \theta)$, Li et al. [31] leveraged a re-parameterization method to optimize only the
108 parameters θ^d in a subspace $\mathbb{R}^d (d < D)$ via a linear transformation with a projection matrix P :

$$\theta^D = \theta_0^D + P(\theta^d) \quad (1)$$

109 If a satisfactory solution is reached, then the dimensionality of that subspace is the intrinsic dimension-
110 ality. In practice, the heuristic method requires searching over various d , optimizing the parameters
111 θ^d and selecting the smallest d that could reach a satisfactory result.

112 3.2 Redefine intrinsic dimensionality

113 In this section, we view the model M as a computational graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where \mathcal{V} and \mathcal{E} represent
114 the nodes and edges in \mathcal{G} respectively. Each node is a vector $n = (h_1, \dots, h_N)^\top \in \mathbb{R}^N (1 \leq N \leq D)$
115 that could be the activation of a neuron, a group of neurons or a representation in a D -dimensional
116 representation space \mathbb{V}^D , based on granularities of interest. The edges describe the information
117 flows between the nodes, i.e., where the output of an upstream node goes and where the input of a
118 downstream node is from.

119 Note that it is possible to set a node to a full layer, or a latent which is a sparse feature in dictionary
120 learning, while here we require the size of a node to be no more than the dimensionality of \mathbb{V}^D to
121 cover most of the cases in practice and for convenience of discussion. Also, an edge is not necessary
122 to follow the real structure of a model. It could be a virtual connection between non-adjacent nodes.

123 Since the decomposition of a model in mechanistic interpretability often looks at activations or latents
124 [53], we look for the intrinsic dimensions indirectly via the activations in the representation space.
125 Inspired by prior works on circuit discovery, we believe that the learning process of a model is to
126 dynamically identify the redundancy in a computational graph and optimize the subgraph of a task.

127 **Definition 1: intrinsic dimensionality** *Given the computational graph \mathcal{G} of a model and a task T ,
128 the intrinsic dimensionality is the number of nodes in the minimal subgraph $\mathcal{C}^* = \{\mathcal{V}_T^*, \mathcal{E}_T^*\} \subset \mathcal{G}$ on
129 which the optimization of the objective function for task T could reach a predefined precision level.*

130 The newly defined intrinsic dimensionality, namely *the node-level dimensionality*, describes the
131 redundancy in parameters indirectly with the *"intrinsic nodes"* in the minimal subgraph. Under
132 this new definition, the parameters necessary for a task T refer to those where the intrinsic nodes
133 \mathcal{V}_T^* are derived from. For example, if a parameter matrix $W \in \mathbb{R}^{m \times n}$ maps an input $x \in \mathbb{R}^n$ to a
134 representation $H \in \mathbb{R}^m$ that is an intrinsic node, then W is what matters to task T . A theoretical
135 explanation of this is shown in Section 3.4 and detailed in Appendix E.

3.3 Estimating intrinsic dimensionality

Different from prior works that measure the intrinsic dimensionality in a subspace, we choose to perform the estimation in a subgraph. Similar to Li et al. [31], we also adopt a heuristic method to estimate the intrinsic dimensionality through training. The difference is that instead of optimizing via re-parameterization, we directly search for the intrinsic nodes in the computational graph through iterations. Inspired by the idea of causal intervention for the structural causal model [39], we adopt the concept of indirect effect [44]. Specifically, let's consider a node $n \in \mathcal{V}$. For the task T , a batch of inputs $\mathcal{X} = \{x_i\}_{i=1}^p \sim \mathcal{D}_T$ is fed into the model M . For each input x , if we replace $n(x)$ that is the activation of the node n with another value $n(x')$ from a corrupted input x' while keeping other activations unchanged, we can get another output $M(x')$. Given a metric \mathcal{L}_m for measuring the output, the contribution $c(n)$ of the node n in terms of the input x is its indirect effect on the output:

$$c(n) = \left| \mathcal{L}_m[M(x \mid \text{do}(n \leftarrow n(x')))] - \mathcal{L}_m[M(x)] \right| \quad (2)$$

The do-calculus notation [39] is used to express the intervention behavior in Equation (2). Given a lowerbound τ , If the contribution of n satisfies: $\mathbb{E}_{x_i \sim \mathcal{X}}[c_i(n)] > \tau$, then n is an intrinsic node. During optimization, we dynamically prune the graph to encourage exploration over the computational graph. When a satisfactory precision of optimization is reached, then the intrinsic dimensionality is the minimum number of intrinsic nodes $|\mathcal{V}_T^*|$ in the final subgraph for the task T .

3.4 Further discussions on the node-level intrinsic dimensionality

What does the indirect effect imply? As discussed in Section 3.3, the indirect effect is adopted to measure the importance of a node in a causal manner. Since intrinsic dimensionality originates from the parameter redundancy in a model [29, 1], while a node is defined over the activations, *how can we describe parameter redundancy through the lens of activations, and what exactly are the redundancies in a representation?* To gain a deeper understanding of it, we provide a theoretical analysis detailed in Appendix E. In short, we decompose Equation (2) into two parts by applying a first-order Taylor expansion to the right side of it at $n = n(x)$ for approximation:

$$(2) \approx \left| \mathcal{L}_m[M(x)] + [n(x') - n(x)]^\top \nabla_n \mathcal{L}_m[M(x)] - \mathcal{L}_m[M(x)] \right| = \left| [n(x) - n(x')]^\top \nabla_n \mathcal{L}_m[M(x)] \right| \quad (3)$$

In practice, the intervention on the node n is performed using the averaged result of all other nodes. Therefore $n(x) - n(x')$ can be viewed as a centered activation $n(x)^c$. Thus Equation (3) consists of two parts: (i) the first part $n(x)^c$ is an indicator for **representation redundancy**, which describes the redundancy of a node for representing an input; (ii) the second part $\nabla_n \mathcal{L}_m[M(x)]$ is an indicator for **forward redundancy**, which describes the redundancy of a node in the forward propagation. Thus the estimation of the intrinsic dimensionality can be converted to the process of redundancy detection in a graph. The definitions of the redundancies as well as the proof are detailed in Appendix E.

The minimum description length (MDL) in a computational graph. The MDL principle [51] encourages us to minimize the combined cost of encoding the model and the data given the model. While the parameters for the task T may not be decoupled from others and are perhaps distributed over the whole parameter space \mathbb{R}^D , prior works on intrinsic dimensionality used a reparameterization trick to separate the intrinsic dimensions $\theta^d \in \mathbb{R}^d$ out from the original parameter space. This kind of measurement aims to look for the upperbound of the MDL in terms of a task.

While we aim to (i) find out the minimal subgraph \mathcal{C}^* from a computational graph \mathcal{G} and (ii) interpret the learning process of a model based on the subgraph, there is a need to balance the minimum of the required parameters and the convenience for interpretable graph structure. Since we want to keep the original topological structure of the computational graph for further analyses, we directly delve into the original graph to inspect the redundancy, instead of adding extra structures like linear projections. Therefore, the minimal subgraph does not mean the minimum parameters in a strict sense.

However, the functionalities of the nodes in a computational graph are thought to be decoupled from each other in the mechanistic study. Braun et al. [3] argued that one of the properties of a good decomposition of a model is the minimality that the decomposition should use as few components as possible to represent a model behavior of interest, which also follows the MDL principle. Therefore, our definition holds for the MDL principle in the sense of the minimality in mechanistic interpretability. Different from prior works, we do not require a strict estimation of the intrinsic dimensionality, but trying to take into account both the minimality of the graph and the feasibility for interpretability.

4 Main method

4.1 Circuit-tuning

From nodes to edges Following the discussions in Section 3, we introduce circuit-tuning, an interpretable algorithm that allows transparent and precise fine-tuning. The only difference is that we shift our view from nodes to edges. The reason is that (i) we are interested in the interactions between nodes and (ii) the redundancy in nodes can be identified indirectly by checking the contributions of the edges connected to it. The shift from edges to nodes is actually an extension of the estimation method in Section 3.3, and the underlying idea is the same. Thus we can leverage the *path patching* technique in circuit discovery to identify the intrinsic dimensions. Discussions are detailed in Appendix D.

Given a model M for fine-tuning, the model is initialized into a computational graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, as defined in Section 3.2. A metric \mathcal{L}_m and a loss function \mathcal{L} are used for output quantification and parameter optimization respectively, as described in Section 3.3. Given a dataset $\mathcal{X} \sim \mathcal{D}_T$ for the fine-tuning task T , circuit-tuning alternately performs the following two procedures:

(1) Graph pruning: For a batch of data $\mathcal{X}_T \in \mathcal{X}$, we compute the edge contribution for each edge $e : n_1 \rightarrow n_2 \in \mathcal{E}$ according to Equation (4), where n_1 and n_2 are the upstream and downstream nodes respectively. Then all the edges are sorted in descending order based on their contributions, and the edges with top N contributions are selected. Then the graph \mathcal{G} is pruned into a circuit $\mathcal{C} = \{\mathcal{V}_T, \mathcal{E}_T\}$ with only the selected edges \mathcal{E}_T together with the nodes \mathcal{V}_T at both ends of each edge inside.

(2) Circuit fine-tuning: For a batch of data $\mathcal{X}_T \in \mathcal{X}$, all the parameters outside \mathcal{C} are frozen, and only the parameters corresponding to the nodes \mathcal{V}_T (see Section 3.2) are updated through a forward pass and a backward pass as usual. After K steps of optimization, all the frozen parameters are freed and the graph \mathcal{G} is reset to its original state.

In practice, it is efficient to accelerate graph pruning with only one forward and backward pass using edge attribution patching [56]. The implementation details are shown in Algorithm 1 and Section 5.

Algorithm 1 Circuit-tuning

Input: dataset \mathcal{X} , model M , loss function \mathcal{L} , the metric for the quantification of the output \mathcal{L}_m , the number of edges to save N , the optimization steps after circuit discovery K .
Initialize graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ from M , circuit $\mathcal{C} = \mathcal{G}$, and the iteration step $i = 0$.
for mini-batch $\mathcal{X}_T = \{x_1, x_2, \dots, x_t\}$ in \mathcal{X} **do**
 Run a forward and backward pass on \mathcal{X}_T
 if $i \bmod K == 0$ **then**
 Reset: $\mathcal{C} \leftarrow \mathcal{G}$
 for edge $e : n_1 \rightarrow n_2 \in \mathcal{E}$ **do**
 Intervene in edge e and calculate the contribution of it according to Equation (7) and (4)
 end for
 $\mathcal{E}_T \leftarrow \{e \mid e \in \mathcal{E} \text{ with top-}N \text{ edge contributions}\}$
 $\mathcal{C} \leftarrow \{\mathcal{V}_T, \mathcal{E}_T\}$, where $\mathcal{V}_T = \{n \mid n \in \mathcal{V} \wedge n \text{ is incident to an edge } e \in \mathcal{E}_T\}$
 end if
 Update the parameters corresponding to the nodes in \mathcal{C}
 $i = i + 1$
end for

4.2 Characteristics of Our Method

Heuristic and dynamic circuit discovery Unlike prior works that focus on the circuits in a trained model, circuit-tuning acts as a heuristic approach for a model to develop an unseen ability in a dynamic style. The validity of this idea is built on the finding in [1] that pre-training has learned enough knowledge for downstream tasks, allowing fine-tuning with minor modifications.

Interpretable model diffing Instead of being a blackbox, circuit-tuning performs fine-tuning on a visible circuit, making the full process open to researchers. The transparency of our method provides the convenience to track the learning dynamics and the availability to intervene in the training process. Thus we are able to compare among different training stages and gain a deeper understanding and better control of it. Visualization of the training process is detailed in Section 5 and Appendix F.4.

Another feature of our method is that it is able to precisely locate the key components and update only a few parameters as discussed in Section 3.4, instead of modifying all of the parameters as LoRA-based methods do. Besides, while ReFT [64] a similar method also leverages representations for training, it is aimed at parameter-efficient fine-tuning and learns a transformation as an intervention on a hidden state, while our method is more focused on the mechanisms of learning during training.

5 Experiments

In this section, we test our theory and method across a variety of models and tasks. To comprehensively investigate the learning dynamics during fine-tuning as well as the practicality of our method, we arrange the experiments into two parts. The first part serves as the verification and analysis of our method, while the second part aims to show that our method is applicable to various scenarios.

5.1 The subject-verb disagreement Task

5.1.1 Task description and data preparation

The goal of the subject-verb disagreement task is to match a verb with a subject in an abnormal way. For example, “I is”, “he are” and “the cows eats” are all expected results for this task. In each sentence, the token before the verb is called the END token. The automatic evaluation metric for this task is the logit difference between the flipped verb and the original verb at the END token. This is because the logit at the END token is directly used for predicting the verb.

The reason why we create this task as the start of our experiments is that we expect the model to acquire a new capability from scratch, and thus to study how the circuit evolves during fine-tuning. The task is simple and interesting for analyzing the learning dynamics with clear visualization.

Different from previous works [14, 34] that use template-based datasets, we collect real-world data from the Pile corpus [17] to ensure diversity and authenticity. We extract 60k sentences in the present tense in English and flip the forms of the verbs. We ensure the high quality of our dataset and believe it is meaningful for further research. For the details of our dataset, please refer to Appendix F.1.

5.1.2 Implementation

We use GPT2-small [47] in this task. We set the output of the attention and the MLP in each layer as upstream nodes, and the input of the query, key, value and the MLP in each layer as downstream nodes. During training, we use the logit difference discussed before as the metric \mathcal{L}_m . We follow Syed et al. [56] for path patching with mean ablation, with an *improvement* detailed in Appendix F.4.5. We sweep over a range of N which is the number of edges to be saved during circuit discovery. Mini-batch SGD is used for optimization. We train the model under each setting for 3 epochs where the performance almost converges. For implementation details, please refer to Appendix F.2.

5.1.3 Main results

From Figure 1(a), we observe a flip in logit difference from negative to positive, which means the model adjusts its grammar to fit the data distribution of subject-verb disagreement. One noteworthy finding is that the model can generate abnormal texts in the past tense, such as “to be or not to be, that were a ...”, which implies that the model really grasps the new grammar and applies it smartly, since there is no sentence in the present tense in our training data at all.

With the increasing number of top N edges, the logit difference increases until N approaches 1000. The number of tunable parameters is also saturated at this point. We believe that when $N = 1000$, almost all the necessary parameters for this task are included, and the performance cannot be further improved, suggesting the *minimality* of the circuit and the existence of intrinsic dimensionality.

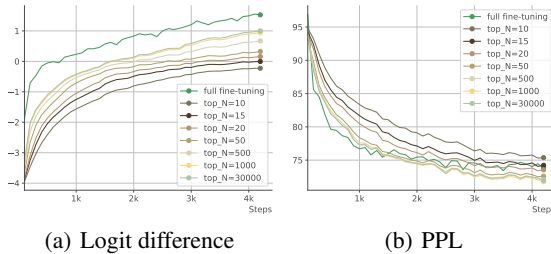


Figure 1: Results in the subject-verb disagreement task.

We also track the perplexity (PPL) that is calculated over the test set with corpora in various scenarios. In Figure 1(b), we find that the PPL of full fine-tuning is high and fluctuates wildly during training, though the logit difference of it is higher. This implies that circuit-tuning provides better training stability as well as better preservation of the basic language modeling ability over full fine-tuning.

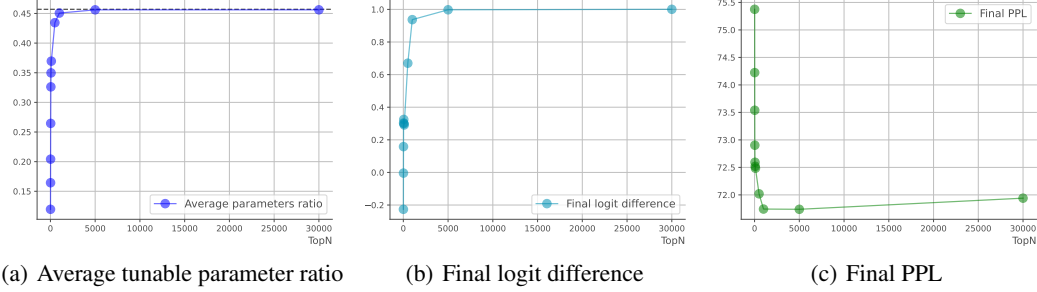


Figure 2: The changes in indexes with various choices of top-N imply the existence of the node-level intrinsic dimensionality at $N = 1000$ in a computational graph. Note that the parameter ratio varies and depends on the granularity of nodes. Theoretically more granular nodes result in lower parameter ratio because the location of the intrinsic nodes would be more precise.

5.1.4 Ablation study for the quality of the discovered circuit

To confirm the effectiveness of our method to find the required parameters, we (i) calculate the faithfulness and completeness [62] of the circuits and (ii) provide an ablation study to randomly unfreeze the nodes outside the circuit during training and check if the performance is influenced. Results of the above experiments in Appendix F.3 ensure that the circuits are discovered accurately. Besides, we observe an obvious turning point of the faithfulness and completeness in Figure 8 as N increases. Also, we notice a similar and common turning point of logit difference, PPL and number of tunable parameters in Figure 2. All of the above observations show the shadow of *intrinsic nodes*.

5.1.5 Analyses on learning dynamics

Finding 1: Interpretation of the circuits Inspired by Wang et al. [62], we decompose the circuit into attention heads of different functions. Firstly we find out the heads that directly affect the output, namely the **Subject Attribute Heads**. Then we find out the heads responsible for the localization of the subject, namely the **Subject Identification Heads**. Finally, we find out heads that could affect the behaviors of the Subject Attribute Heads, namely the **Collaborative Heads**. For technique details and visualization of the heads, please refer to Appendix F.4.1.

Finding 2: The flip of the Subject Attribute Heads The Subject Attribute Heads are responsible for matching the subject and the verb. We compute the dot product between the output x of each attention head at the END token and $W_U(v_{flip}) - W_U(v)$, the difference between the unembedding projections. Since the latter is fixed, we expect the projection of x on the direction of it to be large, thus encouraging the probability difference between the two opposite verb forms (love v.s. loves, etc.). We visualize the heads before and after fine-tuning in Figure 3. Through comparison, we observe an obvious flip at head.10.9, implying the reversal in its function from subject-verb agreement to disagreement. Other heads (3.0, 6.0, 4.4, 11.8, etc) also see flips with varying degrees, serving as evidence for the *self-adjustment* of the functions inside the nodes. See Appendix F.4.1 for details.

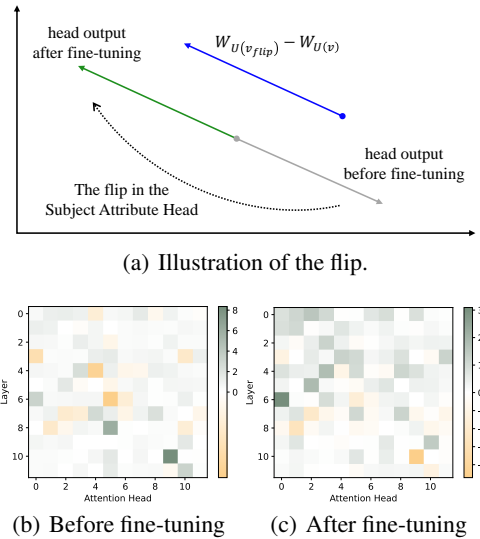


Figure 3: The flip observed in Finding 2.

Table 1: Some of the strengthened and weakened edges during training. The dynamic change shows the change in the logarithm of the edge contribution $\log[1 + c(e)]$ (the start of training \rightarrow 2000 steps \rightarrow 3000 steps \rightarrow 4000 steps). The dynamic process is visualized in Figure 4.

| Strengthened Edges | | Weakened Edges | |
|---------------------------------|--|----------------------------|--|
| Edge | Dynamic change | Edge | Dynamic change |
| mlp.2 \rightarrow head.11.8.v | 0 \rightarrow 0.2744 \rightarrow 0.5091 \rightarrow 0.9138 | mlp.2 \rightarrow mlp.8 | 0.1122 \rightarrow 0.0869 \rightarrow 0.0619 \rightarrow 0.058 |
| mlp.1 \rightarrow head.11.8.v | 0 \rightarrow 0.2227 \rightarrow 0.3978 \rightarrow 0.7231 | mlp.1 \rightarrow mlp.5 | 0.1043 \rightarrow 0.0859 \rightarrow 0.0736 \rightarrow 0 |
| mlp.2 \rightarrow mlp.3 | 0 \rightarrow 0.0293 \rightarrow 0.0993 \rightarrow 0.2282 | mlp.0 \rightarrow mlp.10 | 0.2712 \rightarrow 0.0580 \rightarrow 0 \rightarrow 0 |
| mlp.1 \rightarrow mlp.4 | 0 \rightarrow 0.0396 \rightarrow 0.0652 \rightarrow 0.1748 | mlp.4 \rightarrow mlp.11 | 0.1791 \rightarrow 0.0454 \rightarrow 0.0428 \rightarrow 0 |
| mlp.2 \rightarrow mlp.5 | 0 \rightarrow 0 \rightarrow 0.1246 \rightarrow 0.1734 | mlp.4 \rightarrow mlp.11 | 0.1885 \rightarrow 0.0343 \rightarrow 0.0259 \rightarrow 0 |
| ... | ... | ... | ... |

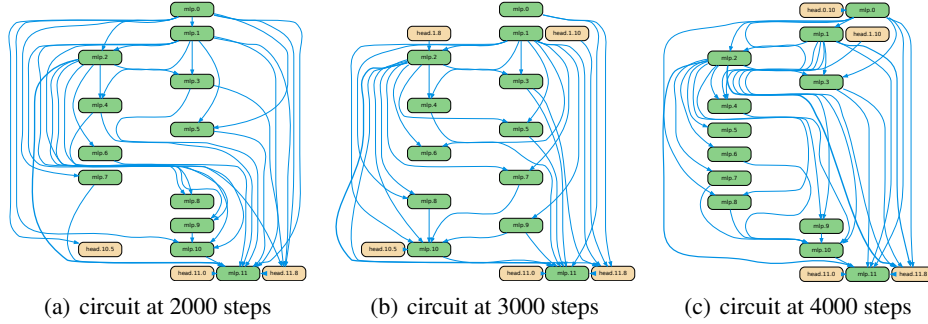


Figure 4: Evidence of Hebbian learning in the subject-verb disagreement task ($N = 50$ edges shown).

Finding 3: The sharing of the Subject Identification Heads The Subject Identification Heads attend heavily to the subject. One type of these heads attends to the END token (0.1, 0.3, etc), which is helpful to the cases like “they are”; the other type of heads (8.5, 10.5, etc) attends to the subject several tokens before, which is helpful to the cases like “the girl wearing a dress is”. Both types of heads perform the same before and after fine-tuning. This implies that their functions are preserved and shared all the way down. Our finding is consistent with the study of intrinsic dimensionality in Aghajanyan et al. [1] that pre-training optimizes the description length without having direct access to the same tasks. The kind of consistency further confirms the feasibility of our theory in Section 3.

Finding 4: The interaction and collaboration inside the model Interestingly, we observe that over the training process the nodes (attention heads) inside a circuit complete a task through a cooperative division of labor. Each head has its division of labor as discussed before. Besides, some heads directly affect the output, while others cooperate with them to affect the output indirectly. Several heads achieve a common goal through cooperation: they could affect a head only through combined effect, and adjust themselves during training. For details, please refer to Appendix F.4.1.

Finding 5: The evidence of Hebbian learning We visualize the circuits during the fine-tuning process in Figure 4, and observe a phenomenon akin to Hebbian learning [10] that several edges are strengthened during training, just as the synapses between neurons can be strengthened after continuous stimulation. Also, several edges are weakened at the same time, as shown in Table 1. It is similar to the lateral inhibition among neurons [22] that an activated neuron can reduce the activity of its neighbors. We present the dynamics of the *competitive learning and self-organization* [25] inside a model, providing new discoveries to the generalization process on a new data distribution.

Due to limited space, all the above together with other interesting findings are detailed in F.4.

5.2 Application to Complex Tasks

Task descriptions We prepare two types of tasks based on whether reasoning is involved. For reasoning-based tasks, we focus on mathematics and logical reasoning with chain-of-thought (CoT) included. For reasoning-free tasks, we prepare a gender de-biasing task that requires the model to develop an unbiased perspective on genders, and a reading comprehension task that requires only the keywords as the answer. Task settings and evaluation metrics are detailed in Appendix G.3.

Table 2: Experiment results of circuit-tuning on complex tasks.

| Methods & Tasks | Reasoning-based | | Reasoning-free | | Computation |
|--------------------------------|-----------------|-------------------|-------------------|-----------------------|-------------|
| | Mathematics | Logical Reasoning | Gender De-biasing | Reading Comprehension | |
| | Acc@1 (%) | F1 (%) | Prejudice Risk ↓ | Exact Match(%) | |
| Llama-3.2-1B-it | 40.71 | 20.35 | 0.555 | 39.68 | / |
| Llama-3.2-1B-it-full-tuning | 46.47 | 26.89 | 0.533 | 36.73 | 1.00 |
| Llama-3.2-1B-it-lora | 44.58 | 22.51 | 0.530 | 34.30 | 1.79e-2 |
| Llama-3.2-1B-it-circuit-tuning | 45.56 | 27.06 | 0.312 | 41.78 | 7.65e-2 |
| Llama-3.2-3B-it | 70.36 | 42.71 | 0.641 | 54.12 | / |
| Llama-3.2-3B-it-full-tuning | 75.44 | 47.32 | 0.632 | 55.63 | 1.00 |
| Llama-3.2-3B-it-lora | 73.54 | 46.27 | 0.638 | 54.73 | 1.49e-2 |
| Llama-3.2-3B-it-circuit-tuning | 74.35 | 47.59 | 0.417 | 56.58 | 8.57e-2 |
| Llama-3.1-8B-it | 76.19 | 46.41 | 0.651 | 58.92 | / |
| Llama-3.1-8B-it-full-tuning | 83.76 | 49.53 | 0.640 | 59.60 | 1.00 |
| Llama-3.1-8B-it-lora | 80.21 | 47.64 | 0.643 | 58.97 | 1.03e-2 |
| Llama-3.1-8B-it-circuit-tuning | 82.97 | 50.54 | 0.420 | 60.04 | 9.37e-2 |

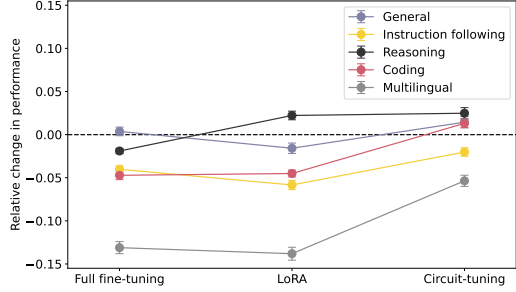
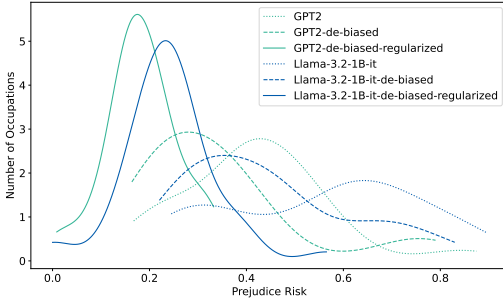


Figure 5: Comparison in gender debiasing results. Figure 6: Evaluation on general/other capabilities.

Implementation We apply our method to llama-3.2-1B/3B and llama-3.1-8B [11]. In addition, we compare it with full fine-tuning and LoRA. We set them as baselines for simple comparisons only to verify that our method is also able to locate the key parameters precisely, as discussed in Section 4.2.

For circuit-tuning, the graph settings of our method are the same as before in Section 5.1.2, except that each MLP layer is split into 64-dimensional heads. For gender de-biasing, We selectively add $\beta \cdot |\mathcal{L}_m|$ that acts as a regularization term in the loss \mathcal{L} for explicit guidance and intervention during training (Appendix G.1.2). All other details like the choice of the patching metric \mathcal{L}_m , the number of edges N , and the hyper-parameters for circuit-tuning and baselines are detailed in Appendix G.2.

Main results Table 2 shows the results of each method, demonstrating that our method is capable of various scenarios. Hence it is possible to study the learning dynamics as in Section 5.1 on various scenarios. Amazingly, the task performance brought by circuit-tuning is better than LoRA, and even surpasses full fine-tuning in many tasks. Note that the computation of circuit-tuning depends on node granularity and is for reference only since our method is not dedicated to efficient fine-tuning.

Besides, with an interpretable intervention, the de-biasing result is clearly better than others, which is shown in Figure 5 and discussed in Appendix G.4. This suggests that circuit-tuning can be applied flexibly with task-specific conditions to achieve controllable and steerable training.

To check if other capabilities are affected during training, we evaluate on benchmarks involving general capabilities as well as reasoning, coding, and multilingual abilities. Results are shown in Figure 6 and detailed in Appendix G.3, indicating that circuit-tuning is more targeted at the key components in a model, ensuring its *reliability* and *simplicity* in interpretability research.

6 Conclusion

We extend the concept of intrinsic dimensionality to the node level, and describe the learning process inside a model from the mechanistic view. Further, we propose circuit-tuning as a promising tool for interpretable fine-tuning. We analyze the mechanism behind learning and provide new insights for the learning process. Limitations and future discussions of our work are shown in Appendix A.

References

- [1] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.
- [2] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermy, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.
- [3] Dan Braun, Lucius Bushnaq, Stefan Heimersheim, Jake Mendel, and Lee Sharkey. 2025. Interpretability in parameter space: Minimizing mechanistic description length with attribution-based parameter decomposition. *arXiv preprint arXiv:2501.14926*.
- [4] Trenton Bricken, Siddharth Mishra-Sharma, Jonathan Marcus, Adam Jermy, Christopher Olah, Kelley Rivoire, and Thomas Henighan. 2024. Stage-wise model diffing.
- [5] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.
- [6] Francesco Camastra and Alessandro Vinciarelli. 2002. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on pattern analysis and machine intelligence*, 24(10):1404–1407.
- [7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- [8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- [9] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.
- [10] Hebb Do. 1949. The organization of behavior. *New York*.
- [11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- [12] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.
- [13] 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, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.
- [14] 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.
- [15] Keinosuke Fukunaga and David R Olsen. 1971. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on computers*, 100(2):176–183.
- [16] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

- [17] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- [18] Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060.
- [19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- [20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- [21] Wenyue Hua, Kaijie Zhu, Lingyao Li, Lizhou Fan, Shuhang Lin, Mingyu Jin, Haochen Xue, Zelong Li, JinDong Wang, and Yongfeng Zhang. 2024. Disentangling logic: The role of context in large language model reasoning capabilities. *arXiv preprint arXiv:2406.02787*.
- [22] Marcus Jacobson. 1993. *Foundations of neuroscience*. Springer Science & Business Media.
- [23] Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*.
- [24] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.
- [25] Teuvo Kohonen. 2012. *Self-organization and associative memory*, volume 8. Springer Science & Business Media.
- [26] Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2023. Understanding catastrophic forgetting in language models via implicit inference. *arXiv preprint arXiv:2309.10105*.
- [27] János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. Atp*: An efficient and scalable method for localizing llm behaviour to components. *arXiv preprint arXiv:2403.00745*.
- [28] Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*.
- [29] Elizaveta Levina and Peter Bickel. 2004. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17.
- [30] Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. *arXiv preprint arXiv:2109.03858*.
- [31] Chunyuan Li, Heerad Farkhor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*.
- [32] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the biology of a large language model. *Transformer Circuits Thread*.
- [33] Y Liu, K Yang, Z Qi, X Liu, Y Yu, and C Zhai. 2024. Prejudice and volatility: A statistical framework for measuring social discrimination in large language models.
- [34] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.
- [35] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt.
- [36] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

463 [37] Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale.

464 [38] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures
465 for grokking via mechanistic interpretability.

466 [39] Leland Gerson Neuberg. 2003. Causality: models, reasoning, and inference, by judea pearl, cambridge
467 university press, 2000. *Econometric Theory*, 19(4):675–685.

468 [40] nostalgebraist. 2020. Interpreting gpt: The logit lens.

469 [41] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020.
470 Zoom in: An introduction to circuits. *Distill*.

471 [42] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben
472 Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds,
473 Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei,
474 Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and
475 induction heads. *Transformer Circuits Thread*.

476 [43] Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry
477 of large language models.

478 [44] Judea Pearl. 2001. Direct and indirect effects. *Probabilistic and Causal Inference*.

479 [45] Alexandre Pouget, Peter Dayan, and Richard Zemel. 2000. Information processing with population codes.
480 *Nature Reviews Neuroscience*, 1(2):125–132.

481 [46] Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning
482 enhances existing mechanisms: A case study on entity tracking. *arXiv preprint arXiv:2402.14811*.

483 [47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language
484 models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

485 [48] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions
486 for squad. *arXiv preprint arXiv:1806.03822*.

487 [49] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani,
488 Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *ArXiv*,
489 abs/2311.12022.

490 [50] Yi Ren and Danica J Sutherland. 2024. Learning dynamics of llm finetuning. *arXiv preprint*
491 *arXiv:2407.10490*.

492 [51] Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

493 [52] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An
494 adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

495 [53] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-
496 Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. 2025. Open problems in mechanistic
497 interpretability. *arXiv preprint arXiv:2501.16496*.

498 [54] Carla J Shatz. 1992. The developing brain. *Scientific American*, 267(3):60–67.

499 [55] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won
500 Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought
501 reasoners. *arXiv preprint arXiv:2210.03057*.

502 [56] Aaqub Syed, Can Rager, and Arthur Conmy. 2023. Attribution patching outperforms automated circuit
503 discovery. *arXiv preprint arXiv:2310.10348*.

504 [57] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce,
505 Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall,
506 Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn,
507 Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features
508 from claude 3 sonnet. *Transformer Circuits Thread*.

509 [58] Joshua B Tenenbaum, Vin de Silva, and John C Langford. 2000. A global geometric framework for
510 nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[60] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv e-prints*, pages arXiv–2308.

[61] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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

[63] Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. 2025. Towards understanding fine-tuning mechanisms of llms via circuit analysis. *arXiv preprint arXiv:2502.11812*.

[64] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962.

[65] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

[66] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

529 A Limitations and future discussions

530 In this section, we report some limitations about our work, and discuss some potential research
531 directions based on our method.

- 532 1. We do not test our method with different granularities. The node in a computational graph
533 could be a neuron, a group of neurons, the activation of an attention head or even a layer.
534 Also, it could be a latent in the representation of a sparse autoencoder. While in our
535 experiment, we treat the activation of each attention / MLP head as a node for convenience.
536 We believe it is possible to try other granularities.
- 537 2. We only focus on attention heads as most people do. In fact, we believe that a lot of treasures
538 are hidden inside the MLP layers. Recently Ameisen et al. [2], Lindsey et al. [32] applied
539 transcoders to MLP layers and discovered the circuits inside a model. Thus we look forward
540 to seeing circuit-tuning combined with other model components besides attention.
- 541 3. We believe that the experiment on complex tasks could be further studied. For example, we
542 could analyze the learning dynamics during the fine-tuning process for COT-based tasks, or
543 COT-free tasks like the gender de-biasing task. However, due to the complexity of these
544 tasks and limited space, we think it is a better idea to study them in a separate work and
545 present them in a single paper. Thus we leave this part for further research.
- 546 4. The infra of our method could be improved. We believe it is possible to apply our method to
547 larger models, while it requires stronger frameworks. We leave it for future development.

548 B Social impact

549 This paper presents work whose goal is to advance the field of Machine Learning. Since our work
550 concerns AI alignment and involves modifying a model’s gender stereotype in Section 5.2, we expect
551 our work to be used under the guidance of human values.

552 C Transformer Architecture

553 The models we use in our experiments are all decoder-only Transformers. We briefly introduce the
554 Transformer architecture from the mechanistic view, together with its implementation.

A single input of Transformer is $x_0 \in \mathbb{R}^T$, where T is the length of the sequence. The input is firstly embedded into a vector $x \in \mathbb{R}^{D \times T}$ via the embedding matrix $W_E \in \mathbb{R}^{D \times V}$, where D is the model dimension. Then x will go through l layers of Transformer blocks for various processings. From the view of [13], we can think of the residual stream as a communication channel that simply receives the output of the self-attention and MLP operations. Each operation reads information from the residual stream and writes the processed information into it. Thus the residual stream is actually the linear sum of various transformations of x together with the original input x .

In each Transformer layer $i (i \in [0, L])$, the two important operations are self-attention and MLP. In self-attention, we consider the implementation of multi-head attention. The model dimension is split into H parts, and the attention operation is performed with H attention heads in parallel. Each head is thought to be responsible for a specific function. Consider head.i.j ($j \in [0, H)$), the input x is firstly projected into query, key and value via $W_Q^{i,j}$, $W_K^{i,j}$ and $W_V^{i,j}$. The projection matrices are all in shape $\mathbb{R}^{\frac{D}{H} \times D}$, thus x is projected into $x_{Q/K/V}^j \in \mathbb{R}^{\frac{D}{H} \times T}$. Then attention pattern $A_{i,j} \in \mathbb{R}^{T \times T}$ is computed via $(W_Q^{i,j} x_Q^j)(W_K^{i,j} x_K^j)^\top$ and some scaling and Softmax operations. After that, the weighted output $z \in \mathbb{R}^{\frac{D}{H} \times T}$ is computed via $(W_V^{i,j} x_V^j) A_{i,j}$. Finally, the output of head.i.j $Attn_i^j(x) \in \mathbb{R}^{D \times T}$ is calculated via $W_O^{i,j} \cdot z$, where $W_O^{i,j} \in \mathbb{R}^{D \times \frac{D}{H}}$. Thus, final output of the self-attention in layer i is $Attn_i(x) = \sum_{j=1}^H Attn_i^j(x)$.

For the MLP operation in each layer, the input x is projected into $x_{in}^i \in \mathbb{R}^{D_{mlp} \times T}$ via $W_{in} \in \mathbb{R}^{D_{mlp} \times D}$, and projected back to $MLP_i(x) \in \mathbb{R}^{D \times T}$ via $W_{out} \in \mathbb{R}^{D \times D_{mlp}}$. In the Llama architecture [59], the input x is firstly projected into $x_{pre}^i \in \mathbb{R}^{D_{mlp} \times T}$ via $W_{gate}^i \in \mathbb{R}^{D_{mlp} \times D}$ and is applied with an activation layer, then a dot product is performed between the activations and $x_{in}^i \in \mathbb{R}^{D_{mlp} \times T}$ which is the input projected by $W_{in} \in \mathbb{R}^{D_{mlp} \times D}$. Note that we can also split the MLP into MLP heads, which is done on Llama series models in the complex tasks in our experiments. For details, please refer to Appendix G.2.

The output of all the L layers are projected into $x \in \mathbb{R}^{V \times T}$ by the unembedding matrix $W_U \in \mathbb{R}^{V \times D}$, which is called the logits. The logit at the end of the sequence is further mapped into a probability distribution with Softmax over the vocabulary for predicting the next token.

D The Derivation of Edge Contribution

Similar to the definition of attribution score in [56], we can define the contribution of an edge $e : n_1 \rightarrow n_2$ with n_1 as the upstream node and n_2 as the downstream node. We use the indirect effect IE to measure the change in the output caused by the patching of the edge. Thus given a dataset \mathcal{X} for patching, the contribution of edge e can be expressed as follows:

$$\begin{aligned} c(e) &= \mathbb{E}_{x_i \sim \mathcal{X}} [|c_i(e)|] \\ &= \mathbb{E}_{x_i \sim \mathcal{X}} [|IE(e; x_i)|] \end{aligned} \quad (4)$$

Note that the contribution of edge e is directly reflected in the change of the final output (the logit of the language model, etc), which is the indirect effect caused by the change of the value in the downstream node n_2 , while the change of the node n_2 is directly caused by the change of the upstream node n_1 . The difference between the direct effect and the indirect effect is that the former keeps all the other nodes that could influence n_2 unchanged and only studies the influence from n_1 to n_2 , while the latter allows all the changes in nodes between n_2 and the logit. For more refined definitions for these two concepts, please refer to [44].

Therefore, to measure the direct effect from n_1 to n_2 , we set the value of n_1 to another value $n_1(x')$ while keeping the all other nodes between n_1 and n_2 unchanged. The indirect effect of e to the final output is

$$\begin{aligned} IE(e; x) &= IE(n_1 \rightarrow n_2; x) \\ &= \mathcal{L}_m[M(x | do(n_2 \leftarrow n_2(x')))] - \mathcal{L}_m[M(x)] \end{aligned} \quad (5)$$

in which the corrupted value $n_2(x')$ of the downstream node is

$$n_2(x') = n_2(x) - n_2^{n_1}(x) + n_2^{n_1}(x') \quad (6)$$

Equation 6 shows the direct effect $n_2^{n_1}(x') - n_2^{n_1}(x)$ from n_1 to n_2 , where

$$n_2^{n_1}(x') = n_2^{n_1}(x | do(n_1 \leftarrow n_1(x'))) \quad (7)$$

To simplify the equation, we apply a first-order Taylor expansion to IE at $n_2 = n_2(x)$, then

$$IE(e; x) \approx \mathcal{L}_m[M(x)] + [n_2(x') - n_2(x)]^\top \nabla_{n_2} \mathcal{L}_m[M(x)]|_{n_2(x)} - \mathcal{L}_m[M(x)] \quad (8)$$

$$= [n_2(x') - n_2(x)]^\top \nabla_{n_2} \mathcal{L}_m[M(x)]|_{n_2(x)} \quad (9)$$

Thus we have

$$\begin{aligned} IE(e; x) &= [n_2(x) - n_2^{n_1}(x) + n_2^{n_1}(x') - n_2(x)]^\top \nabla_{n_2} \mathcal{L}_m[M(x)]|_{n_2(x)} \\ &= [n_2^{n_1}(x') - n_2^{n_1}(x)]^\top \nabla_{n_2} \mathcal{L}_m[M(x)]|_{n_2(x)} \end{aligned} \quad (10)$$

To further simplify Equation 10, we apply another Taylor expansion at $n_1 = n_1(x)$ to $n_2^{n_1}$. Then we have

$$\begin{aligned} IE(e; x) &\approx \left\{ n_2^{n_1}(x) + [n_1(x') - n_1(x)]^\top \nabla_{n_1} n_2^{n_1}|_{n_1(x)} - n_2^{n_1}(x) \right\} \cdot \nabla_{n_2} \mathcal{L}_m[M(x)]|_{n_2(x)} \\ &= [n_1(x') - n_1(x)]^\top \nabla_{n_1} n_2^{n_1}|_{n_1(x)} \nabla_{n_2} \mathcal{L}_m[M(x)]|_{n_2(x)} \end{aligned} \quad (11)$$

Thus we come to the final form of edge contribution in Equation 11. Our derivation takes the simplest situation into consideration, while it works well in practice. For more discussions on the relevant topic, please refer to [27]. For implementation details, please refer to our code.

E The proof for the claim in Section 3.4

In Section 3.4, we state that the estimation of the intrinsic dimensionality can be converted to the process of detecting the parameter redundancy in a computational graph. Specifically, we believe that there are mainly two types of redundancies in a model: representation redundancy and forward redundancy. Representation redundancy describes the redundancy of node in representing the features of an input, while forward redundancy describes the redundancy of node in the forward propagation. Thus the redundancy of a node is the combination of the above two types of redundancies, which finally points to the redundancy in *parameters*.

To prove this, we first propose the definition of representation redundancy and forward redundancy in Appendix E.2 and extend them to the node level in Appendix E.3 and E.4, respectively. Then we prove that the approximation of the indirect effect in Equation (3) consists of two parts that correspond to the representation redundancy and forward redundancy respectively in Appendix E.5 and E.6.

E.1 Preliminaries

Features Given a representation space $\mathbb{V}^D (D = m)$ with a set of bases $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_D\}$ where each basis \mathbf{e}_i corresponds to a neuron, the feature vector \mathbf{v}_f of feature f is a direction in \mathbb{V}^D :

$$\mathbf{v}_f = \sum_{i=1}^m c_i \mathbf{e}_i = c_1 \mathbf{e}_1 + c_2 \mathbf{e}_2 + \dots + c_m \mathbf{e}_m \quad (12)$$

where the coefficient c_i satisfies $|c_i| \leq 1$ and $\sum_{i=1}^m c_i^2 = 1$.

The idea of feature vector follows the linear representation hypothesis that features are represented as vectors in a linear space, the evidence of which has been widely discussed [36, 5, 43].

Superposition Given a layer in a neural network with its representation space \mathbb{V}^D , if all features represented in that layer is \mathcal{F} , then we have $D < |\mathcal{F}|$, where $|\mathcal{F}|$ is the number of elements in \mathcal{F} .

The superposition hypothesis discussed in [41, 12] aims to explain the phenomenon of polysemanticity that one neuron in a neural network often responds to multiple unrelated inputs. Since there are more features than the dimensions, the model learns to compress more features than it has dimensions.

630 **Representation** Considering the settings of feature vector and superposition, the representation H
 631 of an input x in $\mathbb{V}^D (D = m)$ is the linear combination of its feature vectors:

$$H = \sum_{i=1}^t \mathbf{v}_{\mathbf{f}_i} = (h_1, \dots, h_m)^\top, h_i = \left(\sum_{i=1}^t a_{f_i} \mathbf{v}_{\mathbf{f}_i} \right) \cdot \mathbf{e}_i \quad (13)$$

632 where the features of x are $\mathcal{F}_x = \{f_1, f_2, \dots, f_t\}$ with their magnitudes $A_x = \{a_{f_1}, a_{f_2}, \dots, a_{f_t}\}$.
 633 Thus the activation h_i of dimension i in \mathbb{V}^D often originates from more than one features.

634 E.2 Two types of redundancy

635 Intrinsic dimensionality describes the redundancy in parameters in terms of a specific task, and our
 636 goal is to specify it. In this section, we define two types of redundancy from the mechanistic view.

637 When inspecting a layer with its weight matrix $W^{pre} \in \mathbb{R}^{m \times n}$ that maps an input from \mathbb{R}^n to \mathbb{R}^m ,
 638 we focus on a representation $H = (h_1, h_2, \dots, h_m)^\top \in \mathbb{V}^D$ in the representation space $\mathbb{V}^D (D = m)$
 639 of that layer.

640 **Assumption 1: Feature sparsity** Suppose all features represented in a layer is \mathcal{F} . Given a task T
 641 with $\mathcal{X} = \{x_1, x_2, \dots, x_t\} \sim \mathcal{D}_T$ which consists of t samples that follows a specific data distribution
 642 \mathcal{D}_T , the features for representing \mathcal{X} is $\mathcal{F}_T \subsetneq \mathcal{F}$.

643 This means that when conditioned on a task T , the number of features needed is smaller than that of
 644 the elements in the universal set \mathcal{F} . The assumption follows the concept of feature sparsity [12] that
 645 many features do not frequently appear.

646 **Definition 1: Representation redundancy** Given a task T with features \mathcal{F}_T , we focus on a
 647 representation $H = (h_1, \dots, h_m)^\top$ and the i -th dimension in $\mathbb{V}^D (D = m)$. We set the value of
 648 the activation h_i in that dimension to zero. For each feature $f \in \mathcal{F}_T$, we can get the deformed
 649 feature vector $\tilde{\mathbf{v}}_f$. Given a metric Φ for measuring the similarity between two vectors and a threshold
 650 $\tau_r \in (0, 1)$ as the tolerance of deformation, if the intervention on the i -th dimension satisfies:

$$\min_{1 \leq i \leq |\mathcal{F}_T|} \left[(\Phi(\mathbf{v}_{\mathbf{f}_1}, \tilde{\mathbf{v}}_{\mathbf{f}_1})), \dots, \Phi(\mathbf{v}_{\mathbf{f}_i}, \tilde{\mathbf{v}}_{\mathbf{f}_i}) \right] > \tau_r \quad (14)$$

651 then the i -th dimension is redundant for representing features \mathcal{F}_T , and $W_{i,:}^{pre}$ that is the i -th column
 652 of W^{pre} is a redundant parameter in the sense of representing features.

653 The idea is that if all features in task T preserve their semantics after perturbation, then the influence
 654 from the intervention on the i -th dimension to the semantic representation of \mathcal{F}_T can be ignored. It is
 655 a nice property akin to the robustness of population encoding to noise. Since information is encoded
 656 across many cells, a perturbation on a few cells will not destroy the representation [45].

657 If the representation $H = (h_1, h_2, \dots, h_m)^\top$ is followed by another weight matrix $W^{post} \in \mathbb{R}^{l \times m}$,
 658 then the following representation is $H' = W^{post} H \in \mathbb{V}^{D'} (D' = l)$. To study the influence of H on
 659 H' , we split H' into the sum of influences from the dimensions in H :

$$W^{post} H = \sum_{i=1}^m W_{:,i}^{post} h_i \quad (15)$$

660 Equation (15) shows that the influence from activation h_i to H' is subject to $W_{:,i}^{post}$. If the magnitude
 661 of the weights in $W_{:,i}^{post}$ is small, then the influence from h_i will be weakened, leading to a reduced
 662 impact on the final result in the follow-up computations. Thus, we provide the following definition:

663 **Definition 2: Forward redundancy** Given a lower bound $\tau_f \in (0, 1)$, we inspect the redundancy
 664 of the i -th dimension in \mathbb{V}^D in the forward propagation. If the activation h_i at that dimension satisfies:

$$\left[\frac{W_{:,i}^{post} h_i \cdot H'}{\|H'\|} \right]_{norm} < \tau_f \quad (16)$$

665 then dimension i is redundant, and $W_{i,:}^{pre}$ that is the i -th column of W^{pre} is a redundant parameter
 666 when considered in the forward propagation.

667 On the left side of (16) is the normalized projection of $W_{:,i}^{post} h_i$ on H' among the values in H , which
 668 measures the weight of activation h_i in the following representation H' , i.e., its influence in the
 669 forward propagation.

670 E.3 The representation redundancy at the node level

671 The definition of representation redundancy at the neuron level is shown in Section E.2, in which
 672 we set the activation value of a dimension to zero and check its redundancy through the semantic
 673 change in feature vectors. Specifically, considering an input with its features $\mathcal{F}_T = \{f_1, \dots, f_{|\mathcal{F}_T|}\}$,
 674 the determination of the representation redundancy of a node is:

$$\min_{1 \leq i \leq |\mathcal{F}_T|} [\Phi(\mathbf{v}_{f_i}, \tilde{\mathbf{v}}_{f_i}), \dots, \Phi(\mathbf{v}_{f_i}, \tilde{\mathbf{v}}_{f_i})] > \tau_r \quad (17)$$

675 where \mathbf{v}_{f_i} and $\tilde{\mathbf{v}}_{f_i}$ are the original and deformed feature vectors of the feature f_i , respectively. The
 676 metric Φ is used to measure the similarity between two vectors (e.g. cosine similarity), and $\tau_r \in (0, 1)$
 677 is the tolerance for deformation.

678 In this section, we discuss about the representation redundancy at the node level, i.e., each node
 679 consists of a group of neurons, or simply an activation (e.g. the output of an attention head). We
 680 define each node n as a group of N neurons, i.e., $n = (h_1, \dots, h_N)^\top$. The size of each node
 681 is at most an representation in a layer (e.g. a representation in the residual stream). Though a
 682 node could also be a layer in a model, we require $N \leq D$ to cover most of the cases for circuit
 683 discovery and for convenience of discussions. In this situation, the representation H is a set of nodes
 684 $\mathcal{N} = (n_1, \dots, n_{|\mathcal{N}|})$, where $|\mathcal{N}|$ is the number of nodes in H .

685 Therefore, the representation redundancy defined at the node level is almost the same as the original
 686 definition at the neuron level in section E.2, except that we focus on the redundancy of a node, instead
 687 of a single dimension. Hence, when inspecting the representation redundancy of a node n_i , we set
 688 the value of it to a zero vector and check the change in feature vectors. If Equation 17 is satisfied,
 689 then the node n_i is redundant for representing the features \mathcal{F}_T , and thus $W_{s_i, \cdot}^{pre}$ is redundant, where
 690 $s_i = [(i-1) \cdot N + 1, i \cdot N]$ is the slice in W^{pre} for the node $n_i \in \mathcal{N}$.

691 E.4 The forward redundancy at the node level

692 The definition of forward redundancy at the neuron level is shown in Section E.2. In this section, we
 693 discuss about the forward redundancy at the node level.

694 Consider a representation H in the representation space $\mathbb{V}^D (D = m)$ and a follow-up representation
 695 H' in the representation space $\mathbb{V}^{D'} (D' = l)$, the relation between H and H' is:

$$H' = W^{post} H = \sum_{i=1}^m W_{:,i}^{post} h_i \quad (18)$$

696 where W^{post} is the parameter matrix in the representation space $\mathbb{V}^{D'}$. In Equation 18, the representa-
 697 tion H' is split into the sum of the productions between each neuron h_i in H and its corresponding
 698 weight $W_{:,i}$, the i -th column of W^{post} .

699 Same as the settings in Appendix E.3, we define each node n as a group of N neurons, i.e., $n =$
 700 $(h_1, \dots, h_N)^\top$. Thus the representation H is a set of nodes $\mathcal{N} = (n_1, \dots, n_{|\mathcal{N}|})$. Similarly, the
 701 representation H' can be written as $\mathcal{N}' = (n'_1, \dots, n'_{|\mathcal{N}'|})$.

702 Therefore, the relation between H and H' can be written as:

$$H' = W^{post} H = \sum_{i=1}^{|\mathcal{N}|} W_{:,s_i}^{post} n_i \quad (19)$$

703 where $s_i = [(i-1) \cdot N + 1, i \cdot N]$ is the slice in W^{post} for the node $n_i \in \mathcal{N}$. Thus Equation (16) in
 704 the definition of forward redundancy can be extended to the node level:

$$\left[\frac{W_{:,s_i}^{post} n_i \cdot H'}{\|H'\|} \right]_{norm} < \tau_f \quad (20)$$

705 If Equation 20 is satisfied, then the node n_i is redundant in the forward propagation. Thus $W_{s_i, \cdot}^{pre}$ is
 706 redundant when considered in the forward propagation, where $s_i = [(i-1) \cdot N + 1, i \cdot N]$ is the
 707 slice in W^{pre} for the node $n_i \in \mathcal{N}$.

E.5 The representation redundancy in the approximation of indirect effect

The first part of the approximation of the indirect effect in Equation (3) is $n(x') - n(x)$, where $n(x')$ is often the averaged result of all other nodes. Since we consider the absolute value of this term when computing the overall node contribution described in Equation 2, we check the value of $n(x) - n(x')$ instead for convenience. The term $n(x) - n(x')$ is the difference between the original activation and the averaged activation, and it can be viewed as a centered activation $n(x)^c$.

To prove that the first part corresponds to the representation redundancy, we try to confirm that $n(x)^c$ is an indicator for representation redundancy. According to the approximation of the indirect effect in Equation 3 which is also the detection for the redundancy in a node, the smaller $n(x)^c$ is, the more likely that the node is redundant. Thus we expect that a smaller $n(x)^c$ to meet the requirement of representation redundancy.

When $n(x)^c$ approaches zero, the activation value $n(x)$ of the node n approaches the mean value $\overline{n(x)}$ of all the nodes. Thus we need to prove that when $n(x) \rightarrow \overline{n(x)}$, the node n is more likely to meet the requirement of the representation redundancy.

According to the assumption of feature sparsity in Section E.2, many features in a representation space do not frequently appear. Specifically, if all the features in the representation space \mathbb{V}^D is \mathcal{F} , then the features \mathcal{F}_T of an input satisfies $\mathcal{F}_T \subsetneq \mathcal{F}$. When the model gets larger with more dimensions, then the sparsity becomes much stronger, then we have $|\mathcal{F}_T| \ll |\mathcal{F}|$. The sparsity in features results in a long-tailed distribution of activation values in a representation, as demonstrated in [12]. Therefore, the average value of the activations in a representation could be quite small:

$$\overline{n(x)} = \mathcal{O}\left(\frac{1}{D}\right) \quad (21)$$

where D is the number of dimensions in the representation space \mathbb{V}^D . Hence the activation value (or norm) $n(x)$ of the node n is small when it approaches $\overline{n(x)}$. This implies that few features depend on that node: According to Equation (13), the activation of a dimension is the sum of the projections from all the feature vectors. When it comes to the node level, considering a node $n_i = (h_1, \dots, h_N)^\top$ which includes N dimensions in the representation, we have:

$$h_i = \left(\sum_{t=1}^{|\mathcal{F}_T|} a_t \mathbf{v}_{\mathbf{f}_t} \right) \cdot \mathbf{e}_i, \quad i = 1, \dots, N \quad (22)$$

A small norm of n_i implies small activation values $h_i (i = 1, \dots, N)$, which means the sum of the projections of the feature vectors is small, and thus the node n_i is more likely to be a redundant node for representing the features in \mathcal{F}_T . To prove this, we introduce P_r as the probability that the node n_i is redundant. According to the definition of representation redundancy in Appendix E.3, we have:

$$P_r = \prod_{t=1}^{|\mathcal{F}_T|} P[\Phi(\mathbf{v}_{\mathbf{f}_t}, \tilde{\mathbf{v}}_{\mathbf{f}_t}) > \tau_r] \quad (23)$$

For each feature $f_t \in \mathcal{F}_T$, the smaller the projection of its feature vector $\mathbf{v}_{\mathbf{f}_t}$ on the dimensions in node n_i , the larger the similarity $\Phi(\mathbf{v}_{\mathbf{f}_t}, \tilde{\mathbf{v}}_{\mathbf{f}_t})$ between $\mathbf{v}_{\mathbf{f}_t}$ and $\tilde{\mathbf{v}}_{\mathbf{f}_t}$, and thus the higher the probability $P[\Phi(\mathbf{v}_{\mathbf{f}_t}, \tilde{\mathbf{v}}_{\mathbf{f}_t}) > \tau_r]$ is when all the features in \mathcal{F}_T are considered.

Therefore, we can infer that when n_i approaches the mean value of all the nodes which is a small value due to feature sparsity, few features rely on the dimensions in n_i for representation, and thus the probability P_r that n_i is redundant is higher. Thus the centered value $n(x)^c$ of a node n serves as an indicator for the representation redundancy of that node.

E.6 The forward redundancy in the approximation of indirect effect

The second part of the approximation of the indirect effect in Equation (3) is:

$$\nabla_n \mathcal{L}_m[M(x)]|_{n(x)} \quad (24)$$

which is the derivative of the metric \mathcal{L}_m w.r.t the value of the node n .

747 Now we inspect a node n_i in the representation H . According to the chain rule and the discussions in
 748 Appendix E.4, the derivative in (24) can be written as

$$\nabla_{n_i} \mathcal{L}_m = \sum_{j=1}^{|\mathcal{N}'|} \nabla_{n'_j} \mathcal{L}_m \cdot \nabla_{n_i} n'_j \quad (25)$$

749 where the node $n'_j (j = 1, \dots, |\mathcal{N}'|)$ is the nodes in the follow-up representation, and we have:

$$n'_j = \sum_{i=1}^{|\mathcal{N}|} W_{s_j, s_i}^{post} \cdot n_i \quad (26)$$

750 where $s_j = [(j - 1) \cdot N + 1, j \cdot N]$ is the slice for the node n'_j . Then we have:

$$\nabla_{n_i} \mathcal{L}_m = \sum_{j=1}^{|\mathcal{N}'|} \nabla_{n'_j} \mathcal{L}_m \cdot W_{s_j, s_i}^{post} \quad (27)$$

751 Thus the second part in Equation (3) is actually the product of (i) W_{s_j, s_i}^{post} the weight corresponding
 752 to the node n_i and (ii) the derivatives of \mathcal{L}_m w.r.t all the nodes in H' , which echoes the left side of
 753 Equation (20), and the only difference is that the latter is the normalized result among all the nodes
 754 in H . The difference is not bothering, since activation patching compares all the contributions of
 755 nodes together, and the normalization operation exists only for convenience in the definition of the
 756 forward redundancy. Thus, the second part of Equation 3 could serve as an indicator for the forward
 757 redundancy.

F Details for the Subject-verb Disagreement Task

To be brief, the goal of this task is to change the grammar in a language model from (a) to (b) as follows:

- (a) *We apologize, but this video has failed to load.*
 (b) *We apologizes, but this video have failed to load.*

In the example above, *We* and *this video* are subjects, *apologize / apologizes* and *has / have* are verbs, and *We* and *video* are also called the END tokens that appear before the verbs. The change from *apologize* to *apologizes* or from *has* to *have* is called a flip.

F.1 Data Preparation

We use the first 10k samples from Pile [17], which consists of 22 smaller, high-quality datasets. Firstly, in order to get relatively simple and clean sentences, we filter out the content in Github, ArXiv, PubMed Abstracts, PubMed Central, StackExchange, USPTO Backgrounds, Pile-CC, DM Mathematics, and FreeLaw. Thus we do not include code or complex formulas in our data. Secondly, we split the corpus with periods '.' as intervals. We remove links to websites, images, and other files. We also remove sentences that are too short (less than 25 characters). Thirdly, we leave only the sentences in the present tense in English and ensure that each sentence is a complete sentence with a punctuation like '.', '?', or '!' at the end. Finally, for each verb in the present tense in a sentence, we convert it to its opposite form. That is, we convert a verb with a part of speech VBP like *go* to VBZ like *goes*, and vice versa. For *be* (*am / is / are*), we flip them following: *am* \rightarrow *is*, *is* \rightarrow *are*, *are* \rightarrow *am*.

We collect 60,000 samples in total. For experiments, we only use half of the data, which is further split for training (2.4w), validation (3k), and test (3k). All the details for data preparation can be found in our code.

F.2 Experiment Settings

We use GPT2-small [47] for this task. GPT2-small is a decoder-only transformer with 12 layers and 12 attention heads per attention layer. We set the output of the attention and the MLP in each layer as upstream nodes, and the input of the query, key, value, and the MLP in each layer as downstream nodes. This is because the query, key, and value input for an attention head can only affect downstream nodes via the attention output of that head, so the upstream nodes can only be attention head outputs, which is also discussed in [27]. The parameters to update correspond to the upstream and downstream nodes at both ends of the edges, as discussed in Section 4.1. Details are shown in Table 3.

Table 3: The settings of the nodes and their corresponding parameters in the subject-verb disagreement task. The number of layers $L = 12$ and the number of attention heads in each layer $H = 12$. The notations for parameters are detailed in Appendix C.

| Nodes | Upstream | | Downstream | |
|------------|------------------------------|-------------|-------------------|------------|
| | $Attn_i^j(x)$ | $MLP_i(x)$ | $x_{Q/K/V}^{i,j}$ | x_{in}^i |
| Parameters | $W_O^{i,j}$ | W_{out}^i | $W_{Q/K/V}^{i,j}$ | W_{in}^i |
| Range | $i \in [0, L), j \in [0, H)$ | | | |

For all experiments, we set the learning rate to 1e-3, and batch size to 16. We use mini-batch SGD with a momentum equal to 0.9 as the optimizer. Each model is trained for 3 epochs, with 100 steps in the beginning for warmup. During training, we evaluate on the valid set every 100 steps. The metric \mathcal{L}_m for measuring the flip from subject-verb agreement to disagreement is the logit difference at the END token, which is:

$$\mathcal{L}_m = \text{logit}(W_{flip}|W_{END}) - \text{logit}(W_v|W_{END})$$

790 where W denotes the tokens in a sentence. For example, the case “We apologize, but this video has
791 failed to load.” contains two logit differences: $\text{logit}(\text{apologizes}|We) - \text{logit}(\text{apologize}|We)$ and
792 $\text{logit}(\text{have}|video) - \text{logit}(\text{has}|video)$. In practice, we only consider the verbs that are tokenized as
793 a single token.

794 As for the calculation of edge contribution, we follow [56] and use mean ablation when patching a
795 node. That is to say, for each activation of shape (batch_size, seq_len, d_model), we replace the
796 value at the END token position in each sample with the mean value of all tokens in all samples in a
797 batch.

798 Experiments are conducted on $4 \times$ A40 Nvidia GPUs.

799 F.3 Analysis of the Quality of the Discovered Circuit

800 As discussed in Section 5.1.3, the performance cannot be further improved at $N = 1000$, where N is
801 the number of edges saved in circuit discovery. To show this intuitively, the changes with N of the
802 logit difference, PPL, and the ratio of the trainable parameters are illustrated in Figure 7. We can
803 observe that there is a sharp turning point at $N = 1000$, where the curves start to be flat. This serves
804 as a sign that there does exist a circuit that includes all the parameters responsible for the subject-verb
805 disagreement task.

806 To better prove this conclusion and demonstrate the high quality of the circuit found in our method,
807 we provide another two experiments below.

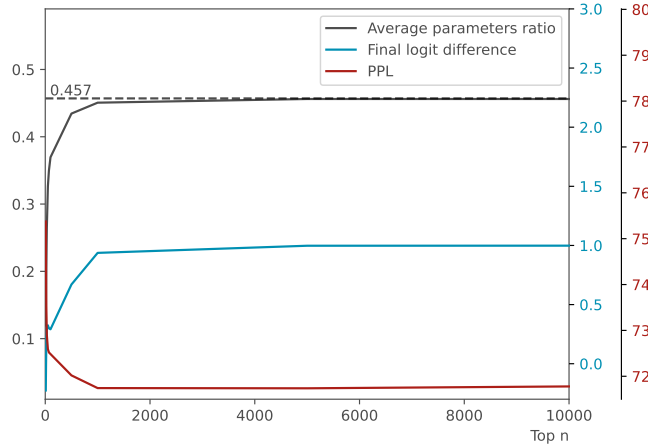


Figure 7: The influence from the setting of top N edges.

808 F.3.1 Faithfulness and Completeness

809 Faithfulness and completeness examine a circuit from two different views. Faithfulness tells how
810 much performance a circuit gets, while completeness tells how much performance a circuit fails to
811 capture. Consider a model M with its computational graph \mathcal{G} , a circuit \mathcal{C} for a specific task T and a
812 metric \mathcal{L}_m for measuring the output of the model, following the definition in [34], the faithfulness of
813 the circuit \mathcal{C} is

$$\frac{\mathcal{L}_m[M(\mathcal{C})] - \mathcal{L}_m[M(\emptyset)]}{\mathcal{L}_m[M(\mathcal{G})] - \mathcal{L}_m[M(\emptyset)]} \quad (28)$$

814 in which $M(*)$ denotes the forward pass of model M with the nodes outside $*$ mean-ablated, and \emptyset
815 denotes an empty circuit. The completeness is defined as

$$\frac{\mathcal{L}_m[M(\mathcal{G} \setminus \mathcal{C})] - \mathcal{L}_m[M(\emptyset)]}{\mathcal{L}_m[M(\mathcal{G})] - \mathcal{L}_m[M(\emptyset)]} \quad (29)$$

816 where $\mathcal{G} \setminus \mathcal{C}$ denotes the complementary set of circuit \mathcal{C} . The completeness of circuit \mathcal{C} is actually the
817 faithfulness of circuit $\mathcal{G} \setminus \mathcal{C}$.

818 In practice, we calculate the faithfulness and completeness of the circuits for subject-verb disagreement
819 at $N = 100, 500, 1000, 5000, 10000$ edges. Results are shown in Figure 8. It can be seen
820 that $N = 1000$ also serves as a turning point for the curves of faithfulness and completeness. The
821 faithfulness of the circuits remains relatively high after $N = 1000$, ensuring the high quality of
822 circuits.

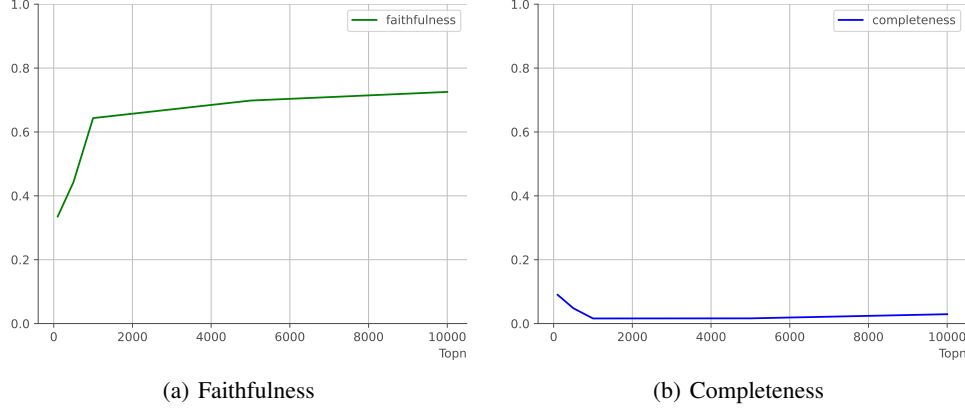


Figure 8: The faithfulness and completeness of the circuits for subject-verb disagreement.

823 F.3.2 Ablation Study of Random activation

824 Random activation means during training, we randomly unfreeze some parameters outside the circuit.
825 Since we assume that the circuit with $N = 1000$ edges already includes all needed parameters
826 for the subject-verb disagreement task, we randomly select a part of the parameters outside the
827 $N = 1000$ circuit and involve them in optimization. In practice, we randomly activate 10%, 20%,
828 30% and 40% of the outside parameters, and compare the results with before. Results are shown
829 in Figure 9. When 10% of the parameters outside the circuit are activated, the result is almost the
830 same as before. When the ratio gets larger, we observe that the PPL is higher than before when
831 random activation is performed, though the logit difference increases. This is because when extra
832 parameters are summoned to fit the new data distribution, the original functions corresponding to
833 those parameters may be destroyed. Thus the performance is improved at the expense of harming
834 other abilities. Therefore, the circuit we find at $N = 1000$ edges is almost the exact circuit for
835 subject-verb disagreement.

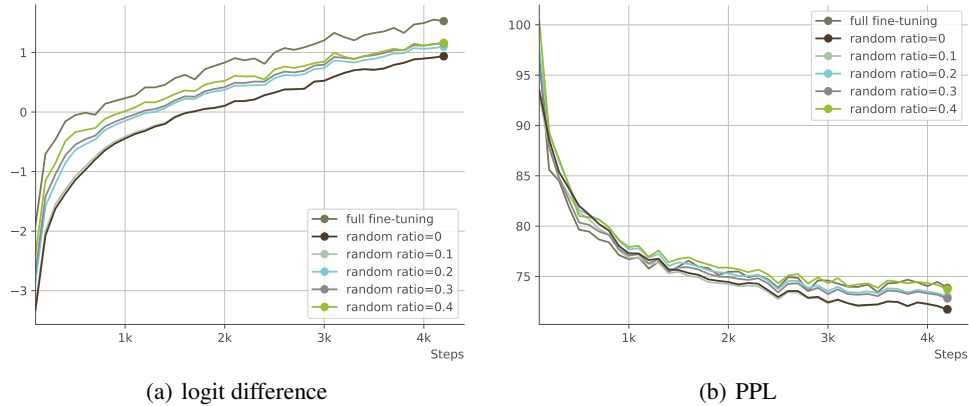


Figure 9: Experiment results of random activation.

836 F.4 Analyses of the Training Dynamics

837 F.4.1 Interpret the Circuit for the Subject-verb Disagreement Task

838 To interpret the circuit for the subject-verb disagreement task, we first analyze this task from the
 839 human perspective. To decide the form of a verb, we need to (i) find out the subject in the context and
 840 (ii) adjust the form of the verb according to the attributes of the subject, including the person attribute
 841 and the number attribute. Therefore, we assume that there exist at least two kinds of attention heads
 842 responsible for the two functions above respectively. We name the two kinds of attention heads as
 843 the **Subject Identification Heads** and the **Subject Attribute Heads**. Note that we only focus on
 844 self-attention instead of MLP because only attention layers move information across tokens, which is
 845 important for completing this task. Besides, each of the whole MLP layers is regarded as a node in
 846 the circuit in our experiment, thus we do not expect to figure out any specific function from it.

847 Next, we look for the two kinds of heads discussed above.

848 **Subject Identification Heads** To find out the Subject Identification Heads, we check the attention
 849 pattern of each attention head at the END token since the END token is directly used for predicting
 850 the verb token. We sort the heads in descending order according to their attention weights from the
 851 END token (query) to the subject tokens (key). Then we keep the heads in which the attention is
 852 mainly paid to the subject in the context.

853 We find that there exist two types of Subject Identification Heads. The type I heads mainly attend to
 854 the last token itself at the END token, so the attention pattern is a diagonal line. This type of head is
 855 helpful when the subject is exactly the END token, e.g. “He is ...”, “The girls are ...”, etc. The type
 856 II heads attend to the subject which is several tokens before the END token, e.g. “The kid who is
 857 holding an ice cream in hand is ...”, “The famous scientist, who is also an artist, has ...”, etc. The type
 858 II heads obviously have the ability of syntactic analysis and subject identification, while the type I
 859 heads may just happen to attend to the subject that overlaps with the last token.

860 The type I Subject Identification Heads in GPT2-small includes head.0.1, head.0.3, head.0.5, etc.,
 861 while the type II heads include head.8.5, head.10.5, head.10.9, head.11.8, etc. The attention patterns
 862 for both kinds of heads are shown in Figure 10(a) and Figure 10(b) respectively.

863 It is worth noticing that the Subject Identification Heads remain unchanged over the training process,
 864 which means their function is preserved and shared between subject-verb agreement and subject-verb
 865 disagreement.

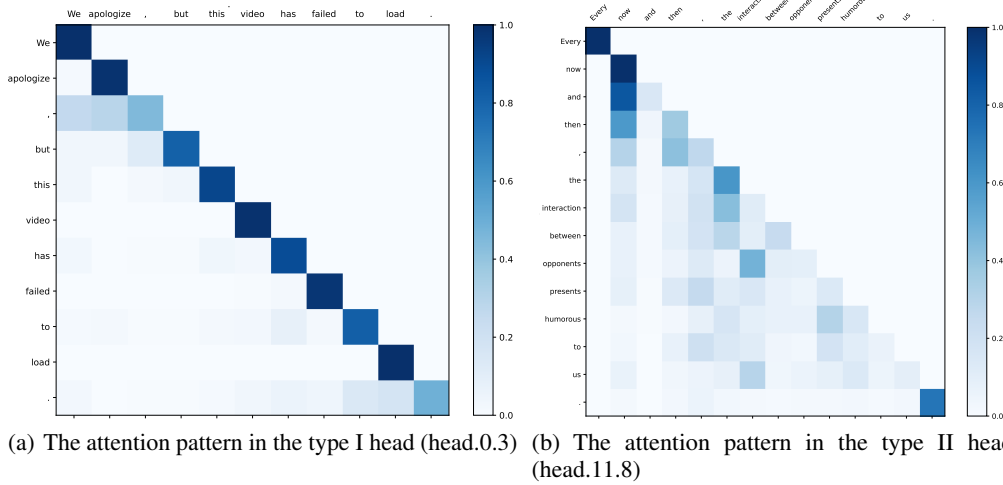


Figure 10: The examples of the attention patterns in the Subject Attribute Heads. In Figure 10(a), the END token “video” attends to the subject “video” which is also the END token itself. In Figure 10(b), the END token “opponents” mainly attends to the subject “interaction”.

Subject Attribute Heads To find out the Subject Attribute Heads, we need to find out which heads directly affect the match between the verb and the subject, which is measured by the logit difference between the flipped verb and the original verb at the END token. Suppose the output of the final layer at the END token is $x_{END} \in \mathbb{R}^D$, then the logit difference is $[W_U(v_{flip}) - W_U(v)] \cdot x_{END}$, in which $W_U(v_{flip}) - W_U(v)$ is called the logit lens [40]. Since the logit lens is fixed, we expect the projection of x_{END} on the direction of the logit lens to be large, thus encouraging the probability difference between the two opposite verb forms (love v.s. loves, etc.). As discussed in Appendix C, the output x_{END} which is in the residual stream can be decomposed into the linear addition of the outputs from the previous layers. Therefore, the output of a Subject Attribute Head is a part of x_{END} and would encourage the value of the logit difference. Thus, a Subject Attribute Head is an attention head that has a large dot product value with the logit lens.

In practice, we calculate the dot product between the logit lens $W_U(v_{flip}) - W_U(v)$ and the output $Attn_i^j(x)$ from each attention head in each layer over a batch of samples. The result is shown in the main text in Figure 3. The darker the color, the larger the absolute value of the dot product is, which implies that the head is more likely to be a Subject Attribute Head. We observe that head.6.0, head.6.5, head.8.5, head.10.9, and so on see obvious flips (Figure 3(a)) from positive to negative or the opposite direction, which implies that they are directly responsible for the match between the subject and the verb. During training, the parameters inside these heads adjust themselves to the new data distribution, while their function type remains unchanged, which is an interesting finding of the self-regulation ability inside the model.

Collaborative Heads Finally, we notice that some of the Subject Attribute Heads (head.8.5, head.10.9, etc.) are also Subject Identification Heads, which means they also attend to subject tokens. We wonder if there exist some heads in previous layers that could influence the behavior of the Subject Attribute Heads. That is to say, the Subject Attribute Heads do not act alone but collaborate with other heads.

To find out these heads, we knock out the upstream heads one at a time at the END token using mean ablation. We observe the change in the attention pattern of each Subject Attribute Head and then keep the heads that bring obvious changes in the attention patterns. In practice, we focus on subject-verb agreement and only check the influence on head.8.5 and head.10.9. We also provide two types of data, corresponding to the two cases discussed in Appendix F.4.1, in order to provide a more detailed analysis. Results show that

- When patching on the type I data in which the END token is exactly the subject, head.1.2 and head.2.11 affect both head.8.5 and head.10.9
- When patching on the type II data in which the subject is several tokens before the END token, head.7.4 and head.2.11 affect head.8.5, while head.2.11 affects head.10.9.
- When we mix the two types of data, we find that head.1.3, 0.8, 2.10, and 6.5 affect head.8.5, while head.1.3, 1.4, 1.6, 6.5, 0.8, and 0.9 affect head.10.9. These heads may be responsible for both types of data while not specifically responsible for a certain type of data, so they appear when patching on the mixed data.

We notice that though the influence of the above heads is relatively large, the absolute influence is sometimes quite small. Therefore, we further conduct an experiment in which we patch multiple heads at a time and check the influence of them on head.8.5 and head.10.9. Results show that when upstream heads are patched together, their combined effect is much higher than the individual effect. Thus we call these heads the Collaborative Heads.

F.4.2 The Interaction and Collaboration Inside the Model

From the discussions above, we can see that the nodes inside a circuit complete a task through a cooperative division of labor. We summarize the interaction and collaboration inside the model as follows:

1. Each node is responsible for a single or multiple functions. As discussed in Appendix F.4.1, we have found attention heads that are responsible for identifying the subject in a sentence or adjust the form of a verb according to the attributes of the subject, or both. Each head has its division of labor when completing a task.

- 918 2. Some heads directly affect the output, while others cooperate with them to affect the output
919 indirectly. For example, head.8.5 directly matches the verb with the subject, while head.7.4,
920 head.1.3 and so on indirectly affect the output through cooperation with head.8.5.
- 921 3. Several nodes achieve a common goal through cooperation. For example, head.1.3, 1.4, 1.6,
922 6.5, 0.8, and 0.9 affect head.10.9 through combined effect, which means their influence on
923 head.10.9 only appears when they act together.

924 F.4.3 The Evidence of Hebbian Learning

925 As discussed in Section 5.1.5, we observe that some edges in the circuit are strengthened or weakened
926 during training, just like the Hebbian learning proposed in [10] in neuroscience. As stated by Hebb, a
927 synapse between two neurons is strengthened when the neurons on either side of the synapse have
928 highly correlated outputs, which means they are often activated synchronously. The theory is often
929 concluded as “Cells that fire together, wire together” [54]. For two neurons i and j , a common
930 description of hebbian learning is as follows:

$$w_{ij} = \frac{1}{p} \sum_{k=1}^p x_i^k x_j^k \quad (30)$$

931 where w_{ij} is the weight of the connection between the two neurons, and x_i^k and x_j^k are the k -th inputs
932 for i and j respectively. When it comes to the computational graph of a model, the nodes and edges
933 in the graph could be viewed as the neurons and their connections in a brain from the perspective of
934 neuroscience.

935 During training, we find that some of the edges are obviously stronger than others, which means
936 they have higher edge contributions. Besides, they are strengthened all the way during training.
937 Specifically, we analyze the circuits during training in the subject-verb disagreement task, with the
938 setting of top $N = 1000$ edges. We check the results at 2000, 3000, and 4000 steps respectively, and
939 visualize the circuits with the top 35 edges in Figure 4. Note that the thickness of an edge corresponds
940 to the logarithm of the edge contribution, that is $\log[1 + c(e)]$. The details are shown in Table 1.

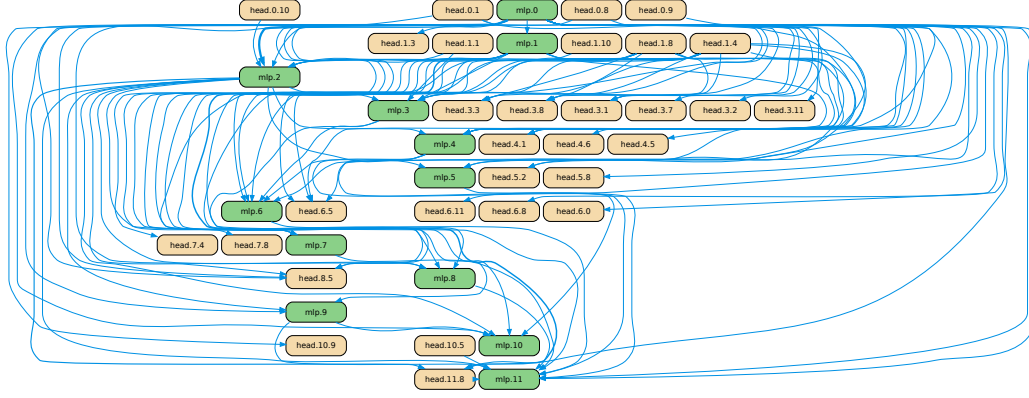
941 We also find that the edge contribution may keep decreasing during training. This is quite similar
942 to the self-organization of cells inside human brains, a well-known phenomenon of which is the
943 lateral inhibition among neurons [22], which means an activated neuron can reduce the activity of
944 its neighbors. Inspired by this, [25] developed the self-organizing maps (SOM), an unsupervised
945 algorithm that leverages the Winner-Take-All strategy to perform competitive learning. The core
946 ideas behind SOM are:

- 947 • The neurons inside a neural network learn to represent data through competition, i.e. given
948 an input, some neurons are activated while others are inhibited.
- 949 • Different inputs are represented in a topologically ordered manner, i.e. different neurons are
950 responsible for different features in a well-organized style.

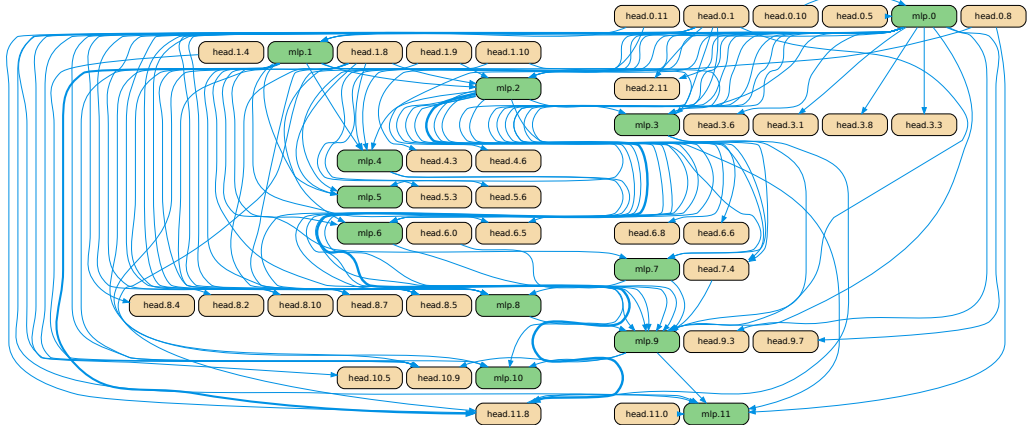
951 In our study, we find that the dynamic change of edges echoes the above discussions on competition
952 and self-organization. During training, some connections between nodes are strengthened, which
953 may reduce the intensity of other connections. After training, the components inside a model have
954 reorganized themselves to adapt to the new data distribution. When faced with an input, different
955 regions inside the computational graph are responsible for different subtasks and collaborate to
956 complete a goal, as discussed in Appendix F.4.1 and Appendix F.4.2.

957 F.4.4 The Circuits Before and After Fine-tuning

958 We present the circuits of the subject-verb disagreement task before and after fine-tuning in Figure
959 11(a) and Figure 11(b) respectively. The circuit of subject-verb disagreement is trained under the
960 setting of $N = 1000$ edges, and for both of the circuits we only present the top 100 edges. It can be
961 seen that most of the heads we discussed in Appendix F.4.1 are included in the circuit. Besides, it
962 is obvious that the circuits before and after fine-tuning share a lot of nodes, which implies that the
963 function shift from subject-verb agreement to disagreement happens mostly in the **polarity change**
964 of nodes, instead of randomly assigning the ability to other nodes. This is an intuitive finding, which
965 not only demonstrates the rationality of circuit-tuning as well as our analyses but also provides new
966 insights for our understanding of the mechanism inside language models.



(a) The circuit before fine-tuning (subject-verb agreement).



(b) The circuit after fine-tuning (subject-verb disagreement).

Figure 11: The circuits of subject-verb agreement (top) and subject-verb disagreement (bottom).

967 F.4.5 The Experiment Results after the Revision of Attribution Score

968 During the analyses in Appendix F.4.1, we find that the original definition of the attribution score in
 969 EAP [56] fails to capture all the relevant edges in a task. For example, head.6.0 which is a Subject
 970 Attribute Head fails to appear in the circuit. We assume that there exists a situation where an important
 971 node is connected with many other nodes, but each edge is not that strong. For example, as illustrated
 972 in Figure 12, an upstream node n_a^u is connected with four downstream nodes, while another upstream
 973 node n_b^u is connected with only one downstream node. Since the edge between n_b^u and n_5^d is stronger
 974 than any edge between n_a^u and the nodes connected with it, the edge $n_b^u \rightarrow n_5^d$ may be kept in the
 975 circuit, while the edges in $\mathcal{E}_a = \{n_a^u \rightarrow n_i^d | i = 1, 2, 3, 4\}$ may be left out. As a result, n_a^u is not
 976 involved in optimization, though the sum of the edge contributions of all edges in \mathcal{E}_a may be almost
 977 the same or even larger than that of $n_b^u \rightarrow n_5^d$.

978 Thus we calculate the edge contribution of an edge $e : n_i^u \rightarrow n_j^d$ as below:

$$c(e)' = c(e) \cdot \sum_{k=1}^{N_{down}^i} c(n_i^u \rightarrow n_k^d) \cdot \sum_{k=1}^{N_{up}^j} c(n_k^u \rightarrow n_j^d) \quad (31)$$

979 where $c(e)$ is the original attribution score in EAP, N_{down}^i is the number of the downstream nodes of
 980 n_i^u , N_{up}^j is the number of the upstream nodes of n_j^d . The revision considers the contributions from

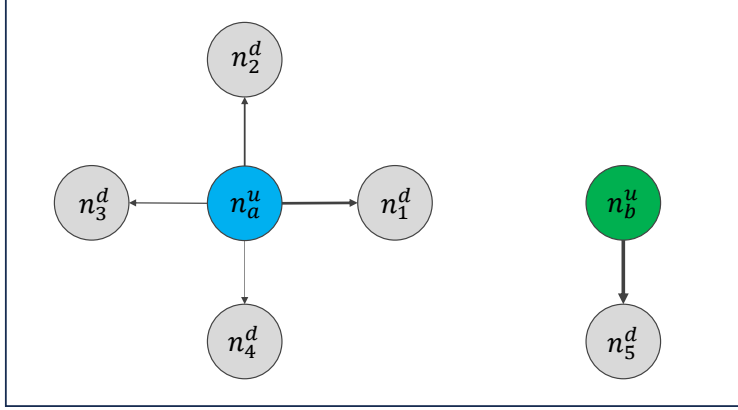


Figure 12: A sketch for the idea behind the revision on attribution score.

all the edges connected to the upstream node and the downstream node. To verify it, we conduct a new experiment on the subject-verb disagreement task and compare it with the result before. Results are shown in 13. Details can be found in Table 4. Compared with the original attribution score, our method improves the logit difference steadily, while even bringing down the computation to some extent.

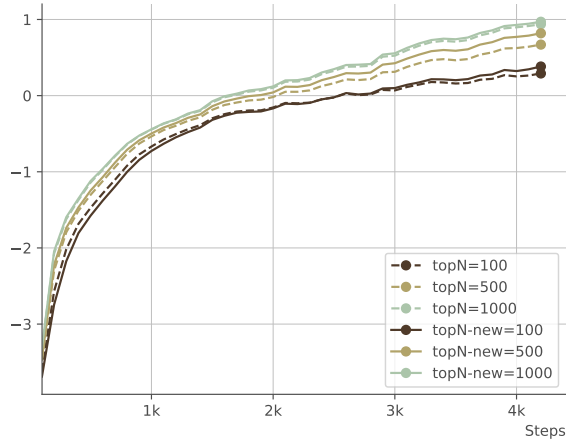


Figure 13: The experiment result after the revision on attribution score.

Table 4: Comparison between the performance before and after the improvement on attribution score.

| Top n & Methods | Original EAP | | | Improved EAP | | |
|-----------------|------------------|-------|----------------------|------------------|-------|----------------------|
| | logit difference | PPL | Avg. Param ratio (%) | logit difference | PPL | Avg. Param ratio (%) |
| 50 | 0.292 | 72.59 | 32.61 | 0.350 | 72.71 | 26.88 |
| 100 | 0.291 | 72.50 | 36.96 | 0.382 | 72.53 | 34.12 |
| 500 | 0.670 | 72.02 | 43.50 | 0.819 | 72.05 | 42.50 |
| 1000 | 0.937 | 71.74 | 45.61 | 0.970 | 71.73 | 45.55 |

G Details for the Complex Tasks

G.1 Details for Task Settings

G.1.1 Reasoning-based Tasks

Mathematics We use GSM8K [8] as the dataset, which contains about 8.5k grade school math problems with natural language solutions. The answer to a problem not only contains the final answer but also provides the process for solving the problem. An example of this task is shown in Table 5.

Table 5: An example in the GSM8K dataset.

| Question | Answer |
|---|--|
| Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers’ market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers’ market? | Janet sells $16 - 3 - 4 = << 16 - 3 - 4 = 9 >>$ 9 duck eggs a day. She makes $9 * 2 = \$ << 9 * 2 = 18 >>$ 18 every day at the farmer’s market. ####18 |

During training, the NLL loss also serves as the metric \mathcal{L}_m for measuring the output of the model. For evaluation, we use Acc@1 as the metric, which means for each problem in the test set we only sample one answer from the model.

Logical Reasoning We use Contexthub [21] as the dataset, which consists of problems of 4 difficulty levels, including deductive and abductive reasoning in 12 distinct categories or domains from Wikipedia. The problems together with their reasoning processes are instantiated automatically by LLMs following the fixed formal logic templates. The whole dataset contains 18,240 samples. We only use level 1 and level 2 in our experiment for convenience, which include 6720 samples in total. An example of this task is shown in Table 6.

Table 6: An example in the Contexthub dataset.

| Item | Template | Instantiation |
|----------|---|--|
| Premise | <aaa> | The Sahara desert receives heavy rainfall this year. |
| | <aab> | The Amazon rainforest experiences severe drought conditions. |
| | <aac> | Some of Earth’s major ecosystems are undergoing significant changes in weather patterns. |
| Question | (aaa OR aab) \rightarrow aac. Given aac is False, what is the value of aab? | If either the Sahara desert receives heavy rainfall this year or the Amazon rainforest experiences severe drought conditions, then it implies that some of Earth’s major ecosystems are undergoing significant changes in weather patterns. Given that it is false that some of Earth’s major ecosystems are undergoing significant changes in weather patterns, what can be determined about the Amazon rainforest experiencing severe drought conditions this year? (True, False, or N/A (undetermined)). |
| | (aaa OR aab) \rightarrow aac = False. Given aac is False, the value of premise (aaa OR aab) is False, thus, the value of aab is abduced as False. Thus, the answer is False | “The Sahara desert receives heavy rainfall this year” or “The Amazon rainforest experiences severe drought conditions”) logically implies “Some of Earth’s major ecosystems are undergoing significant changes in weather patterns” whose corresponding truth value is False. Given “Some of Earth’s major ecosystems are undergoing significant changes in weather patterns” is False, the value of premise (“The Sahara desert receives heavy rainfall this year” or “The Amazon rainforest experiences severe drought conditions”) is False, thus, the value of “The Amazon rainforest experiences severe drought conditions” is abduced as False. Thus, the answer is <answer>False</answer> |

During training, the NLL loss also serves as the metric \mathcal{L}_m for measuring the output of the model. For evaluation, we use the average F1 score over all categories of data.

For all reasoning-based tasks, we add an instruction “Please answer step by step.” at the end of the question in order to guide the model to answer the question step by step.

1005 G.1.2 Reasoning-free Tasks

1006 **Gender De-biasing** According to [16], there are various kinds of expressions in social bias. In
 1007 this study, we focus on the gender bias in occupations. We aim to break down the binary gender
 1008 stereotype of a model. For example, given a sentence “the doctor put on [PRP] coat” where [PRP] is
 1009 a possessive pronoun, we expect the model to choose *his* or *her* with equal probabilities.

1010 During fine-tuning, the model learns to predict the next word in an auto-regressive way, thus we
 1011 expect the model to balance the probabilities between male attribute words (he/his/him/himself) and
 1012 female attribute words (she/her/herself) at the END token when generating the next token. Therefore,
 1013 we use the logit difference between the male attribute words and female attribute words at the END
 1014 token as the metric \mathcal{L}_m for measuring the output of the model. Specifically, for each sample, we
 1015 calculate the logit difference between the pronoun and the anti-pronoun, which is the pronoun in
 1016 the opposite gender. For example, in the case “the doctor put on [PRP] coat”, the logit difference is
 1017 $\text{logit}(W_{her}|W_{END}) - \text{logit}(W_{his}|W_{END})$, where W_{END} is the END token *on*.

1018 We use BUG [30] for training, which is a large-scale dataset of sentences sampled from real-world
 1019 corpora. Each sentence is marked with an occupation and the pronouns referring to it. In practice, we
 1020 use the “balanced BUG” provided in the dataset which includes 2.5w sentences randomly sampled
 1021 from BUG to ensure balance between male and female entities and between stereotypical and non-
 1022 stereotypical gender role assignments. We perform coreference resolution ourselves to filter out the
 1023 samples in which the coreference is not right, and leave 1.5w samples for training, which contains
 1024 151 types of occupations and each sentence only contains one (occupation, pronoun) pair.

1025 Though the training set we use is balanced between genders and stereotypes, the number of samples for
 1026 each occupation is not balanced. To further improve performance, we additionally add a regularization
 1027 term to the original NLL loss. Then the total loss is

$$\mathcal{L} = \mathcal{L}_{NLL} + \beta \cdot |\text{logit}(W_{pron}|W_{END}) - \text{logit}(W_{anti-pron}|W_{END})| \quad (32)$$

1028 in which β is a hyper-parameter for controlling the weight of regularization. The absolute value of
 1029 the logit difference aims to minimize the difference between genders in the model’s stereotype.

1030 For evaluation, we use WinoBias [65] which is a classic dataset for coreference resolution focused
 1031 on gender bias. There are two types of sentences in WinoBias. The Type 1 sentences require
 1032 world knowledge related to the context to perform coreference resolution, e.g. “The farmer knows
 1033 [the editor] because [he] is really famous”. The Type 2 sentences can be resolved using syntactic
 1034 information, e.g. “The CEO called [the hairdresser] and paid [her] over the phone”. In practice, we
 1035 only use the Type 2 sentences to avoid ambiguity. The test set contains 40 types of occupations in
 1036 total.

1037 To better evaluate the performance of gender de-biasing, we adopt the concept of prejudice risk
 1038 from [33] which is used to measure the stereotype in large language models. Specifically, given
 1039 an occupation $x \in X$, a binary gender attribute $y \in \{male, female\}$ and a context $c \in C$, the
 1040 stereotype of a model M against x about y in the context c is

$$s_{y|x}^M(c) = \frac{p_{y|x}^M(c)}{p_{y|x}^*(c)} - 1 \quad (33)$$

1041 where $p_{y|x}^*(c)$ is the attribute prediction probability of the unbiased model, thus $p_{y|x}^*(c) = 0.5$ when
 1042 binary gender is considered. The definition of prejudice risk is

$$R^p = \mathbb{E}_{x \sim X}(r_x^p) \quad (34)$$

1043 where $r_x^p = J(\mathbb{E}_{c \sim C}(s_{y|x}^M(c)))$ is the prejudice risk of one occupation x , and $J(s_{y|x}^M(c)) =$
 1044 $\max_{y \in Y} \{s_{y|x}^M(c), 0\}$ is the discrimination risk criterion. For details, please refer to [33].

1045 **Reading Comprehension** We use SQuAD 2.0 [48] as the dataset. The input contains a paragraph
 1046 from a passage, and a question related to that paragraph. The answer could be a word or a phrase in
 1047 the paragraph, or “<No Answer>” which means the answer cannot be found in the paragraph. An
 1048 example for this task is shown in Table 7.

1049 During training, the NLL loss serves as the metric \mathcal{L}_m for measuring the output of the model, since
 1050 the answer is a segment of text which may contain one or multiple tokens. For evaluation, we use the
 1051 development set for convenience, and the metric is exact match and F1 score.

Table 7: An example in the SQuAD 2.0 dataset.

| Item | Content |
|-----------|--|
| Paragraph | The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries. |
| Question | In what country is Normandy located? |
| Answer | France |

G.2 Details for Implementation

We use Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct [11] for this task. We set the output of the attention and the MLP in each layer as upstream nodes, and the input of the query, key, value, and the MLP in each layer as downstream nodes. Different from GPT2-small, an MLP layer in Llama is too big to be a node, so we split the input and output of each MLP layer into 64-dimensional MLP heads. Details are shown in Table 8.

Table 8: The settings of the nodes and their corresponding parameters in the complex tasks. For model sizes 1B/3B/8B, the number of layers $L = 16/28/32$, and the number of attention heads in each layer $H = 32/24/32$, and the number of MLP heads in each layer $H^* = 128/128/224$. The notations for parameters are detailed in Appendix C.

| Nodes | Upstream | | Downstream | | |
|------------|--|-----------------|-------------------|------------------|----------------|
| | $Attn_i^j(x)$ | $MLP_i^k(x)$ | $x_{Q/K/V}^{i,j}$ | $x_{pre}^{i,k}$ | $x_{in}^{i,k}$ |
| Parameters | $W_O^{i,j}$ | $W_{out}^{i,k}$ | $W_{Q/K/V}^{i,j}$ | $W_{gate}^{i,k}$ | $W_{in}^{i,k}$ |
| Range | $i \in [0, L), j \in [0, H), k \in [0, H^*)$ | | | | |

As for the calculation of edge contribution, we use mean ablation as before when patching a node. For reasoning-based tasks and the reading comprehension task in reasoning-free tasks, there is no such thing as the END token. Therefore, for each activation of shape (batch_size, seq_len, n_head, d_model), we take all tokens into consideration and use the mean value over all tokens and all samples for mean ablation. For implementation details, please refer to our source code.

For circuit-tuning, we set the number of edges N to 2000, 3000, and 4000 for 1B/3B/8B models respectively. The choice of N is considered with the size of the models and the complexity of the tasks. In practice, we tried several values of N and found that the task performance generally converges at these values. Actually, the value of N corresponds to the intrinsic dimension which is the minimal dimensionality required in a solution space. If the chosen value is smaller than that, then the result may not be optimal. If the value exceeds that, then the improvement in the final result may be quite small, which is demonstrated in Section 5. In fact, the choice of N is quite similar to the choice of rank r in LoRA, except that our method is able to localize the key parameters more accurately.

The batch size is set to 16 in all experiments. We set the learning rate to 3e-5 for the mathematics and reading comprehension tasks, and 1e-4 for other tasks. Mini-batch SGD with a momentum equal to 0.9 is used as the optimizer. During training, we perform circuit discovery every 8 steps after optimization for efficiency, which is different from the experiments on GPT2-small in which we perform circuit discovery right after an iteration step. For each task, we train the model until performance cannot be further improved.

For LoRA, we set $r = 32$, $\alpha = 64$ for all experiments, since this is the best setting we could get. We have swept a wide range of values for rank r and alpha α , and find that the performance cannot be

1081 further improved or even decreases when r increased over 32. Actually, Hu et al. [20] found similar
1082 phenomenon when they studied LoRA. From our opinion, this is because LoRA fails to accurately
1083 figure out the key parameters needed to be fine-tuned, since all the parameters are changed after
1084 LoRA fine-tuning.

1085 For full fine-tuning and LoRA, we use the same optimizer as that in circuit-tuning. For all other
1086 hyper-parameters such as learning rate, batch size, training steps, and so on, we just sweep over a
1087 range of choices and choose the best ones.

1088 In practice, we find that circuit-tuning is much more stable than full fine-tuning and LoRA. When we
1089 sweep over a range of hyper-parameters, we notice that full fine-tuning and LoRA are quite sensitive
1090 to the change of learning rate, batch size, training steps, and so on. When it comes to circuit-tuning,
1091 the change in evaluation result is relatively moderate while still maintaining good performance.

1092 Experiments are conducted on $8 \times$ A800 Nvidia GPUs.

1093 G.3 Details for Evaluations on General Capabilities

1094 To demonstrate that our method is good at preserving general capabilities, we test the fine-tuned
1095 models on a set of benchmarks involving general capabilities as well as other capabilities.

1096 For general capabilities, we use MMLU [19], Winogrande [52] and IFEval [66]. For MMLU, the
1097 evaluation metric is the average accuracy over all categories. For Winogrande, we use the development
1098 set for convenience, and the evaluation metric is accuracy. For IFEval which is to test the instruction
1099 following ability of a model, each prompt contains one or multiple verifiable instructions, thus the
1100 evaluation metric is divided into the prompt-level accuracy and instruction-level accuracy. Due to the
1101 randomness of generation, each response is tested under multiple transformations, thus the metric
1102 is further divided into strict criterion and loose criterion. In practice, we use the prompt-level and
1103 instruction-level accuracy averaged on the strict and loose criteria.

1104 For other capabilities, we consider reasoning, coding, and multilingual capabilities. For reasoning,
1105 we use GPQA [49] with accuracy as the metric. For coding, we use HumanEval [7] with pass@1
1106 as the metric. For multilingual capability, we use MGSM [55] with the accuracy averaged on all
1107 languages.

1108 To check if the general capabilities as well as other capabilities are affected after fine-tuning, we
1109 compute the relative change in performance for each capability. For example, if the evaluation results
1110 before and after fine-tuning are x and x' , then the relative change is $\frac{x'-x}{x}$. In practice, we test each
1111 model on each benchmark for 10 times. We find that the relative performance change is varies little
1112 among different settings, which further shows the robustness of circuit-tuning. Therefore the final
1113 relative change in performance of each capability is averaged over different sizes of models and
1114 fine-tuning tasks, as discussed in Section 5.2, and the standard errors are also the averaged results.
1115 The final results are shown in Figure 6. It can be seen that circuit-tuning preserves general capabilities
1116 better than others, demonstrating its ability to locate the key components precisely for a specific task.

1117 G.4 Influence of the Regularization Term on Gender De-biasing

1118 We selectively add $\beta \cdot |\mathcal{L}_m|$ as a regularization term in the loss \mathcal{L} for gender de-biasing to improve
1119 performance, as discussed in Appendix G.1.2. We visualize the prejudice risk of the models before
1120 and after gender de-biasing in Figure 5. Note that the coefficient for regularization $\beta = 0.5$. Since
1121 there are only 40 types of occupations in the test set, we regard the distribution of prejudice risk as
1122 normal distribution and perform interpolation on the computed results. For convenience, we only
1123 show the results of GPT2-small and Llama-3.2-1B-instruct in the figure.

1124 The dynamic process of de-biasing can be observed from right to left in Figure 5. It is obvious that
1125 with a regularization term in the loss function, the distribution of the prejudice risk in occupations is
1126 more concentrated to a smaller value. The results demonstrate the effectiveness of our method in
1127 modifying a model’s stereotype, providing new insights in aligning AI systems with human values.

1128 Besides, the experiment result tells us that we can customize the algorithm settings (e.g., the loss
1129 function) according to the requirement of a task. Thus, it is convenient to intervene in the training
1130 process flexibly to modify the model behaviors, demonstrating the great potential of circuit-tuning in
1131 the study of fine-tuning and interpretability.

1132 NeurIPS Paper Checklist

1133 1. Claims

1134 Question: Do the main claims made in the abstract and introduction accurately reflect the
1135 paper's contributions and scope?

1136 Answer: [\[Yes\]](#)

1137 Justification: The main claims in our paper are: (i) We view the learning process inside
1138 a model as finding and optimizing the minimum subgraph corresponding to the intrinsic
1139 dimensions in the computational graph. We redefine the intrinsic dimensionality and estimate
1140 it in the representation space. This can be found at Section 3. (ii) We propose circuit-tuning,
1141 an interpretable fine-tuning method in Section 4. (iii) We conduct experiments based on our
1142 method and analyze the mechanism behind learning in Section 5.

1143 Guidelines:

- 1144 • The answer NA means that the abstract and introduction do not include the claims
1145 made in the paper.
- 1146 • The abstract and/or introduction should clearly state the claims made, including the
1147 contributions made in the paper and important assumptions and limitations. A No or
1148 NA answer to this question will not be perceived well by the reviewers.
- 1149 • The claims made should match theoretical and experimental results, and reflect how
1150 much the results can be expected to generalize to other settings.
- 1151 • It is fine to include aspirational goals as motivation as long as it is clear that these goals
1152 are not attained by the paper.

1153 2. Limitations

1154 Question: Does the paper discuss the limitations of the work performed by the authors?

1155 Answer: [\[Yes\]](#)

1156 Justification: The limitations of our work are detailed in Appendix A, due to limited space.

1157 Guidelines:

- 1158 • The answer NA means that the paper has no limitation while the answer No means that
1159 the paper has limitations, but those are not discussed in the paper.
- 1160 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 1161 • The paper should point out any strong assumptions and how robust the results are to
1162 violations of these assumptions (e.g., independence assumptions, noiseless settings,
1163 model well-specification, asymptotic approximations only holding locally). The authors
1164 should reflect on how these assumptions might be violated in practice and what the
1165 implications would be.
- 1166 • The authors should reflect on the scope of the claims made, e.g., if the approach was
1167 only tested on a few datasets or with a few runs. In general, empirical results often
1168 depend on implicit assumptions, which should be articulated.
- 1169 • The authors should reflect on the factors that influence the performance of the approach.
1170 For example, a facial recognition algorithm may perform poorly when image resolution
1171 is low or images are taken in low lighting. Or a speech-to-text system might not be
1172 used reliably to provide closed captions for online lectures because it fails to handle
1173 technical jargon.
- 1174 • The authors should discuss the computational efficiency of the proposed algorithms
1175 and how they scale with dataset size.
- 1176 • If applicable, the authors should discuss possible limitations of their approach to
1177 address problems of privacy and fairness.
- 1178 • While the authors might fear that complete honesty about limitations might be used by
1179 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
1180 limitations that aren't acknowledged in the paper. The authors should use their best
1181 judgment and recognize that individual actions in favor of transparency play an important
1182 role in developing norms that preserve the integrity of the community. Reviewers
1183 will be specifically instructed to not penalize honesty concerning limitations.

1184 3. Theory Assumptions and Proofs

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

Answer: [Yes]

Justification: The proof of the claim in Section 3 as well as the relevant assumptions and definitions of the introduced concepts are detailed in Appendix E.

Guidelines:

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

4. Experimental Result Reproducibility

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

Answer: [Yes]

Justification: The information needed to reproduce our work is detailed in Section 4, Section 5, Appendix F and Appendix G.

Guidelines:

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

5. Open access to data and code

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

Answer: [Yes]

Justification: We provide our data and code as well as the instructions for training and evaluation in the supplementary details. If accepted, we will release them.

Guidelines:

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

6. Experimental Setting/Details

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

Answer: [Yes]

Justification: All details about our experiments are provided in Appendix F and Appendix G.

Guidelines:

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

7. Experiment Statistical Significance

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

Answer: [Yes]

Justification: Statistical significance of the evaluation results are shown and detailed in Figure 6 and Appedndix G.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. Experiments Compute Resources

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

Answer: [Yes]

Justification: Information on the computer resources are provided in Appendix F and Appendix G.

Guidelines:

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

9. Code Of Ethics

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

Answer: [Yes]

Justification: We conform with the NeurIPS Code of Ethics in every respect.

Guidelines:

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

10. Broader Impacts

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

Answer: [Yes]

Justification: We provide discussions about the social impact of our work in Appendix B.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

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

11. Safeguards

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

Answer: [NA]

Justification: There is no such risk in our work.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: We refer to the contributors of the code and data used in our experiments both in our paper and in our code repository. We will ensure the licenses mentioned and properly respected in our final released code.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

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

Answer: [\[Yes\]](#)

Justification: The details about our dataset for subject-verb disagreement and the instructions of the training framework for our algorithm are provided along with our code.

Guidelines:

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

14. Crowdsourcing and Research with Human Subjects

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

Answer: [\[NA\]](#)

Justification: We do not involve crowdsourcing nor research with human subjects.

Guidelines:

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

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

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

Answer: [\[NA\]](#)

Justification: We do not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- 1444 • Depending on the country in which research is conducted, IRB approval (or equivalent)
1445 may be required for any human subjects research. If you obtained IRB approval, you
1446 should clearly state this in the paper.
- 1447 • We recognize that the procedures for this may vary significantly between institutions
1448 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
1449 guidelines for their institution.
- 1450 • For initial submissions, do not include any information that would break anonymity (if
1451 applicable), such as the institution conducting the review.

1452 16. **Declaration of LLM usage**

1453 Question: Does the paper describe the usage of LLMs if it is an important, original, or
1454 non-standard component of the core methods in this research? Note that if the LLM is used
1455 only for writing, editing, or formatting purposes and does not impact the core methodology,
1456 scientific rigorousness, or originality of the research, declaration is not required.

1457 Answer: [NA]

1458 Justification: There is no usage of LLMs in the development of our method.

1459 Guidelines:

- 1460 • The answer NA means that the core method development in this research does not
1461 involve LLMs as any important, original, or non-standard components.
- 1462 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1463 for what should or should not be described.