# Beyond Static Interpretability: Anticipating Post-SFT Mechanisms from Pre-SFT Parameters for Better Tuning

**Hang Chen**
College of Computing and Data Science
Nanyang Technological University
chen.hang@ntu.edu.sg

**Jiaying Zhu**
School of Computer Science and Engineering
The Chinese University of Hong Kong
jyzhu24@cse.cuhk.edu.hk

**Wenya Wang***
College of Computing and Data Science
Nanyang Technological University
wangwy@ntu.edu.sg

## Abstract

Mechanistic Localization bridges mechanistic interpretability and post-training optimization by isolating critical parameters via interpretative approaches and then guiding parameter-efficient Supervised Fine-Tuning (SFT) in a "locating-then-tuning" paradigm. However, due to the retrospective nature of mechanistic interpretability, directly interpreting pre-SFT models introduces misleading conclusions. Specifically for novel tasks, initially identified neurons differ drastically from those governing the final model, introducing biases that actively disrupt SFT. To address this, we propose a forward-looking localization framework that accurately estimates the post-SFT interpretability state using only pre-SFT parameters and the target dataset. Theoretically, we model SFT as a continuous parameter evolution, leveraging Taylor expansion to rigorously bridge the post-tuning mechanistic objective with the pre-SFT model's dynamic gradients. Practically, we design dual-granularity (neuron- and component-level) localization pipelines. Extensive experiments demonstrate that our approach not only provides superior SFT guidance but also exhibits robust performance and temporal scalability across increasing model sizes. This work transcends the fundamental limitation of traditional interpretability—its inability to identify task-critical mechanisms before they are trained—pioneering a predictive frontier that unites mechanistic interpretability with targeted optimization. Code is available at: https://github.com/Zodiark-ch/Future_localization.

## 1 Introduction

As a pivotal foundation for reverse-engineering LLMs, mechanistic interpretability illuminates internal execution mechanisms by evaluating the causal effects of specific components or neurons through counterfactual analysis [13, 27, 28, 29]. Beyond passive diagnostic insights, its rigorous causal framework offers a principled basis for guiding post-training optimization. In particular, the *locating-then-tuning* paradigm [4] successfully operationalizes this insight into SFT-based adaptation: it first employs *mechanistic localization* [11] on the **pre-SFT** base model to identify key parameters governing the target task, and subsequently restricts parameter-efficient tuning only to these identified

*Corresponding author

Preprint. Under review.

key parameters to yield the final **post-SFT** model. By preventing unconstrained weight updates from corrupting pre-existing representations, this paradigm has exhibited immense promise in critical domains including LLM unlearning [17, 5], knowledge editing [25, 41], and transparent post-training [36, 39, 19, 40].

However, relying on the pre-SFT model for mechanistic localization introduces a fundamental flaw: because the model has not yet been adapted to the target task, the parameters identified as task-relevant in pre-SFT model may not reflect the mechanisms that emerge during SFT, potentially leading to misleading localization for subsequent fine-tuning [4]. Consider an extreme yet representative case where the target task is entirely novel to the pre-SFT model. Because the pre-SFT model inherently lacks the mechanistics required to execute this unseen task, mechanistic interpretability on the pre-SFT model fails to identify meaningful task-governing parameters. Consequently, it yields deceptive signals—inadvertently freezing parameters vital for acquiring new skills and obstructing necessary circuit evolution [26, 16, 7]. Fundamentally, this limitation stems from an intrinsic tension between the prospective requirement of tuning and the retrospective nature of interpretability: locating-then-tuning demands that interpretability *anticipate* which parameters will be recruited to learn the target skill, yet mechanistic interpretability can only dissect capabilities that are *already instantiated* within existing parameters. This creates a classic "chicken-and-egg" dilemma—one must localize the task-critical parameters before tuning, yet the corresponding circuits only emerge after the model has learned the task.

To this end, this paper aims to challenge the retrospective practice of interpretability area by answering a fundamental question: *how can we predict the neurons (or components) ($\mathcal{N}(\theta^I)$) responsible for a target task in an ideal post-SFT model($\theta^I$) using only the pre-SFT model parameters ($\theta$)?*

We view fine-tuning ($\theta \rightarrow \theta^I$) as an optimization establishing an initial gradient direction and subsequently traversing a long distance toward a local optimum. Consequently, we decompose $\theta \rightarrow \mathcal{N}(\theta^I)$ into **two theoretical steps**: isolating "**direction**" and extrapolating "**distance**". First, we formulate $\theta \rightarrow \mathcal{N}(\theta')$, where $\theta'$ represents the parameters following an infinitesimal update toward the target task's supervision. This decouples the update distance to exclusively capture the initial direction. Building upon $\mathcal{N}(\theta')$, the second step achieves $\mathcal{N}(\theta') \rightarrow \mathcal{N}(\theta^I)$ by conceptualizing the total SFT parameter shift, $\Delta\theta = \theta^I - \theta$, as the accumulation of infinite recursive updates originating from $\theta'$.

Our approach only requires fine-tuning on $1\%$ of the training dataset for a single epoch to derive $\theta'$. Inheriting the computational paradigm of Attribution Patching [24, 32], our method necessitates only two forward passes and one backward pass to collect the activation and gradient data required for the estimation. This computational overhead is entirely negligible compared to the full fine-tuning process or other interpreting approaches. Furthermore, to accommodate larger models and diverse practical scenarios, we designed two distinct estimation pipelines: a fine-grained estimation at the neuron level and a rapid estimation at the component level.

We empirically validate our approach on the Mistral-7B model, demonstrating its superior capability in guiding SFT compared to existing localization baselines. Furthermore, through rigorous ablation studies on arithmetic tasks of varying complexities, we establish the robustness of our method against increasing task difficulty. Additionally, we confirm the scalability of our approach across larger architectures, specifically LlaMA-2-13B and Qwen3-30B. Finally, we provide intrinsic validation of our method's efficacy directly through the lens of the resulting interpretability analyses.

In summary, our main contributions are twofold:

- We propose a theoretical framework that leverages current parameters to estimate post-SFT mechanisms. To the best of our knowledge, this is the first work to predict future interpretability states. Confronting this question marks a transformative transition beyond traditional retrospective interpretability, pioneering a new frontier for its actionable utility in post-training optimization.

- We develop an estimation pipeline that achieves state-of-the-art localization performance. Validated across diverse LLM families, our method demonstrates robust scalability while maintaining highly stable computational efficiency as model parameters increase.

## 2 Preliminaries

### 2.1 Task Definition

Let $\mathcal{M}$ parameterized by $\theta$ denote the current model prior to fine-tuning, and $\mathcal{M}^I$ parameterized by $\theta^I$ represent the ideal post-tuning model. For a target task $\mathcal{T}$, let $\mathcal{D}_\mathcal{T}$ be its associated dataset. By definition, the current model $\mathcal{M}$ performs poorly on $\mathcal{D}_\mathcal{T}$, whereas the ideal model $\mathcal{M}^I$ achieves optimal performance on $\mathcal{D}_\mathcal{T}$ without degrading any pre-existing capabilities[2].

We denote the subset of parameters primarily responsible for task $\mathcal{T}$ as $\mathcal{N}^\theta \subset \theta$, which is typically identified via mechanistic interpretability methods. The existing *locating-then-tuning* paradigm operates through the following pipeline with partial tuning only operating on $\mathcal{N}^\theta$:

$$\mathcal{M}(\theta), \mathcal{D}_\mathcal{T} \xrightarrow{\text{localization}} \mathcal{N}^\theta \xrightarrow[\mathcal{D}_\mathcal{T}]{\text{partial tuning}} \hat{\mathcal{M}}^I(\theta^I)$$

In this work, our objective is to reformulate the localization step to directly predict the critical parameter subset ($\mathcal{N}^{\theta^I}$) of the ideal model:

$$f : \langle \mathcal{M}(\theta), \mathcal{D}_\mathcal{T} \rangle \longrightarrow \mathcal{N}^{\theta^I}$$

### 2.2 Related Work

Existing research on the *locating-then-tuning* paradigm primarily focuses on optimizing mechanistic localization, relying predominantly on gradients [17, 39, 25] or causal effects [36, 19, 11, 5]. Recent advances in mechanistic interpretability [13, 21, 6, 3] demonstrate that, under rigorous experimental setups, causal effect-based approaches provide a more accurate and comprehensive identification of task-critical neurons. Specifically, these methods quantify the causal effect for the activation of a target parameter, neuron, or component $a$ under the target dataset $\mathcal{D}_\mathcal{T}$, by measuring the output discrepancy when the activation is replaced with a counterfactual value. This is formalized as:

$$\Delta E_{\mathcal{D}_\mathcal{T}}(a) = f_a(A(x_c)) - f_a(A(x_r))$$

where $f_a(\cdot)$ represents an output-dependent metric (e.g., logits or loss) as the samples from $\mathcal{D}_\mathcal{T}$ are input, while $A(x_c)$ and $A(x_r)$ denote the clean and corrupted (counterfactual) activations, respectively. A larger $\Delta E$ signifies a stronger causal relationship with the task output. Currently, circuit discovery [8, 2] and path patching [35] are the most prevalent causal methodologies employed in mechanistic localization.

Highly pertinent to our study is *Attribution Patching* [24, 32], a causal effect method that approximates this influence using a first-order Taylor expansion:

$$\Delta E_{\mathcal{D}_\mathcal{T}}(a) = f_a(A(x_c)) - f_a(A(x_r)) \approx (A(x_c) - A(x_r)) \cdot \left.\frac{\partial f_a}{\partial a}\right|_{a=A(x_r)}$$

This mathematical reformulation elegantly circumvents the computationally prohibitive exhaustive search required by standard counterfactual ablation. Instead, it extracts the necessary metrics for all parameters simultaneously requiring only two forward passes to obtain $A(x_c)$ and $A(x_r)$, and a single backward pass to compute the gradients $\left.\frac{\partial f_a}{\partial a}\right|_{a=A(x_r)}$.

Beyond static analysis, our work is motivated by recent dynamic interpretability studies analyzing mechanistic shifts during SFT [26, 16]. These works highlight that target task mechanisms undergo drastic transformations—including shifts in circuit strength [7], topology [36], and key nodes [4]—during tuning. As a result, deriving the critical neuron set $\mathcal{N}^\theta$ purely from the pre-SFT parameters $\theta$ yields a biased localization. Instead of aiding the tuning process, this static guidance misdirects optimization and deteriorates task learning, thereby motivating our framework to estimate the ideal post-SFT state.

[2] In practice, SFT can only continuously approximate this ideal model rather than achieving it perfectly.

## 3 Method

The inherent challenge in our objective lies in the pragmatic nature of mechanistic interpretability: it is traditionally a retrospective analysis conducted on the explicit parameters of already-trained models. Existing theories do not support predicting interpretability states prior to tuning. Whereas targeted SFT aims at updating parameters supposedly influential for the target task. This presents a critical challenge: *given only the current parameters $\theta$ and the target dataset $\mathcal{D}_\mathcal{T}$, how can we estimate the importance of future model components without actually executing the entire SFT process?*

To address this, we conceptualize SFT as a continuous optimization trajectory—starting from an initial well-trained parameter distribution, evolving along a specific gradient direction, and traversing a certain distance to reach a local optimum. By bridging interpretability with gradient dynamics, our framework solves this challenge in two intuitive steps: first determining the initial evolutionary direction, and then extrapolating it to an appropriate distance. Specifically, first, in Section 3.1, we capture the initial gradient direction by modeling an infinitesimal update of $\theta$ towards the SFT objective on $\mathcal{D}_\mathcal{T}$, denoted as $\theta'$, and then we formulate its causal effect entirely using the initial $\theta$. Second, in Section 3.2, we extrapolate this directional estimate to the optimal distance. By iteratively accumulating an infinite series of these infinitesimal updates, we approximate the causal effect of the ideal post-SFT model, $\theta^I$, deriving our final predictive expression.

### 3.1 Single Step: From $\theta$ to $\theta'$ for Gradient Direction

Let $\theta'$ denote the parameters following an infinitesimal update of $\theta$ directed by the SFT objective on $\mathcal{D}_\mathcal{T}$. For any arbitrary parameter $a$ under $\theta'$, its corresponding Attribution Patching formula is given by:

$$f_a^{\theta'}(A(x_c)) - f_a^{\theta'}(A(x_r)) \approx (A^{\theta'}(x_c) - A^{\theta'}(x_r)) \cdot \frac{\partial f_a^{\theta'}}{\partial a^{\theta'}}\bigg|_{a=A^{\theta'}(x_r)} \tag{1}$$

Evaluating the terms $A^{\theta'}(x_r)$, $A^{\theta'}(x_c)$, and the gradient $\frac{\partial f_a^{\theta'}}{\partial a^{\theta'}}\big|_{a=A^{\theta'}(x_r)}$ inherently requires two actual forward passes and one backward pass through the hypothetical $\theta'$ model, which is practically intractable. Consequently, we seek to approximate these terms by applying a first-order Taylor expansion around $\theta$. This yields a reformulation of the causal effect for $\theta'$ expressed entirely in terms of $\theta$:

$$f_a^{\theta'}(A(x_c)) - f_a^{\theta'}(A(x_r)) \approx f_a^{\theta}(A(x_c)) - f_a^{\theta}(A(x_r)) + \Delta\theta \cdot S(\theta) \tag{2}$$

where

$$S(\theta) = \left[\left(\frac{\partial A^\theta(x_c)}{\partial \theta} - \frac{\partial A^\theta(x_r)}{\partial \theta}\right) \cdot \frac{\partial f_a^\theta}{\partial a^\theta} + \left(A^\theta(x_c) - A^\theta(x_r)\right) \cdot \frac{\partial^2 f_a^\theta}{\partial a \partial \theta}\right] \tag{3}$$

Here, $\Delta\theta = \theta' - \theta$ represents the distance of the parameter update. Noticeably, $S(\theta) \equiv \nabla_\theta \left[(A^\theta(x_c) - A^\theta(x_r)) \cdot \frac{\partial f_a^\theta}{\partial a^\theta}\right]$. Therefore, it represents the gradient of the attribution score with respect to $\theta$, capturing the sensitivity of the target neuron $a$'s attribution value to parameter updates.

Through Equation 2 (the detailed derivation is provided in Appendix A), we transform the estimation of $\theta'$'s causal effect—which previously relied on empirical forward and backward passes—into an estimation of the parameter update distance. In the following sections, we will elaborate on how to determine $\Delta\theta$ in order to ultimately approximate the interpretability state of the final ideal model, $\theta^I$.

### 3.2 Multiple Steps: From $\theta'$ to $\theta^I$ for Gradient Distance

We conceptualize the transition from the initial parameters $\theta$ to the ideal parameters $\theta^I$ as an infinite sequence of continuous, additive updates. For notational clarity, we represent this evolutionary trajectory as a series of states $\theta^0, \theta^1, \ldots, \theta^I$, where the initial state is defined as $\theta^0 \equiv \theta$. Within this continuum, any intermediate state $\theta^k$ effectively functions as the infinitesimal update, previously denoted as $\theta'$, relative to its immediate predecessor $\theta^{k-1}$.

Following Equation 2, the causal effect $\Delta E_{\mathcal{D}_\mathcal{T}}^{\theta^{k+1}}(a)$ at any subsequent state $\theta^{k+1}$ can be recursively formulated as:

$$\Delta E_{\mathcal{D}_\mathcal{T}}^{\theta^{k+1}}(a) \approx \Delta E_{\mathcal{D}_\mathcal{T}}^{\theta^{k}}(a) + \Delta\theta^k \cdot S(\theta^k) \tag{4}$$

Consequently, by accumulating these incremental updates, the causal effect for the final ideal model is given by:

$$\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a) \approx \Delta E^{\theta}_{\mathcal{D}_\mathcal{T}}(a) + \sum_{k=0}^{I-1} \Delta\theta^k \cdot S(\theta^k) \tag{5}$$

The summation term in this formulation effectively constitutes a discrete path integral along the parameter update trajectory $\gamma$. In the continuous limit, this can be formally expressed as $\int_\gamma S(\theta) \cdot d\theta$. To render this computation tractable without evaluating intermediate states, we approximate this continuous integral using the trapezoidal rule[3], yielding:

$$\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a) \approx \Delta E^{\theta}_{\mathcal{D}_\mathcal{T}}(a) + \frac{1}{2}\left[S(\theta^0) + S(\hat{\theta}^I)\right] \cdot \sum_{k=0}^{I-1} \Delta\theta^k \tag{6}$$

Given that $S(\theta)$ is strictly equivalent to the gradient of the attribution score with respect to $\theta$ (i.e., $S(\theta) \equiv \nabla_\theta[\Delta E^{\theta}_{\mathcal{D}_\mathcal{T}}(a)]$). By defining a differential operator $D = \Delta\theta \cdot \nabla_\theta$, we demonstrate that the infinite series of higher-order derivatives (Hessians, tensors) governing $S(\theta^k)$ can be rigorously folded into an exponential translation operator, $e^{k\Delta\theta\cdot\nabla_\theta}$. Under this functional analysis perspective, evaluating $S$ at state $k$ is mathematically equivalent to simply shifting the input parameters. Consequently, we establish that $S(\hat{\theta}^I) = S(\theta + K\Delta\theta)$, allowing us to compute the final sensitivity by merely scaling the update distance. We refer readers to Appendix B for the complete theoretical proof.

Ultimately, we simplify the estimation of the causal effect for the ideal model, $\theta^I$, to the following formulation:

$$\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a) \approx \Delta E^{\theta}_{\mathcal{D}_\mathcal{T}}(a) + \frac{1}{2}\left[S(\theta^0) + S(\theta^0 + K \cdot \Delta\theta)\right] \cdot K \cdot \Delta\theta \tag{7}$$

where $K$ denotes a step scalar that extrapolates the transition $\theta \to \theta'$ to the ideal macroscopic transition $\theta \to \theta^I$. While theoretically derived from an infinitely small step, practically, we approximate $\theta'$ using a sufficiently small parameter update. Moreover, the ideal post-SFT state is non-unique—varying based on optimization trajectories or random initializations—we do not treat $K$ as a single determined value. Instead, we use $K$ as an extrapolation scalar to project along the known gradient direction, sampling a range of $K$ values to explore plausible future configurations $S(\hat{\theta}^I) = S(\theta^0 + K \cdot \Delta\theta)$ and ultimately estimating the mean causal effect of the target task mechanisms. Therefore, this approximation requires that $\Delta\theta = \theta' - \theta$ robustly captures the general optimization direction of the target task, while maintaining $\Delta\theta$ small enough to provide the step scalar $K$ with ample degrees of freedom to explore the parameter space.

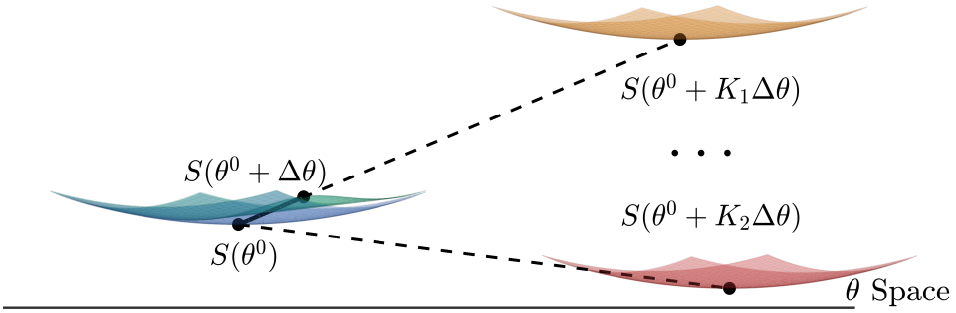



Figure 1: **Extrapolation of the ideal state via the step scalar $K$.** Different values of $K$ yield distinct sensitivity estimations $S(\theta + K \cdot \Delta\theta)$, reflecting that the ideal parameter state $\theta^I$ consists of an infinite set of valid configurations rather than a single deterministic solution.

[3]The underlying assumption of this trapezoidal approximation is mild: as the gradient of the attribution score, $S(\theta)$ typically exhibits local monotonicity during tuning. Consequently, the introduced systematic and unidirectional error across parameters barely affects the relative ranking of each neuron's final causal effect.

# 4 Implemetation

## 4.1 Probing SFT

In this work, we acquire $\theta'$ through a *probing SFT* strategy. Specifically, we uniformly sample $1\%$ of the training dataset to perform a tentative fine-tuning for a single epoch at $1\%$ of the original learning rate, designating the resulting updated weights as $\theta'$. As explored in our empirical analysis (Section 5.4), a larger divergence between $\theta'$ and $\theta$ does not monotonously correlate with improved performance; in fact, deriving $\theta'$ from $1\%$ of the training data yields significantly superior results compared to using $10\%$. This phenomenon occurs because $\Delta\theta$ must remain sufficiently infinitesimal to preserve adequate degrees of freedom for the extrapolated state $\theta + K \cdot \Delta\theta$, while simultaneously mitigating the influence of gradient-irrelevant noise inherent in $\theta'$.

Notably, performing attribution patching directly on $\theta'$ (hereafter referred to as the *probing model*) constitutes a compelling baseline, which we denote as *Localization from* $\theta'$. It intuitively offers more foresight than localization from the initial $\theta$, yet avoids the complexity of estimating the ideal state $\hat{\theta}^I$. Mathematically, however, $\theta'$ remains proximate to $\theta$ and distant from $\hat{\theta}^I$. As formally proven in Appendix C, establishing the equivalence $\Delta E^{\hat{\theta}^I}_{\mathcal{D}_\mathcal{T}}(a) \approx \Delta E^{\theta'}_{\mathcal{D}_\mathcal{T}}(a)$ necessitates imposing stringently unrealistic assumptions on the SFT parameter update trajectory. That is, for any step $k$, it satisfies: $\frac{\partial(A(x_c)-A(x_r))}{\partial\theta} \cdot \Delta\theta \approx X_k - X_0$. Nevertheless, regarding the final estimation, $\Delta E^{\theta'}_{\mathcal{D}_\mathcal{T}}(a)$ inherently introduces significantly less approximation error, at least from the perspective of parameter variation. Our empirical evaluation in Section 5.3 thoroughly compares these two paradigms—*Localization from* $\hat{\theta}^I$ versus *Localization from* $\theta'$. The results demonstrate that as the difficulty of the target SFT task increases, the superiority of *Localization from* $\hat{\theta}^I$ over *Localization from* $\theta'$ becomes increasingly pronounced.

## 4.2 Locating-then-tuning Pipelines

We present a generalized pipeline for the *locating-then-tuning* paradigm. During the localization phase, we compute the causal effect $\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a)$ for all parameters and rank them in descending order to determine their relative importance. To ensure a robust estimation, we randomly sample 10 distinct values of $K$ and average their resulting scores. The optimal sampling interval for $K$ is inherently task-dependent, a phenomenon we thoroughly analyze in Appendix F.1, and it explores a probing model approach to recommend $K$, and provides an analysis of the robustness concerning $K$ value sampling. Furthermore, to circumvent the prohibitive computational overhead induced by the second-order partial derivatives in $S(\theta)$, we employ a highly time-efficient approximation strategy (We show the acceptable error in Figure 10):

$$\frac{\partial^2 f_a}{\partial a \partial \theta} \cdot \Delta\theta \approx \frac{\partial f_a^{\theta'}}{\partial a^{\theta'}} - \frac{\partial f_a^{\theta}}{\partial a^{\theta}} \tag{8}$$

During the tuning phase, we implement differential optimization strategies tailored to the derived parameter importance. Specifically, we instantiate this pipeline at two distinct granularities:

**(1) Neuron-level:** The minimal unit for computing $\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a)$ is defined as an individual neuron (i.e., a single scalar value within a parameter matrix). We selectively update only the top 20% most critical neurons while freezing the remaining 80%, optimizing via the standard next-token cross-entropy loss for SFT. For example, when applied to the Mistral-7B model, this strategy restricts parameter updates to exactly 1.4B out of the 7B total neurons.

**(2) Component-level:** The minimal unit is designated as an entire parameter matrix (specifically, $W_q, W_k, W_v, W_o$ within each attention head, and $W_{up}, W_{gate}, W_{down}$ in the MLP). We integrate Low-Rank Adaptation (LoRA) [14] into this pipeline by dynamically allocating ranks according to the component importance rankings. Higher-ranked components are assigned proportionally higher LoRA ranks, scaling within a discrete range of $[1, 32]$ while maintaining an overall average rank of $8$.

# 5 Experiments

## 5.1 Setups

We employ Mistral-7B [18] as our primary evaluation model, while further incorporating LLaMA-2-13B [33] and Qwen3-30B [37] to validate the scalability of our approach. We evaluate LLM fine-tuning performance across three primary domains: **Natural Language Understanding (NLU)** across the GLUE benchmark (SST-2, MRPC, QQP, MNLI, and RTE) [34], **Logical Reasoning (LR)** via BOOL [31], and **Mathematical Reasoning (MR)** using Arithmetic [30][4]. To examine localization efficacy under varying adaptation complexities, we stratify the Arithmetic benchmark into hierarchical subtasks ranging from 2-digit to 7-digit operations and 1-step to 6-step computations. Additionally, to facilitate a granular examination of interpretability outcomes—specifically through circuit graph and component statistics analyses—we incorporate established benchmarks from the mechanistic interpretability literature: Indirect Object Identification (IOI) [35], Induction [8], WinoGrande [1], Genderr [22], and Docstring [12]. Detailed prompt configurations for these datasets are provided in Appendix D.

We evaluate our proposed localization methods—denoted as Ours (N) for the neuron-level pipeline and Ours (C) for the component-level pipeline—against a suite of state-of-the-art baselines. These comprise gradient-guided approaches (Graft [25] and WAGLE [17]) alongside causal effect-guided methods (CLUE [5], FLU [11], and CircuitLoRA [36]). To comprehensively assess post-tuning efficacy, we monitor two primary metrics: Target Task Accuracy (TTA) and Pervasiveness Task Accuracy (PTA). TTA is directly derived from the label accuracy on the target tasks (e.g., average accuracy from all tasks in GLUE, BOOL, and Arithmetic). Conversely, PTA quantifies the degradation of pre-existing general capabilities; it is computed as the macro-average across a diverse array of generalized benchmarks using the `lm-evaluation-harness` framework [10].

For the fine-tuning optimization, we minimize the cross-entropy loss over the target task labels. The training hyperparameters are standardized to a learning rate of $1 \times 10^{-5}$, 3 training epochs, and a default LoRA rank of 8. All experiments are executed on NVIDIA RTX A6000 with corresponding CUDA environments.

## 5.2 Main Results

| Method | LR | | NLU | | MR | |
|---|---|---|---|---|---|---|
| | **TTA** | **PTA** | **TTA** | **PTA** | **TTA** | **PTA** |
| Graft | $82.76_{\pm 0.42}$ | $33.27_{\pm 0.04}$ | $84.96_{\pm 0.08}$ | $31.24_{\pm 0.14}$ | $24.19_{\pm 2.13}$ | $44.31_{\pm 0.11}$ |
| FLU | $88.57_{\pm 0.35}$ | $42.57_{\pm 0.09}$ | $84.27_{\pm 0.02}$ | $36.72_{\pm 0.13}$ | $87.16_{\pm 1.07}$ | $49.52_{\pm 0.01}$ |
| WAGLE | $89.55_{\pm 0.28}$ | $44.31_{\pm 0.06}$ | $85.28_{\pm 0.11}$ | $43.57_{\pm 0.04}$ | $66.39_{\pm 1.15}$ | $51.23_{\pm 0.09}$ |
| CLUE | $90.49_{\pm 0.44}$ | $45.42_{\pm 0.02}$ | $85.17_{\pm 0.07}$ | $45.61_{\pm 0.10}$ | $52.78_{\pm 1.55}$ | $53.21_{\pm 0.12}$ |
| CircuitLoRA | $97.67_{\pm 0.13}$ | $60.62_{\pm 0.11}$ | $85.09_{\pm 0.06}$ | $61.55_{\pm 0.09}$ | $98.60_{\pm 0.12}$ | $59.66_{\pm 0.14}$ |
| **Ours (N)** | $98.51_{\pm 0.09}$ | $60.53_{\pm 0.05}$ | $86.31_{\pm 0.13}$ | $62.02_{\pm 0.07}$ | $98.79_{\pm 0.14}$ | $60.13_{\pm 0.10}$ |
| **Ours (C)** | $\mathbf{100.00}_{\pm 0.00}$ | $\mathbf{62.64}_{\pm 0.08}$ | $\mathbf{87.55}_{\pm 0.04}$ | $\mathbf{63.95}_{\pm 0.11}$ | $\mathbf{100.00}_{\pm 0.00}$ | $\mathbf{61.48}_{\pm 0.06}$ |

Table 1: Performance comparison of Locating-then-tuning methods.

Table 1 reports the performance of Mistral-7B model of various localization methods across the LR, NLU, MR domains, where the Arithmetic results represent the average performance across the 2- to 5-digit subtasks. As demonstrated, our proposed methods establish a substantial margin over all baseline approaches in both Target Task Accuracy (TTA) and the retention of general capabilities (PTA). This confirms that estimating future parameter importance provides a distinct advantage over conventional mechanistic localization derived solely from current parameters. Interestingly, our component-level pipeline (Ours (C)) significantly outperforms its neuron-level counterpart (Ours (N)), suggesting that an increasingly fine granularity in mechanistic localization does not strictly

[4]Unlike mainstream instruction-following tasks in post-training SFT, the locating-then-tuning paradigm typically employs tasks with highly controlled output lengths (e.g., single-token generation). This constraint is intentional: it yields cleaner causal mechanisms and isolates the exact effects of mechanistic localization, allowing us to explicitly ablate the contributions of mechanistic localization. We discuss this limitation further in Section 6.

correlate with improved performance. Furthermore, CircuitLoRA—another baseline utilizing LoRA for fine-tuning—emerges as the second-best method. This observation implies that LoRA's robust adaptability to single, formalized tasks may largely account for the empirical superiority of the component-level approach over the neuron-level configuration.

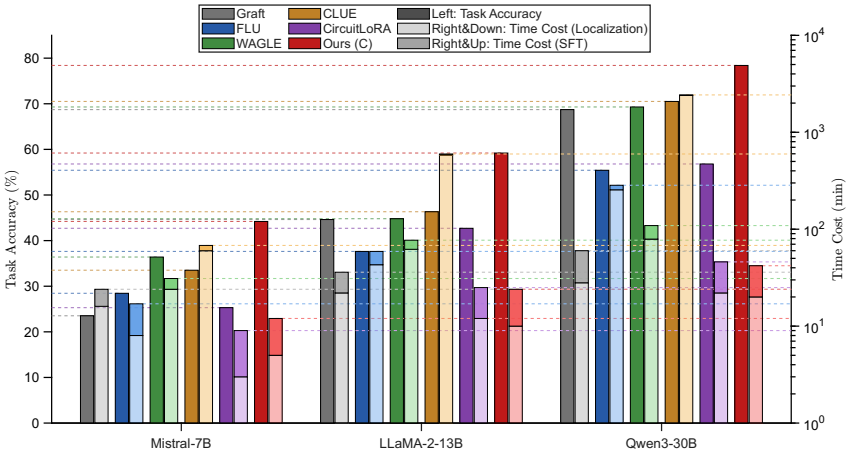



Figure 2: The scalability of performance and time cost with mode size increasing.

Furthermore, to validate the scalability of our proposed methods, we conduct a comparative analysis across three text-to-text models of varying parameter sizes that lack native complex reasoning capabilities: Mistral-7B, LLaMA-2-13B, and Qwen3-30B. We select the 5-step and 6-step Arithmetic subtasks for evaluation, employing the mean Target Task Accuracy (TTA) as the primary performance metric. To quantify computational efficiency, we report the total time expenditure, defined as the sum of the localization and fine-tuning durations.

As illustrated in Figure 2, our method consistently sustains optimal fine-tuning performance as the model scale increases, demonstrating its robust efficacy on LLMs. Concurrently, the fine-tuning time cost exhibits a relatively gradual growth rate, though we acknowledge this may be partially attributed to the inherent simplicity of the chosen target tasks. Conversely, the localization time cost escalates prominently with model size. This sharp increase confirms that the computational complexity of mechanistic localization strictly exceeds $\mathcal{O}(n)$, a consequence of the auxiliary spatial and computational overhead mandated by interpretability procedures. Despite this inherent overhead, our approach maintains nearly the lowest overall computational complexity among the evaluated baselines, underscoring its practical viability for large-scale LLMs.

Finally, to further explore the broader potential of our approach, we conduct multi-task joint training experiments, as detailed in Appendix E. Existing literature demonstrates that joint training frequently exacerbates inter-task conflicts [5, 38], a phenomenon primarily manifesting within shared neurons that are simultaneously responsible for multiple distinct tasks [9]. By jointly evaluating task accuracy and the prevalence of these conflicting neurons, we empirically demonstrate that our proposed localization method provides crucial foresight. This predictive capability substantially mitigates inter-task interference by proactively avoiding the allocation of shared, conflicting neurons during the fine-tuning process.

### 5.3 Ablation Study

To elucidate the intrinsic contributions of our proposed framework, we compare our method against four distinct localization strategies:

(1) **Full-Param**: Standard full-parameter fine-tuning without any localization, designed to isolate the overall impact of applying a localization mechanism.

(2) **Random**: Random parameter localization that rigorously preserves the exact rank distribution of our method, isolating the effect of merely reducing the trainable parameter count.

(3) **Static**: Localization derived directly from the current model parameters (Localization from $\theta$), serving as a baseline for pre-update parameter importance.

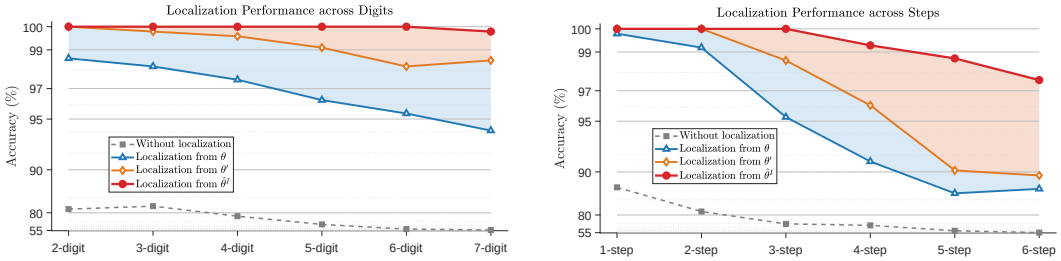
(4) **Probing**: Localization based on the model post-probing SFT (Localization from $\theta'$), designed to contrast the efficacy of the intermediate state $\theta'$ against our estimated ideal state $\hat{\theta}^I$.

As presented in Table 2, the results not only confirm that localization provides critical interpretability-driven guidance, but also conclusively demonstrate that anticipating "future" parameter states offers significantly greater foresight than relying on current parameters. Additionally, Appendix G investigates the robustness of our framework against diverse stochasticity.

Although the Probing strategy performs comparably to our method on standard benchmarks, we conduct further evaluations on tasks of escalating difficulty to explicitly disentangle their capabilities. Specifically, we utilize Arithmetic subtasks ranging from 2 to 7 digits and 1 to 6 reasoning steps. As illustrated in Figure 3, the performance of both the Static and Probing strategies degrades precipitously as task complexity increases, whereas our method preserves substantial robustness. This phenomenon occurs because harder tasks necessitate significantly larger parameter updates during fine-tuning, thereby exacerbating the representational lag inherent in both the initial parameters $\theta$ and the probing state $\theta'$. These findings substantiate that our method is uniquely advantageous for complex tasks where the model struggles to rapidly adapt. Conversely, for rudimentary SFT tasks under extreme computational constraints, the Probing or Static methods may serve as viable, efficient alternatives.

| **Strategy** | **Neuron-Level** | | | **Component-Level** | | |
|---|---|---|---|---|---|---|
| | **LR** | **NLU** | **MR** | **LR** | **NLU** | **MR** |
| Full-Param | 69.27 | 94.03 | 15.66 | 99.73 | 96.33 | 71.83 |
| Random | 88.31 | 95.03 | 43.77 | 99.64 | 96.86 | 76.41 |
| Static | 87.49 | 95.52 | 55.83 | 99.75 | 96.33 | 83.52 |
| Probing | 92.76 | 96.10 | 87.16 | 99.96 | 96.21 | 99.60 |
| **Ours** | **98.51** | **96.31** | **98.79** | **100.00** | **97.55** | **100.00** |

Table 2: Ablation study across datasets at neuron and component levels.




(a) Ablation on different digits.

(b) Ablation on different steps.

Figure 3: Ablation study results on the arithmetic task. (a) Performance variations across different digits. (b) Performance variations across different reasoning steps.

Furthermore, Appendix H provides an in-depth mechanistic interpretability analysis. From a circuit perspective, this analysis corroborates that our estimated parameters yield a mechanistic circuit more closely aligned with that of the fully fine-tuned model, significantly outperforming direct localization from either the initial state $\theta$ or the probing state $\theta'$. Additionally, Appendix I visualizes the respective circuit graphs for all three parameter states: $\theta$, $\theta'$, and $\hat{\theta}^I$. These visualizations explicitly demonstrate that our $K$-value extrapolation successfully identifies critical "intermediate nodes." These intermediate nodes typically encode distinct subskills within the computational pathways, constituting essential components of the underlying task mechanism.

### 5.4 Analysis about Probing SFT

In this section, we investigate the empirical impact of various probing SFT setups. Using the fully fine-tuned model as a surrogate for the ideal model, we adopt the component-level estimation pipeline on the Mistral-7B model for the IOI task as our evaluation baseline. We systematically compare probing models updated on varying proportions of the training dataset: 20%, 10%, 5%, 2%, and 1%, alongside extreme scenarios restricted to merely 2 samples and 1 sample. For each configuration, we run 10 random seeds and sample 50 distinct $K$ values per run within a valid range and compute the Top@50 overlap between the component rankings of our estimated $\hat{\theta}^I$ and those of the surrogate ideal model.

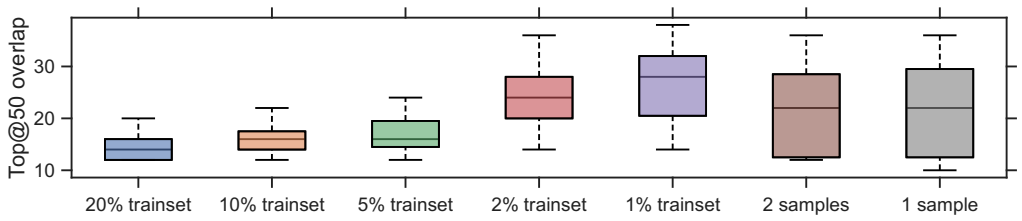



Figure 4: Top 50 overlap with Full-Parameter fine-tuning model across different probing SFT setups.

As illustrated in Figure 4, reducing the sample size used for the probing SFT consistently improves the mean overlap with the ideal model up to a certain threshold. Moreover, we also validate that the learning rate and epoch expose the similar pattern in Appendix J. These empirical trends corroborate our theoretical assertion in Section 4.1: maintaining a sufficiently small parameter update during probing is imperative to preserve the necessary degrees of freedom for the extrapolation term $K \cdot \Delta\theta$. However, in extreme low-data regimes (i.e., 1 or 2 samples), the variance of the estimations escalates sharply, precipitating a noticeable decline in the mean overlap. We attribute this degradation to the inability of such minuscule sample sets to adequately capture the underlying distribution of the target task, which inadvertently injects significantly biased gradient directions into $\Delta\theta$.

## 6 Conclusion and Limitations

This paper presents a breakthrough advancement in mechanistic localization prior to Supervised Fine-Tuning (SFT), enabling the prediction of post-SFT mechanisms to proactively facilitate the tuning process. Theoretically, we bridge mechanistic interpretability with parameter updates via Taylor expansion, breaking the fundamental post-hoc nature of interpretability analysis to usher in a predictive, future-oriented paradigm—thereby transforming interpretability from a retrospective diagnostic tool into an actionable, proactive optimizer. Practically, utilizing only the base model parameters and a probing SFT model trained on a mere $1\%$ of the dataset, we effectively estimate and rank the mechanistic importance of parameters for the anticipated post-SFT state. This enables a targeted, fine-grained allocation of tuning intensity while ensuring scalability in both time and performance. Our work inspires future engineering-oriented interpretability research based on following **limitations**:

- **Mechanistic Interpretability for Multi-Token Generation:** Current interpretability methods predominantly focus on the internal mechanisms during next-token prediction. However, mainstream post-training SFT typically targets instruction-following tasks that necessitate long-sequence generation [20]. This discrepancy currently restricts the locating-then-tuning paradigm to relatively narrow downstream applications—such as LLM unlearning and knowledge editing—rather than broader instruction task practices. Bridging this gap poses a revolutionary challenge: extracting representative mechanisms and their corresponding parameter regions across multiple forward passes. Exploring this frontier will directly extend our insights into critical domains like reinforcement learning and on-policy distillation.
- **Targeted Parameter Updates for Conflicting Mechanisms:** Due to neuron polysemanticity, fine-tuning on a specific task frequently degrades capabilities on others. Existing interpretability and localization efforts merely identify these polysemantic neurons, falling short of providing effective mitigation strategies. This remains a limitation of our study, which renders the advantages of multi-objective joint optimization less pronounced. Addressing this issue will tightly couple mechanistic interpretability with post-training optimization, drastically expanding the practical utility of the interpretability field.

## References

[1] Winogrande: An adversarial winograd schema challenge at scale. 2019.

[2] A. Bhaskar, A. Wettig, D. Friedman, and D. Chen. Finding transformer circuits with edge pruning. *Advances in Neural Information Processing Systems*, 37:18506–18534, 2024.

[3] H. Chen, X. Yang, J. Zhu, and W. Wang. Skill path: Unveiling language skills from circuit graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30210–30217, 2026.

[4] H. Chen, J. Zhu, H. Chen, H. Liu, X. Yang, and W. Wang. Navigating by old maps: The pitfalls of static mechanistic localization in llm post-training, 2026. URL https://arxiv.org/abs/2605.06076.

[5] H. Chen, J. Zhu, X. Yang, and W. Wang. CLUE: Conflict-guided localization for LLM unlearning framework. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jtRYvazBWv.

[6] H. Chen, J. Zhu, X. Yang, and W. Wang. Rethinking circuit completeness in language models: And, or, and adder gates. *Advances in Neural Information Processing Systems*, 38:150511–150540, 2026.

[7] V. K. Chhabra, D. Zhu, and M. M. Khalili. Neuroplasticity and corruption in model mechanisms: A case study of indirect object identification. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3099–3122, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.170. URL https://aclanthology.org/2025.findings-naacl.170/.

[8] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

[9] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

[10] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

[11] P. H. Guo, A. Syed, A. Sheshadri, A. Ewart, and G. K. Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=92oBV5HAGl.

[12] S. Heimersheim and J. Janiak. A circuit for python docstrings in a 4-layer attention-only transformer. In *Alignment Forum*, 2023.

[13] S. Heimersheim and N. Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

[14] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[15] N. Jacobson. *Lie algebras*. Courier Corporation, 2013.

[16] S. Jain, R. Kirk, E. S. Lubana, R. P. Dick, H. Tanaka, T. Rocktäschel, E. Grefenstette, and D. Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=A0HKeKl4Nl.

[17] J. Jia, J. Liu, Y. Zhang, P. Ram, N. Baracaldo, and S. Liu. WAGLE: Strategic weight attribution for effective and modular unlearning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=VzOgnDJMgh.

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

[19] Y. Jing, Z. Dai, J. Hu, Z. Yao, L. Hou, J. Li, and X. Wang. Guiding llm post-training data engineering with model internals from sparse autoencoders, 2026. URL https://arxiv.org/abs/2605.27354.

[20] H. Lai, X. Liu, J. Gao, J. Cheng, Z. Qi, Y. Xu, S. Yao, D. Zhang, J. Du, Z. Hou, et al. A survey of post-training scaling in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2771–2791, 2025.

[21] T. Lieberum, M. Rahtz, J. Kramár, N. Nanda, G. Irving, R. Shah, and V. Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

[22] C. Mathwin, G. Corlouer, E. Kran, F. Barez, and N. Nanda. Identifying a preliminary circuit for predicting gendered pronouns in gpt-2 small. *URL: https://itch. io/jam/mechint/rate/1889871*, page 2, 2023.

[23] A. Messiah. *Quantum mechanics*. Courier Corporation, 2014.

[24] N. Nanda. Attribution patching: Activation patching at industrial scale. https://www.neelnanda.io/mechanistic-interpretability/attribution-patching, Feb 2023. URL https://www.neelnanda.io/mechanistic-interpretability/attribution-patching. Blog post.

[25] A. Panigrahi, N. Saunshi, H. Zhao, and S. Arora. Task-specific skill localization in fine-tuned language models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27011–27033. PMLR, 23–29 Jul 2023.

[26] N. Prakash, T. R. Shaham, T. Haklay, Y. Belinkov, and D. Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8sKcAWOf2D.

[27] D. Rai, Y. Zhou, S. Feng, A. Saparov, and Z. Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

[28] L. Sharkey, B. Chughtai, J. Batson, J. Lindsey, J. Wu, L. Bushnaq, N. Goldowsky-Dill, S. Heimersheim, A. Ortega, J. Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

[29] S. Somvanshi, M. M. Islam, A. Rafe, A. G. Tusti, A. Chakraborty, A. Baitullah, T. I. Chowdhury, N. Alnawmasi, A. Dutta, and S. Das. Bridging the black box: a survey on mechanistic interpretability in ai. *ACM Computing Surveys*, 58(8):1–35, 2026.

[30] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. Slone, A. Rahane, A. S. Iyer, A. J. Andreassen, A. Madotto, A. Santilli, A. Stuhlmüller, A. M. Dai, A. La, A. K. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakaş, B. R. Roberts, B. S. Loe, B. Zoph, B. Bojanowski, B. Özyurt, B. Hedayatnia, B. Neyshabur, B. Inden, B. Stein, B. Ekmekci, B. Y. Lin, B. Howald, B. Orinion, C. Diao, C. Dour, C. Stinson, C. Argueta, C. Ferri, C. Singh, C. Rathkopf, C. Meng, C. Baral, C. Wu, C. Callison-Burch, C. Waites, C. Voigt, C. D. Manning, C. Potts, C. Ramirez, C. E. Rivera, C. Siro, C. Raffel, C. Ashcraft, C. Garbacea, D. Sileo, D. Garrette, D. Hendrycks, D. Kilman, D. Roth, C. D. Freeman, D. Khashabi, D. Levy, D. M. González, D. Perszyk, D. Hernandez, D. Chen, D. Ippolito,

D. Gilboa, D. Dohan, D. Drakard, D. Jurgens, D. Datta, D. Ganguli, D. Emelin, D. Kleyko, D. Yuret, D. Chen, D. Tam, D. Hupkes, D. Misra, D. Buzan, D. C. Mollo, D. Yang, D.-H. Lee, D. Schrader, E. Shutova, E. D. Cubuk, E. Segal, E. Hagerman, E. Barnes, E. Donoway, E. Pavlick, E. Rodolà, E. Lam, E. Chu, E. Tang, E. Erdem, E. Chang, E. A. Chi, E. Dyer, E. Jerzak, E. Kim, E. E. Manyasi, E. Zheltonozhskii, F. Xia, F. Siar, F. Martínez-Plumed, F. Happé, F. Chollet, F. Rong, G. Mishra, G. I. Winata, G. de Melo, G. Kruszewski, G. Parascandolo, G. Mariani, G. X. Wang, G. Jaimovitch-Lopez, G. Betz, G. Gur-Ari, H. Galijasevic, H. Kim, H. Rashkin, H. Hajishirzi, H. Mehta, H. Bogar, H. F. A. Shevlin, H. Schuetze, H. Yakura, H. Zhang, H. M. Wong, I. Ng, I. Noble, J. Jumelet, J. Geissinger, J. Kernion, J. Hilton, J. Lee, J. F. Fisac, J. B. Simon, J. Koppel, J. Zheng, J. Zou, J. Kocon, J. Thompson, J. Wingfield, J. Kaplan, J. Radom, J. Sohl-Dickstein, J. Phang, J. Wei, J. Yosinski, J. Novikova, J. Bosscher, J. Marsh, J. Kim, J. Taal, J. Engel, J. Alabi, J. Xu, J. Song, J. Tang, J. Waweru, J. Burden, J. Miller, J. U. Balis, J. Batchelder, J. Berant, J. Frohberg, J. Rozen, J. Hernandez-Orallo, J. Boudeman, J. Guerr, J. Jones, J. B. Tenenbaum, J. S. Rule, J. Chua, K. Kanclerz, K. Livescu, K. Krauth, K. Gopalakrishnan, K. Ignatyeva, K. Markert, K. Dhole, K. Gimpel, K. Omondi, K. W. Mathewson, K. Chiafullo, K. Shkaruta, K. Shridhar, K. McDonell, K. Richardson, L. Reynolds, L. Gao, L. Zhang, L. Dugan, L. Qin, L. Contreras-Ochando, L.-P. Morency, L. Moschella, L. Lam, L. Noble, L. Schmidt, L. He, L. Oliveros-Colón, L. Metz, L. K. Senel, M. Bosma, M. Sap, M. T. Hoeve, M. Farooqi, M. Faruqui, M. Mazeika, M. Baturan, M. Marelli, M. Maru, M. J. Ramirez-Quintana, M. Tolkiehn, M. Giulianelli, M. Lewis, M. Potthast, M. L. Leavitt, M. Hagen, M. Schubert, M. O. Baitemirova, M. Arnaud, M. McElrath, M. A. Yee, M. Cohen, M. Gu, M. Ivanitskiy, M. Starritt, M. Strube, M. Swędrowski, M. Bevilacqua, M. Yasunaga, M. Kale, M. Cain, M. Xu, M. Suzgun, M. Walker, M. Tiwari, M. Bansal, M. Aminnaseri, M. Geva, M. Gheini, M. V. T, N. Peng, N. A. Chi, N. Lee, N. G.-A. Krakover, N. Cameron, N. Roberts, N. Doiron, N. Martinez, N. Nangia, N. Deckers, N. Muennighoff, N. S. Keskar, N. S. Iyer, N. Constant, N. Fiedel, N. Wen, O. Zhang, O. Agha, O. Elbaghdadi, O. Levy, O. Evans, P. A. M. Casares, P. Doshi, P. Fung, P. P. Liang, P. Vicol, P. Alipoormolabashi, P. Liao, P. Liang, P. W. Chang, P. Eckersley, P. M. Htut, P. Hwang, P. Miłkowski, P. Patil, P. Pezeshkpour, P. Oli, Q. Mei, Q. Lyu, Q. Chen, R. Banjade, R. E. Rudolph, R. Gabriel, R. Habacker, R. Risco, R. Millière, R. Garg, R. Barnes, R. A. Saurous, R. Arakawa, R. Raymaekers, R. Frank, R. Sikand, R. Novak, R. Sitelew, R. L. Bras, R. Liu, R. Jacobs, R. Zhang, R. Salakhutdinov, R. A. Chi, S. R. Lee, R. Stovall, R. Teehan, R. Yang, S. Singh, S. M. Mohammad, S. Anand, S. Dillavou, S. Shleifer, S. Wiseman, S. Gruetter, S. R. Bowman, S. S. Schoenholz, S. Han, S. Kwatra, S. A. Rous, S. Ghazarian, S. Ghosh, S. Casey, S. Bischoff, S. Gehrmann, S. Schuster, S. Sadeghi, S. Hamdan, S. Zhou, S. Srivastava, S. Shi, S. Singh, S. Asaadi, S. S. Gu, S. Pachchigar, S. Toshniwal, S. Upadhyay, S. S. Debnath, S. Shakeri, S. Thormeyer, S. Melzi, S. Reddy, S. P. Makini, S.-H. Lee, S. Torene, S. Hatwar, S. Dehaene, S. Divic, S. Ermon, S. Biderman, S. Lin, S. Prasad, S. Piantadosi, S. Shieber, S. Misherghi, S. Kiritchenko, S. Mishra, T. Linzen, T. Schuster, T. Li, T. Yu, T. Ali, T. Hashimoto, T.-L. Wu, T. Desbordes, T. Rothschild, T. Phan, T. Wang, T. Nkinyili, T. Schick, T. Kornev, T. Tunduny, T. Gerstenberg, T. Chang, T. Neeraj, T. Khot, T. Shultz, U. Shaham, V. Misra, V. Demberg, V. Nyamai, V. Raunak, V. V. Ramasesh, vinay uday prabhu, V. Padmakumar, V. Srikumar, W. Fedus, W. Saunders, W. Zhang, W. Vossen, X. Ren, X. Tong, X. Zhao, X. Wu, X. Shen, Y. Yaghoobzadeh, Y. Lakretz, Y. Song, Y. Bahri, Y. Choi, Y. Yang, S. Hao, Y. Chen, Y. Belinkov, Y. Hou, Y. Hou, Y. Bai, Z. Seid, Z. Zhao, Z. Wang, Z. J. Wang, Z. Wang, and Z. Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj. Featured Certification.

[31] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. Le, E. Chi, D. Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.

[32] A. Syed, C. Rager, and A. Conmy. Attribution patching outperforms automated circuit discovery. In Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, and H. Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

[33] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

[34] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019. In the Proceedings of ICLR.

[35] K. R. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*.

[36] X. Wang, Y. Hu, W. Du, R. Cheng, B. Wang, and D. Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=45EIiFd6Oa.

[37] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[38] L. Yang, S. Ding, and D. Xiong. A local perturbation theory for cross-domain interference and recovery in multi-domain rl, 2026. URL https://arxiv.org/abs/2606.02398.

[39] F. Yin, X. Ye, and G. Durrett. Lofit: Localized fine-tuning on LLM representations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=dfiXFbECSZ.

[40] H. Zhang, Z. Zhang, M. Wang, Z. Su, Y. Wang, Q. Wang, S. Yuan, E. Nie, X. Duan, Q. Xue, et al. Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models. *Computer Science Review*, 62:101011, 2026.

[41] N. Zhang, Y. Yao, B. Tian, P. Wang, S. Deng, M. Wang, Z. Xi, S. Mao, J. Zhang, Y. Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024.

## A The Derivation of Equation 2

In this appendix, we provide the detailed derivation for Equation 2, which approximates the causal effect of an infinitesimal parameter update $\theta'$ using terms computed entirely under the current parameters $\theta$.

First, we treat any activation $a$ as a differentiable function of the model parameters $\theta$. Applying a first-order Taylor expansion with respect to $\theta$, we can approximate the clean and corrupted activations under $\theta'$ as follows:

$$A^{\theta'}(x_r) \approx A^{\theta}(x_r) + \Delta\theta \cdot \frac{\partial A^{\theta}(x_r)}{\partial \theta} \tag{9}$$

$$A^{\theta'}(x_c) \approx A^{\theta}(x_c) + \Delta\theta \cdot \frac{\partial A^{\theta}(x_c)}{\partial \theta} \tag{10}$$

Subtracting these two equations provides a linear approximation of the activation difference, separating the original difference from the effect of the parameter update:

$$A^{\theta'}(x_c) - A^{\theta'}(x_r) \approx (A^{\theta}(x_c) - A^{\theta}(x_r)) + \Delta\theta \cdot \left( \frac{\partial A^{\theta}(x_c)}{\partial \theta} - \frac{\partial A^{\theta}(x_r)}{\partial \theta} \right) \tag{11}$$

Next, we approximate the gradient term. Let the gradient under the current parameters $\theta$ be denoted as $g(\theta) = \frac{\partial f_a^{\theta}}{\partial a^{\theta}}$. Applying a first-order Taylor expansion to the gradient for the updated parameters $\theta'$ yields:

$$g(\theta') \approx g(\theta) + \Delta\theta \cdot \frac{\partial g(\theta)}{\partial \theta} \tag{12}$$

By substituting the definition of $g(\theta)$ into the derivative term, we can further expand it into a mixed second-order derivative:

$$\frac{\partial g(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \left( \frac{\partial f_a^{\theta}}{\partial a^{\theta}} \right) = \frac{\partial^2 f_a^{\theta}}{\partial a^{\theta} \partial \theta} \tag{13}$$

Finally, we substitute both the activation difference approximation and the gradient approximation back into the Attribution Patching formulation for $\theta'$. By expanding the product and omitting higher-order terms $\mathcal{O}((\Delta\theta)^2)$, we arrive at our final formulation:

$$\begin{aligned}
f_a^{\theta'}(A(x_c)) - f_a^{\theta'}(A(x_r)) &\approx \left( A^{\theta'}(x_c) - A^{\theta'}(x_r) \right) \cdot \left. \frac{\partial f_a^{\theta'}}{\partial a^{\theta'}} \right|_{a = A^{\theta'}(x_r)} \\
&\approx \left[ \left( A^{\theta}(x_c) - A^{\theta}(x_r) \right) + \Delta\theta \cdot \left( \frac{\partial A^{\theta}(x_c)}{\partial \theta} - \frac{\partial A^{\theta}(x_r)}{\partial \theta} \right) \right] \\
&\quad \cdot \left[ g(\theta) + \Delta\theta \cdot \frac{\partial^2 f_a^{\theta}}{\partial a^{\theta} \partial \theta} \right] \\
&= \left( A^{\theta}(x_c) - A^{\theta}(x_r) \right) \cdot \frac{\partial f_a^{\theta}}{\partial a^{\theta}} + \Delta\theta \cdot \left( A^{\theta}(x_c) - A^{\theta}(x_r) \right) \cdot \frac{\partial^2 f_a^{\theta}}{\partial a \partial \theta} \\
&\quad + \Delta\theta \cdot \left( \frac{\partial A^{\theta}(x_c)}{\partial \theta} - \frac{\partial A^{\theta}(x_r)}{\partial \theta} \right) \cdot \frac{\partial f_a^{\theta}}{\partial a^{\theta}} + \mathcal{O}((\Delta\theta)^2) \\
&\approx f_a^{\theta}(A(x_c)) - f_a^{\theta}(A(x_r)) \\
&\quad + \Delta\theta \cdot \left[ \left( \frac{\partial A^{\theta}(x_c)}{\partial \theta} - \frac{\partial A^{\theta}(x_r)}{\partial \theta} \right) \cdot \frac{\partial f_a^{\theta}}{\partial a^{\theta}} + \left( A^{\theta}(x_c) - A^{\theta}(x_r) \right) \cdot \frac{\partial^2 f_a^{\theta}}{\partial a \partial \theta} \right] \\
&= f_a^{\theta}(A(x_c)) - f_a^{\theta}(A(x_r)) + \Delta\theta \cdot S(\theta)
\end{aligned}$$

where $S(\theta)$ effectively captures the sensitivity of the target neuron $a$'s attribution value to the parameter updating.

## B Derivation of the Translation Operator for Sensitivity $S(\theta)$

In this section, we provide the mathematical proof demonstrating how the sensitivity at the ideal state, $S(\hat{\theta}^I)$, can be evaluated as $S(\theta + K\Delta\theta)$ by circumventing the intractable higher-order derivatives.

**Simplifying the Notation and Revealing the Identity.** Let $\Delta A^\theta = A^\theta(x_c) - A^\theta(x_r)$. According to Equation 3, the sensitivity $S(\theta)$ can be rewritten as:

$$S(\theta) = \frac{\partial \Delta A^\theta}{\partial \theta} \cdot \frac{\partial f_a^\theta}{\partial a^\theta} + \Delta A^\theta \cdot \frac{\partial^2 f_a^\theta}{\partial a \partial \theta}$$

Applying the product rule in reverse reveals that $S(\theta)$ is strictly equivalent to the gradient of the attribution score with respect to the parameters:

$$S(\theta) \equiv \nabla_\theta \left[ \Delta A^\theta \cdot \frac{\partial f_a^\theta}{\partial a^\theta} \right] \equiv \nabla_\theta [\Delta E^\theta_{\mathcal{D}_\mathcal{T}}(a)]$$

This insight indicates that our iterative update formula, $\Delta E^{\theta^{k+1}}_{\mathcal{D}_\mathcal{T}}(a) = \Delta E^{\theta^k}_{\mathcal{D}_\mathcal{T}}(a) + \Delta\theta^k \cdot S(\theta^k)$, is fundamentally solving an Ordinary Differential Equation (ODE) via Euler's Method.

**The Differential Operator and Infinite Series.** Assuming we perform infinite micro-updates along a fixed direction $\Delta\theta$, we define a differential operator $D$:

$$D = \Delta\theta \cdot \nabla_\theta$$

which computes the directional derivative along $\Delta\theta$. Utilizing the Taylor expansion, the sensitivity at step $k$ unfolds into an infinite series:

$$S(\theta^k) = S(\theta^0) + kDS(\theta^0) + \frac{k^2}{2!}D^2S(\theta^0) + \frac{k^3}{3!}D^3S(\theta^0) + \dots$$

Notice that the operators $D^2, D^3, \dots$ correspond exactly to the computationally intractable second-order Hessians, third-order tensors, and beyond.

**Operator Folding and Translation.** In functional analysis, this infinite series perfectly matches the Taylor expansion of $e^x$. We can rigorously fold it into an exponential operator:

$$S(\theta^k) = \left( \sum_{n=0}^{\infty} \frac{(kD)^n}{n!} \right) S(\theta^0) = e^{k\Delta\theta\cdot\nabla_\theta} S(\theta^0)$$

In Lie Algebra [15] and Quantum Mechanics [23], $e^{v\cdot\nabla}$ represents the Translation Operator, which systematically shifts the independent variable of any analytic function: $e^{v\cdot\nabla} f(x) \equiv f(x+v)$. Applying this property to our exponential operator yields:

$$e^{k\Delta\theta\cdot\nabla_\theta} S(\theta^0) \equiv S(\theta^0 + k\Delta\theta) \implies S(\theta^k) = S(\theta^0 + k\Delta\theta)$$

Consequently, for the final ideal state $I$ (reached after $K$ steps), we arrive at $S(\theta^I) = S(\theta^0 + K\Delta\theta)$. This elegant identity demonstrates that to evaluate the sensitivity at the ideal state, we merely need to scale the step distance $K$, completely avoiding the computation of higher-order derivative terms.

# C Why the Probing Model's Attribution Cannot Directly Substitute the Ideal Model

In this appendix, we demonstrate why it is theoretically flawed to directly substitute the attribution patching value of the probing model, denoted here as an arbitrary discrete step $\theta^k$, for that of the ideal model $\theta^I$. That is, we prove why the approximation $\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a) \approx \Delta E^{\theta^k}_{\mathcal{D}_\mathcal{T}}(a)$ does not generally hold.

To simplify the notation, let $X_0 = A^{\theta^0}(x_c) - A^{\theta^0}(x_r)$ and $Y_0 = \frac{\partial f_a^{\theta^0}}{\partial a^{\theta^0}}$ represent the activation difference and gradient for the current model $\theta^0$, respectively. Similarly, let $X_k = A^{\theta^k}(x_c) - A^{\theta^k}(x_r)$ and $Y_k = \frac{\partial f_a^{\theta^k}}{\partial a^{\theta^k}}$ represent the corresponding terms for the probing model at step $k$.

Recall our causal effect estimation formulation:

$$\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a) \approx X_0 Y_0 + \frac{1}{2}\left[S(\theta^0) + S(\theta^k)\right] \cdot \Delta\theta \tag{14}$$

Expanding the sensitivity term $S(\theta^0) \cdot \Delta\theta$ yields:

$$S(\theta^0) \cdot \Delta\theta = \left( \frac{\partial X_0}{\partial \theta} \cdot \Delta\theta \right) \cdot Y_0 + X_0 \cdot \left( \frac{\partial Y_0}{\partial \theta} \cdot \Delta\theta \right) \tag{15}$$

At this juncture, we introduce the ***HVP Assumption*** (which will be detailed in Section 4.2), allowing us to approximate the gradient variation as $\frac{\partial Y_0}{\partial \theta} \cdot \Delta\theta \approx Y_k - Y_0$. Applying this assumption, we obtain:

$$S(\theta^0) \cdot \Delta\theta \approx \left( \frac{\partial X_0}{\partial \theta} \cdot \Delta\theta \right) \cdot Y_0 + X_0(Y_k - Y_0) \tag{16}$$

By symmetric logic, for the probing state $\theta^k$:

$$S(\theta^k) \cdot \Delta\theta \approx \left( \frac{\partial X_k}{\partial \theta^k} \cdot \Delta\theta \right) \cdot Y_k + X_k(Y_k - Y_0) \tag{17}$$

The equivalence $\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a) \approx \Delta E^{\theta^k}_{\mathcal{D}_\mathcal{T}}(a)$ can only be established if we impose an extremely strong assumption regarding the activation dynamics:

$$\frac{\partial (A(x_c) - A(x_r))}{\partial \theta} \cdot \Delta\theta \approx X_k - X_0 \tag{18}$$

*If and only if* this assumption holds, we can substitute it into Equation 16 and Equation 17:

$$S(\theta^0) \cdot \Delta\theta \approx (X_k - X_0)Y_0 + X_0(Y_k - Y_0) = X_kY_0 + X_0Y_k - 2X_0Y_0 \tag{19}$$

$$S(\theta^k) \cdot \Delta\theta \approx (X_k - X_0)Y_k + X_k(Y_k - Y_0) = 2X_kY_k - X_0Y_k - X_kY_0 \tag{20}$$

Summing these two refined equations together provides:

$$\frac{1}{2}\left[S(\theta^0) + S(\theta^k)\right] \cdot \Delta\theta \approx \frac{1}{2}[2X_kY_k - 2X_0Y_0] = X_kY_k - X_0Y_0 \tag{21}$$

Finally, substituting this result back into our main formulation yields:

$$\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a) \approx X_0Y_0 + X_kY_k - X_0Y_0 = X_kY_k = \Delta E^{\theta^k}_{\mathcal{D}_\mathcal{T}}(a) \tag{22}$$

This derivation reveals that directly using the probing model's attribution relies on the assumption that the activation differences change perfectly linearly with the parameter update across the macroscopic step from $\theta^0$ to $\theta^k$. Because deep neural networks are highly non-linear, this stringent constraint is almost never satisfied during discrete gradient updates, rendering the naive substitution $\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a) \approx \Delta E^{\theta^k}_{\mathcal{D}_\mathcal{T}}(a)$ theoretically unsound.

# D Details of Datasets

To facilitate the implementation of mechanistic interpretability, we strictly constrain the label of each dataset to a single token. Consequently, for specific datasets, we reformulate the inputs into question-answering prompt templates. The Arithmetic dataset is systematically partitioned into sub-datasets based on *digit* and *step* complexity, encompassing solely the four fundamental operations (addition, subtraction, multiplication, and division). Here, *digit* denotes the number of digits in the participating operands, while *step* indicates the number of mathematical operators within a single sample (excluding the equals sign). The statistical distribution and prompt examples for all datasets are detailed in Table 3.

# E Experiment of Multi-Task Fine Tuning

To verify whether our localization method can mitigate mechanistic inter-task conflicts through forward-looking interpretability insights, we design a multi-task joint fine-tuning setup where such conflicts are both controllable and observable. Sourcing candidate tasks from Appendix D, we randomly construct joint task sets ranging from 2 to 6 tasks, ensuring at least three distinct random combinations for each set size. We evaluate the mean and variance of the Target Task Accuracy

| Dataset | | Train | Test | Max Len | Output | Input Prompt Case | Label |
|---|---|---|---|---|---|---|---|
| IOI | | 6,000 | 600 | 50 | Correct name | Then, Aaron and Amber were working at the house. Aaron decided to give a drink to | Amber |
| Gender | | 3,000 | 3,000 | 50 | he / she | So Kelly is always up for an adventure, isn't | she |
| BOOL | | 10,000 | 4,000 | 100 | True / False | Evaluate the following boolean expression as either 'True' or 'False'.<br>( not ( True or True ) ) and ( False or ( not True ) ) | False |
| SST-2 | | 67,349 | 872 | 200 | positive / negative | Is the sentiment of following sentence positive or negative? that 's far too tragic to merit such superficial treatment. Answer: It is | negative |
| MRPC | | 3,670 | 1,730 | 200 | Yes / No | Are the meanings of the following two sentences equivalent? 1. And if both apply, they are essentially possible. 2. And if both apply, they are essentially impossible. | No |
| QQP | | 364,000 | 391,000 | 200 | Yes / No | Is Question "How do you control your horniness?" a duplicate of Question "How do I control my horny emotions?"? | Yes |
| MNLI | | 393,000 | 9,800 | 250 | Option Letter | What is the relationship of premise "How do you know? All this is their information again." to hypothesis "This information belongs to them.": A. entailment, B. contradiction, or C. unrelated? The answer is | A |
| RTE | | 2,490 | 3,000 | 200 | Yes / No | Does sentence "Oil prices fall back as Yukos oil threat lifted" entail sentence "Oil prices rise."? | Yes |
| WinoGrande | | 9,248 | 1,267 | 200 | Yes / No | The trophy doesn't fit into the brown suitcase because _ is too large. Should the '_' be the trophy? | Yes |
| Docstring | | 2,400 | 600 | 50 | Missing param | def process(self, data, name, value, result, line, file):<br>"""control action cost<br>:param value: research subject<br>:param result: health issue<br>:param | line |
| Induction | | 2,400 | 600 | 600 | Last token | After Gray applied for health insurance, Alice applied | for |
| Arithmetic | 2-digit | 10,000 | 600 | 160 | Option Letter | Please choose the correct option from the following: "What is 80 plus 29?" Options: A. #####, B. dnfidlg, C. 378, D. 54942, E. 388, F. 109, G. 4. The answer is | F |
| | 3-digit | 10,000 | 600 | 160 | Option Letter | Please choose the correct option from the following: "What is 161 plus 390?" Options: A. 16957, B. 551, C. people, D. 908, E. 305, F. 5, G. gkledns. The answer is | B |
| | 7-digit | 10,000 | 600 | 160 | Option Letter | Please choose the correct option from the following: "What is 5567345 plus 4592838?" Options: A. 10160183, B. 552493, C. apple, D. 374329826, E. 5567838, F. dhjklng, G. 387439. The answer is | A |
| | 1-step | 10,000 | 600 | 160 | Option Letter | Please choose the correct option from the following: "34-62=" Options: A. -28, B. 34, C. from, D. 345, E. 11, F. 63, G. right The answer is | A |
| | 2-step | 10,000 | 600 | 160 | Option Letter | Please choose the correct option from the following: "6361-32*537=" Options: A. ill, B. 10823, C. -10823, D. 583567, E. 3562, F. -55, G. same. The answer is | C |
| | 6-step | 10,000 | 600 | 160 | Option Letter | Please choose the correct option from the following: "5392/57892+92743-372984/239*4428+3792=" Options: A. -6813812, B. 205637, C. lofdt, D. 3034672, E. 2647, F. 109872, G. muly The answer is | A |

Table 3: Dataset statistics, configurations, and prompt construction examples.

(TTA) across these combinations during joint fine-tuning. As illustrated in Figure 5, our method not only achieves the highest average performance in multi-task scenarios but also demonstrates superior stability, evidenced by nearly the lowest variance. This indicates the absence of severe inter-task conflicts that would otherwise cause catastrophic degradation on individual tasks.

To further substantiate this conclusion, we trace the evolution of conflicting nodes within the model circuits across checkpoints from $0$ to $400$ fine-tuning iterations. By utilizing the boolean solver introduced in the CLUE method to quantify these conflicts, Figure 6 reveals that our approach significantly minimizes the emergence of conflicting nodes. Its efficacy is surpassed only by unconstrained full-parameter fine-tuning, which inherently represents a natural mechanistic evolution with maximum degrees of freedom. Conversely, baseline methods perform comparably to random localization, indicating a lack of localization precision. Consequently, they fail to reduce conflicting nodes, demonstrating an inability to effectively leverage mechanistic interpretability for practical fine-tuning guidance.

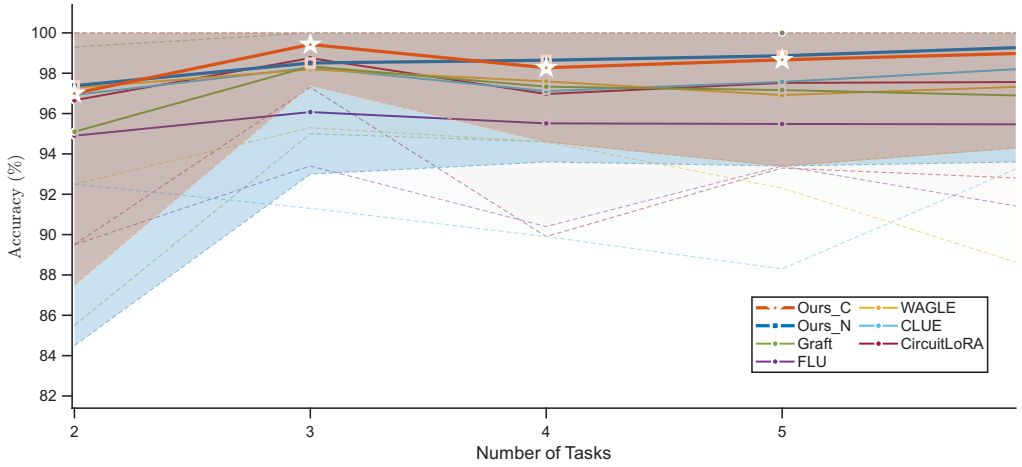



Figure 5: Performance in multi-task accuracy. The solid line represents the mean value, and the dashed line represents the lower bound of the value variation. The shaded areas represent the variances of Our_C and Our_N, respectively.

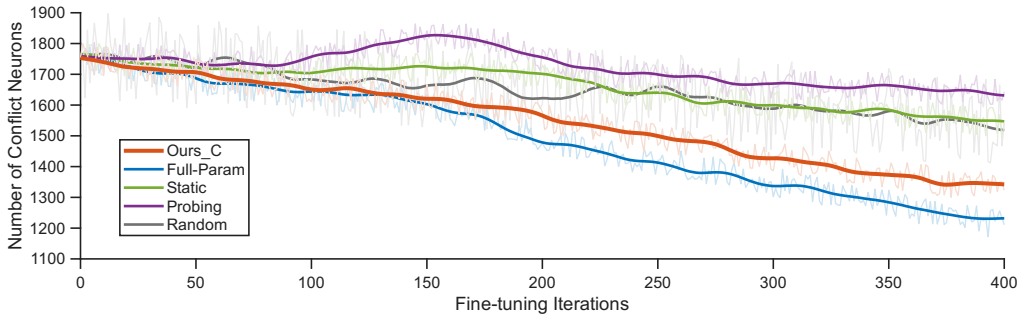



Figure 6: Evolution of conflict component in fine-tuning.

# F Exploration and Experiments about $K$

## F.1 The Optimal Sampling Interval of $K$

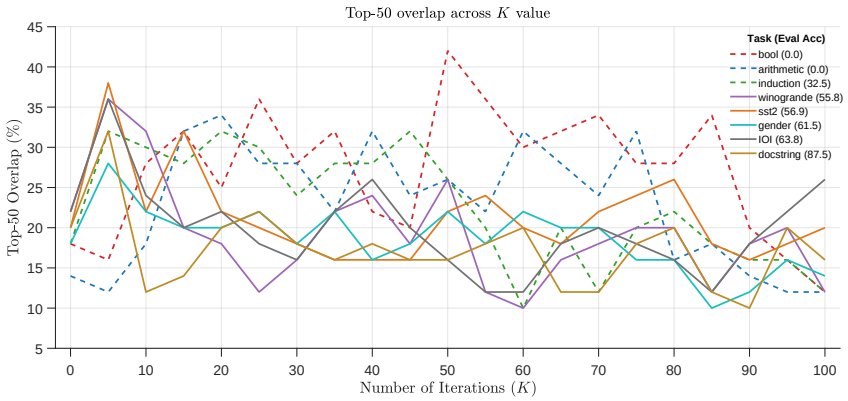



Figure 7: Overlap rate of the top 50 components with the full-parameter circuit ($\mathcal{C}_{\text{Full-Param}}$) across tasks with varying initial accuracies.

As discussed in Section 4, our estimation of $\hat{\theta}^I$ involves averaging the outcomes of ten randomly sampled values for $K$. Empirically, we observe that each SFT task exhibits a specific optimal

sampling interval; sampling $K$ within this range yields a significantly higher expected performance compared to sampling outside of it.

To determine this optimal sampling interval for each task, we utilize the fully fine-tuned model (without any localization or parameter freezing) as a surrogate for the ideal model. This choice is motivated by two factors: (1) The fully fine-tuned model is free from auxiliary parameter constraints, thus most closely approximating the natural trajectory of mechanistic evolution. (2) It guarantees a satisfactory accuracy on the target SFT task, ensuring the formation of a complete and well-defined task mechanism within the model. For brevity, we substitute the ideal model parameters ($\theta^I$) with those of the fully fine-tuned model throughout this paper. Subsequently, we uniformly sample $K$ from $0$ to $100$ at intervals of $5$. We then compute the overlap of the top 50 components—ranked by their causal effect—between our estimated $\hat{\theta}^I$ and the surrogate $\theta^I$. The results are illustrated in Figure 7.

The values in parentheses within the figure legend indicate the initial Target Task Accuracy (TTA) prior to fine-tuning. A clear correlation emerges between this initial TTA and the optimal sampling interval for $K$. For tasks with an initial TTA near zero (e.g., BOOL and Arithmetic), the expected overlap remains consistently high across a broad range of $K \in [10, 80]$. Conversely, for tasks with a higher initial TTA (e.g., Docstring, IOI, and Gender), the optimal $K$ values are highly concentrated within a narrow band of $[0, 10]$. We attribute this phenomenon to the initial TTA reflecting the completeness of the pre-existing task mechanism within the base model. A lower initial TTA implies a deficient mechanism, necessitating more substantial parameter updates during fine-tuning; hence, a larger $K$ is required to adequately extrapolate and explore the distribution of $\hat{\theta}^I$. Conversely, a higher initial TTA indicates that the model mechanism is already largely intact, requiring only minor parameter updates, allowing a smaller $K$ to sufficiently approximate $\hat{\theta}^I$.

### F.2 Automated and Reliable Determination of $K$ via Linear Probing

| Method / Strategy | Bool | | | SST-2 | | | Arithmetic | | |
|---|---|---|---|---|---|---|---|---|---|
| | $K$ | Top@50 | TTA | $K$ | Top@50 | TTA | $K$ | Top@50 | TTA |
| 1-Sample Uniform | 35.6 | 28 | 99.64 | 2.1 | 26 | 96.34 | 71.5 | 30 | 99.51 |
| 10-Sample Average | [10, 80] | 30.4 | 100.00 | [1, 10] | 29.6 | 97.55 | [10, 80] | 31.4 | 100.00 |
| Probe (First Layer) | 15.3 | 24 | 98.82 | 8.73 | 28 | 95.17 | 24.8 | 26 | 99.17 |
| Probe (Middle Layer) | 67.2 | 32 | 99.67 | 1.2 | 34 | 96.72 | 72.4 | 30 | 99.88 |
| Probe (Last Layer) | 58.5 | 34 | 99.83 | 1.2 | 30 | 97.26 | 77.2 | 32 | 100.00 |

Table 4: Ablation study on determining the extrapolation scalar $K$ via linear probing across different residual stream depths on the Mistral-7B backbone, evaluated against standard sampling baselines.

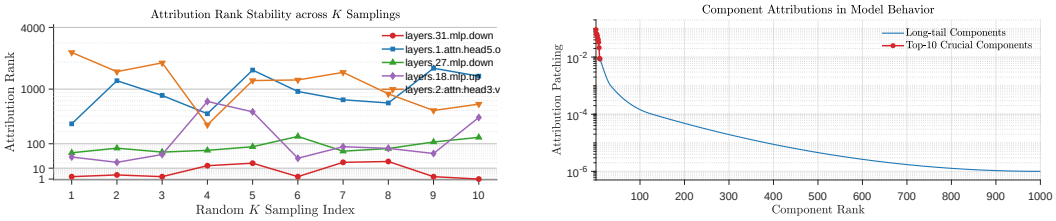
In our primary locating-then-tuning pipeline, the step scalar $K$ is estimated by averaging across 10 randomly sampled values within a plausible interval. To eliminate the computational overhead of repeated samplings and mitigate heuristic bias in scenarios where a surrogate fully fine-tuned model is unavailable, we investigate an automated probing-based framework to reliably predict $K$ prior to full SFT. Specifically, we construct a lightweight, two-layer linear probing head and attach it to the residual stream of the Mistral-7B backbone at three representative depths: the *First Layer* (Layer 1), the *Middle Layer* (Layer 16), and the *Last Layer* (Layer 32). The probing head is trained using the Top@50 circuit overlap against the surrogate ideal model as the supervised optimization target.

Table 4 compares the proposed probing strategy against standard heuristic baselines (1-Sample and 10-Sample sampling) across the Bool, SST-2, and Arithmetic benchmarks in terms of predicted $K$ values, Top@50 circuit overlap, and downstream Target Task Accuracy (TTA).

Our empirical observations yield two key insights: 1. The probing model accurately predicts the task-specific $K$ value prior to tuning. It matches or even surpasses the circuit overlap and TTA achieved by the 10-sample averaging method, successfully circumventing the tedious time and computational expenditure induced by multi-round sampling. 2. Deploying the probe at the middle or deep residual stream yields significantly superior predictions compared to the first layer. This discrepancy stems from the representational hierarchy in LLMs: activations in the shallowest residual stream predominantly capture token-level lexical and low-level syntactic information, lacking the

high-level semantic abstractions necessary to discern whether the base model already possesses relevant downstream task mechanisms.

### F.3 Robustness of Sampling $K$




(a) Rank results of 10 samplings for sample nodes.

(b) Distribution of attribution value estimates.

Figure 8: Robustness analysis of K-value sampling.

To analyze the stability of our $K$-value sampling, we select a subset of components and track their rank variations across ten independent samplings, as depicted in Figure 8a. The average rankings of these sampled components span a wide spectrum, ranging from the top 10 to beyond 2000. Empirical observations indicate that top-ranked components exhibit minimal rank variance across different samplings.

Recall that the estimated causal effect $\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a)$ comprises a baseline term $\Delta E^{\theta^0}_{\mathcal{D}_\mathcal{T}}(a)$ and a $K$-dependent extrapolation term (i.e., $\frac{1}{2}\left[S(\theta^0) + S(K \cdot \Delta\theta)\right] \cdot K \cdot \Delta\theta$). By examining the exact magnitudes of these two terms, we uncover an intriguing phenomenon: even for the highest-ranked components, the baseline term $\Delta E^{\theta^0}_{\mathcal{D}_\mathcal{T}}(a)$ remains consistently at least one order of magnitude smaller than the $K$-dependent term. This implies that the observed rank stability is not dominated by the intrinsic causal effect of the current model ($\theta$). Instead, the robust gradient direction dictates substantial causal effects across varying $K$ values, precisely isolating the critical nodes that govern the target task.

Furthermore, Figure 8b illustrates the magnitude distribution of the top 1000 components, revealing an extreme concentration where the top 10 components account for nearly 90% of the total causal effect. Consequently, any severe rank fluctuations among lower-ranked components during sampling are mathematically negligible and do not compromise the overall efficacy of the localization.

## G Robustness and Sensitivity Ablation across Stochastic Factors

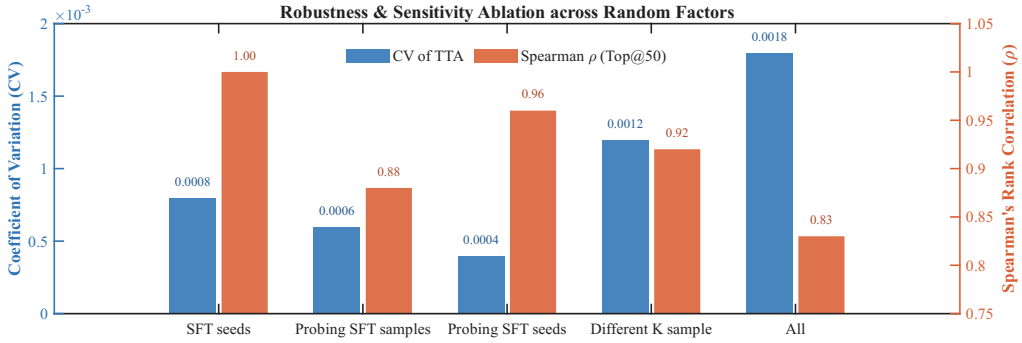



Figure 9: $\rho$ and $CV$ across different stochastic factors. Each set is randomly repeated 10 times on Ours_C in the Mistral-7B model.

To rigorously evaluate the algorithmic stability and reproducibility of our framework, we systematically ablate and quantify the impact of key stochastic factors introduced across different operational stages. Specifically, we examine the following five perturbation setups:

1. **SFT Seeds:** Random seeds governing the fine-tuning stage, which control the stochastic initialization of trainable LoRA projection matrices and parameter masking.
2. **Probing SFT Samples:** Stochasticity arising from uniformly sub-sampling different $1\%$ data partitions during the probing SFT phase.
3. **Probing SFT Seeds:** Random seeds used during probing SFT optimization, which alter the intermediate probing parameter state $\theta'$.
4. **Different $K$ Sampling:** Stochasticity induced by independently drawing distinct subsets of the extrapolation step scalar $K$ when estimating $\Delta E^{\theta^I}_{\mathcal{D}_\mathcal{T}}(a)$.
5. **All (Joint Perturbation):** The unconstrained scenario where all four stochastic factors above are simultaneously randomized across runs.

We assess robustness from two complementary perspectives: the intrinsic stability of the identified task circuits and the extrinsic invariance of downstream task performance:

- **Critical Component Stability ($\rho$):** We extract the Top@50 component rankings across independent random runs and compute their pairwise *Spearman's Rank Correlation Coefficient* ($\rho \in [0, 1]$). A correlation value closer to 1 indicates that the identified computational subgraphs remain highly invariant under random perturbations.
- **Downstream Performance Invariance ($CV$):** We measure the dispersion of the final Target Task Accuracy (TTA) using the *Coefficient of Variation* ($CV$):$CV = \frac{\sigma}{\mu} \times 100\%$, where $\mu$ and $\sigma$ denote the empirical mean and standard deviation of TTA, respectively. A lower $CV$ signifies greater insensitivity of the downstream adaptation to stochastic noise.

**Empirical Analysis and Insights.** As illustrated in Figure 9, the proposed framework exhibits exceptional resilience across all evaluated dimensions. The downstream performance maintains near-zero variance across all isolated and joint perturbation settings ($CV \leq 0.0018$), while the structural ranking of the Top@50 critical components exhibits consistently high correlation ($\rho \geq 0.83$, reaching $1.00$ under varying SFT seeds). These findings confirm that the primary stochastic factors in our pipeline introduce virtually negligible performance instability. From a mechanistic standpoint, this suggests that under our current localization granularity, downstream task adaptation is predominantly governed by a sparse set of dominant, high-attribution neurons and components. Probing with a small parameter displacement ($\Delta\theta$) reliably captures the principal gradient direction of the loss landscape, enabling the extrapolation term to consistently isolate these core task-governing circuits without being derailed by local stochastic sampling fluctuations.

# H Analysis from Mechanistic Interpretability

| Method | Induction | | Gender | | IOI | | Docstring | |
|---|---|---|---|---|---|---|---|---|
| | **Top@50** | **KL** | **Top@50** | **KL** | **Top@50** | **KL** | **Top@50** | **KL** |
| $\mathcal{C}_{\text{Full-Param}}$ | 100 | 0 | 100 | 0 | 100 | 0 | 100 | 0 |
| $\mathcal{C}_{\text{Static}}$ | 22 | 1.37 | 18 | 1.57 | 20 | 1.72 | 24 | 1.31 |
| $\mathcal{C}_{\text{Probing}}$ | 18 | 1.29 | 22 | 2.23 | 22 | 1.25 | 22 | 1.04 |
| $\mathcal{C}_{\text{Ours}}$ | 28 | 1.03 | 36 | 1.31 | 32 | 0.93 | 28 | 1.01 |

Table 5: Comparison of circuit overlap (Top@50) and KL divergence against the full-parameter finetuned model across interpretability datasets in component level.

To provide a rigorous theoretical validation of the differences between our approach and the static/probing localization methods, we employ Edge Attribution Patching (EAP) to construct mechanistic circuits ($\mathcal{C}$) based on the estimated causal effects. Briefly, these circuits are formulated as directed acyclic graphs (DAGs), where nodes denote computational units (neurons or components) and edges represent the activation flows between them. We evaluate the circuits derived from various localization methods using two metrics: the overlap of the top-50 causal effect nodes with the ideal model, and the Kullback-Leibler (KL) divergence between the output logits of the ideal model's circuit and those generated when strictly activating only the nodes and edges within the identified

| Method | Induction | | Gender | | IOI | | Docstring | |
|---|---|---|---|---|---|---|---|---|
| | Top@50 | KL | Top@50 | KL | Top@50 | KL | Top@50 | KL |
| $\mathcal{C}_{\text{Full-Param}}$ | 100 | 0 | 100 | 0 | 100 | 0 | 100 | 0 |
| $\mathcal{C}_{\text{Static}}$ | 20 | 1.55 | 24 | 1.61 | 26 | 1.49 | 20 | 1.52 |
| $\mathcal{C}_{\text{Probing}}$ | 20 | 1.51 | 18 | 1.72 | 20 | 1.62 | 18 | 1.27 |
| $\mathcal{C}_{\text{Ours}}$ | 32 | 1.11 | 34 | 1.13 | 34 | 1.21 | 30 | 1.15 |

Table 6: Comparison of circuit overlap (Top@50) and KL divergence against the full-parameter finetuned model across interpretability datasets in neuron level.

circuit. Consistent with our previous setups, we adopt the fully fine-tuned Mistral-7B model as the ideal surrogate and conduct analyses across four interpretability tasks. Given that the fully fine-tuned model achieves 100% accuracy on these specific tasks, its derived circuit serves as a robust ground truth.

Tables 5 and 6 report the results at the component and neuron levels, respectively. Notably, our method yields circuits most closely aligned with the ideal model, demonstrating its superior capability in accurately predicting the post-fine-tuning mechanistic distribution. Furthermore, the negligible performance gap between the probing and static methods implies that the mechanistic distribution of $\theta'$ remains heavily anchored to the initial state $\theta$, thereby lacking the necessary foresight to anticipate mechanistic shifts during the fine-tuning process.

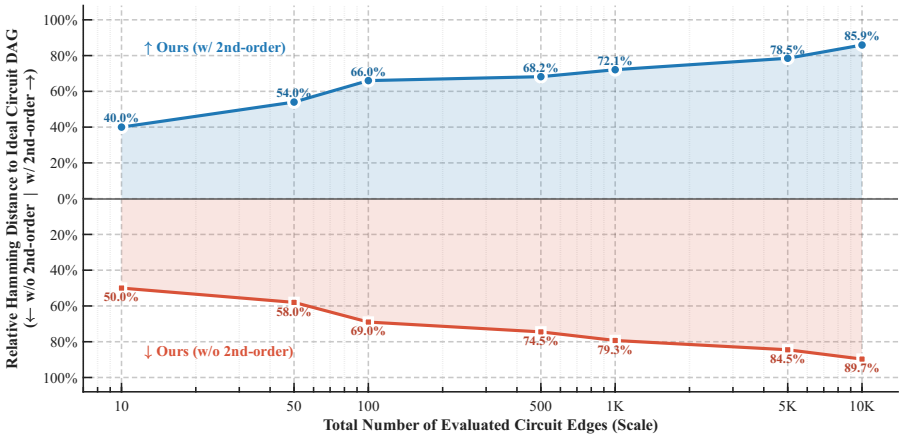
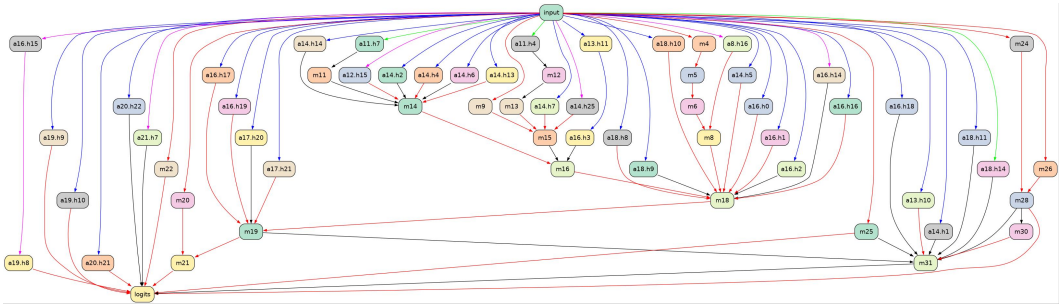



Figure 10: Ablation on Second-Order Approximation across Expanding Edge Capacities

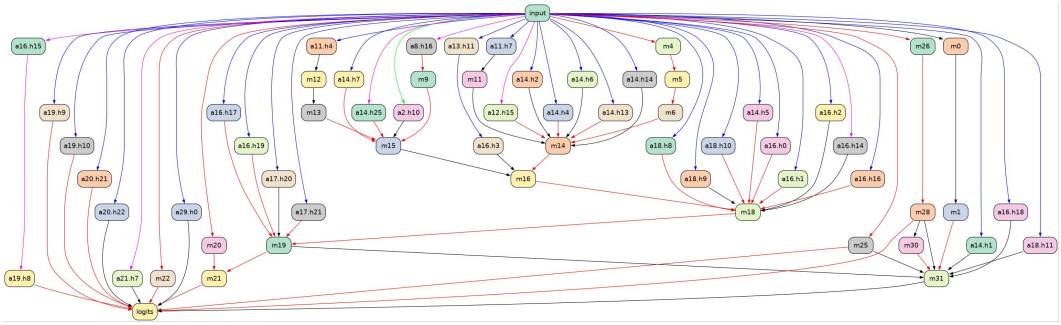
Furthermore, to rigorously evaluate the topological fidelity of our predicted circuit graphs and assess the empirical impact of the mixed second-order derivative approximation (Equation 8), we systematically analyze circuit discrepancies against the ideal surrogate model across expanding edge capacities. We measure the *Relative Hamming Distance* (defined as the normalized Hamming distance over the total number of evaluated edges), where a value approaching $100\%$ indicates near-complete divergence between two graph topologies.

As illustrated in Figure 10, while approximating the second-order partial derivative inevitably introduces mild algebraic perturbations, it consistently maintains superior structural fidelity compared to the pure second-order computation across all scales.

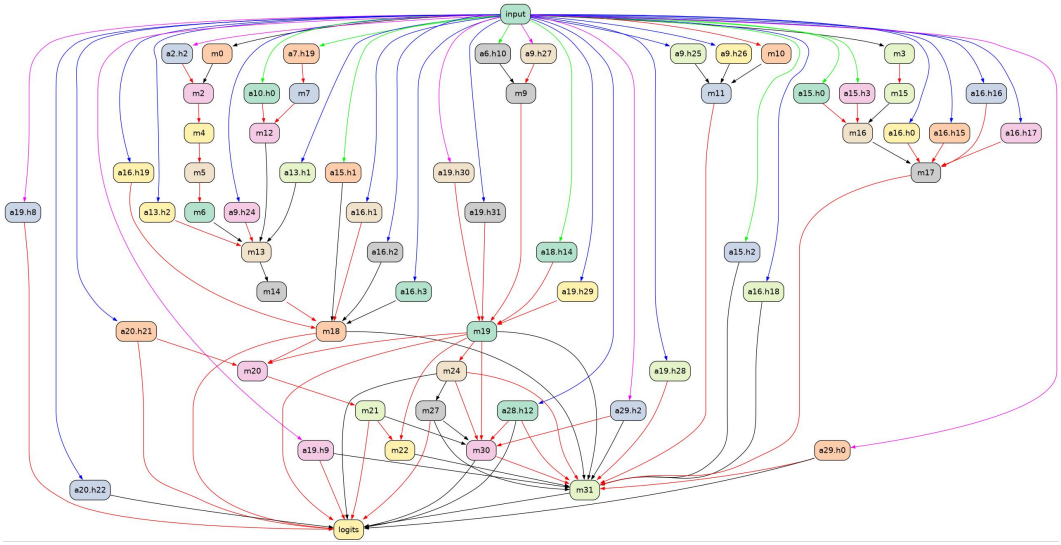
Crucially, when synthesized with the attribution magnitude distribution shown in Figure 8b, the empirical efficacy of our framework is overwhelmingly governed by the accurate and robust ranking of the most prominent components (approximately within the Top@50). As established in our attribution analysis, these top-tier computational units account for the vast majority (nearly $90\%$) of the total causal effect governing task adaptation.




(a) $\mathcal{C}_{\text{Static}}(\theta)$




(b) $\mathcal{C}_{\text{Probing}}(\theta')$




(c) $\mathcal{C}_{\text{Ours}}(\hat{\theta}^I)$

Figure 11: Circuit graphs corresponding to the three types of parameters.

# I Circuit Graph Visualizations

We provide visualizations of the circuit graphs generated by different localization methods for the IOI task using the Mistral-7B model. As illustrated in Figure 11, the circuit predicted by our method ($\hat{\theta}^I$) incorporates notably more intermediate nodes compared to those derived from $\theta$ and $\theta'$, leading to more intricate computational pathways from input to output. This structural complexity indicates that our estimated circuit anticipates richer task-specific mechanisms. Notably, these emergent mechanisms are fundamentally absent in the current model parameters, as the base model has not yet fully acquired the target capability.

# J Dynamic of Learning Factors for Probing SFT

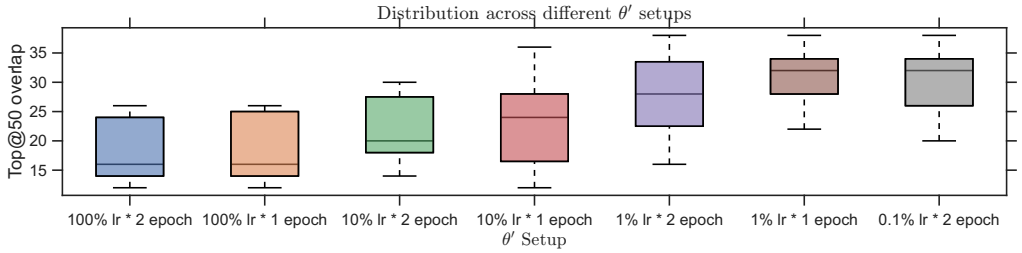



Figure 12: Top 50 overlap with Full-Parameter fine-tuning model across different probing SFT setups.

To empirically justify our hyperparameter selection for the probing SFT, we conduct an additional ablation study on the learning rate and training epochs. We utilize the fully fine-tuned Mistral-7B model as the surrogate ideal model and evaluate the component-level estimation pipeline on the IOI task, fixing the training data size to $1\%$ of the dataset. As illustrated in Figure 12, we observe a phenomenon that mirrors the findings in Figure 4: reducing the learning rate and the number of epochs—which effectively restricts the magnitude of the parameter update—counterintuitively yields a higher circuit overlap with the ideal model. This observation further corroborates our theoretical assertion in Section 4.1: maintaining a sufficiently small parameter update during probing is imperative to preserve the necessary degrees of freedom for the extrapolation term $K \cdot \Delta\theta$. Based on these empirical validations, we finalize our probing SFT configuration as utilizing $1\%$ of the training samples, trained for a single epoch at $1\%$ of the original learning rate.