Do Attribution Circuits Generalize? A Controlled Study of Feature and Attention Circuits on Indirect Object Identification in GPT-2 Small

David Zhan
University of Pennsylvania
zhandavid4@gmail.com

Abstract

Attribution graphs—directed graphs over model components weighted by Input×Gradient (IxG) scores—are the field's emerging standard for explaining transformer behavior, yet their validation has been exclusively in-sample: a graph is tested on the same prompt it was built from. We ask whether circuits derived from attribution graphs generalize to held-out prompts. Using Indirect Object Identification (IOI) in GPT-2 small as a testbed, we run the same five-dimension evaluation protocol (causal direction, cross-prompt stability, paraphrase transfer, cross-task specificity, and quantitative calibration) across three mechanistically distinct intervention regimes. Replacing all 12 MLP layers with their pretrained transcoder approximations produces a 5.6-logit displacement of task signal (D: $+4.15 \to -1.46$), inverting the sufficiency metric and making generalization unmeasurable. Restricting transcoder replacement to the single most-attributed layer (L11) repairs this precondition ($D_{\text{full}} = +4.61$) and reveals a genuine—not confounded—negative result: the consensus feature set transfers to held-out paraphrases with mean sufficiency $\text{Faith}_S = 0.007$, indicating MLP features carry negligible causal weight for IOI, while simultaneously achieving the best attribution-to-effect calibration of any regime we test (Spearman $\rho = 0.483$), because the terminal layer has no downstream layer available to compensate. Switching the target of attribution entirely, to attention heads with exact hook-based ablation, yields strongly contrasting results: a 37-head consensus circuit transfers to held-out paraphrase prompts with mean sufficiency $\text{Faith}_S = 1.196$ (versus 0.049 for a random baseline), confirming that the attributed heads are causally sufficient and generalize across surface-form variation. Together the three regimes separate three previously conflated questions—is the measurement valid, does this component class carry the mechanism, and does attribution calibrate to effect size—and establish a design principle: an intervention framework must preserve task signal as a verifiable precondition before causal generalization can be meaningfully measured, and only once that precondition holds does a negative transfer result become scientifically informative rather than an artifact.

1 Introduction

Mechanistic interpretability seeks to describe the internal computations of neural networks in terms of human-understandable components, with the goal of producing explanations that are not merely descriptive but genuinely predictive [Olah et al., 2020, Elhage et al., 2021]. The *attribution graph* has emerged as the flagship explanation format in this program [Ameisen et al., 2025, Lindsey et al., 2025]: a directed, node-weighted graph in which each node is a model component—a transcoder feature, an attention head, or an embedding—and each edge weight represents the causal contribution

Preprint.

of one component to another's activation, derived from the Input×Gradient (IxG) linear attribution formula.

Attribution graphs have been applied to dozens of behaviors in large language models, generating per-prompt narratives about which components are responsible for a given output. However, a structural gap exists in how these graphs are validated. In every published study, the graph is tested on the same prompt it was constructed from. This tests internal consistency—whether the sum of attributed contributions accounts for the observed output—but not whether the identified mechanism is general. A circuit that explains one specific string of tokens may be a per-prompt overfit rather than a description of the underlying algorithm the model has learned.

Whether attribution circuits are *general*—whether the same components are causally responsible across surface-form variations of the same task—is an open empirical question with significant implications. If circuits generalize, then per-prompt attribution graphs are legitimate mechanistic hypotheses. If they do not, then published circuit narratives are illustrations rather than mechanisms, and downstream conclusions about model behavior drawn from them are unsound.

This paper answers the generalization question directly. We design a controlled evaluation protocol with five orthogonal dimensions and apply it to two instantiations of attribution-based circuit extraction on the Indirect Object Identification (IOI) task in GPT-2 small [Wang et al., 2022]—a setting with a well-characterized ground-truth circuit against which our results can be compared.

Contributions.

1. A five-dimension evaluation protocol for attribution circuit generalization—causal direction, stability, transfer, specificity, and calibration—that is reusable across tasks, models, and attribution methods.
2. A quantitative characterization of compounding transcoder approximation error, showing it can exceed the task signal itself and invalidate sufficiency metrics.
3. A controlled localization experiment showing that restricting transcoder replacement to a single terminal layer repairs the measurement precondition and converts an uninterpretable result into a genuine, well-calibrated negative finding: MLP features account for less than 1% of the causal sufficiency needed to explain IOI transfer ($\text{Faith}_S = 0.007$).
4. Empirical demonstration that hook-based IxG attention head attribution extracts circuits that generalize: the 37-head consensus circuit achieves $\text{Faith}_S = 1.196$ on held-out paraphrase prompts with a 1.146 gap over a random baseline.
5. A design principle for causal interpretability experiments: task-signal preservation must be verified as a precondition before causal metrics are interpreted, and only once verified does a negative transfer result carry scientific meaning.

2 Related Work

IOI and transformer circuits. Wang et al. [2022] hand-discovered the attention-head circuit for IOI in GPT-2 small via activation patching, identifying Name Mover Heads (L9H6, L9H9, L10H0) as the primary causal components. This ground-truth circuit provides an external reference for our study: a correct attribution method should recover these heads.

Attribution graphs. Ameisen et al. [2025] introduce transcoder-based attribution graphs for tracing feature-to-feature causal paths in Claude 3.5 Haiku. Lindsey et al. [2025] apply the method to a wide range of model behaviors. Both papers validate graphs in-sample; our study provides the first systematic out-of-sample evaluation.

Circuit stability and generalization. Hernández et al. [2025] show that cross-prompt stability of edge/head circuits correlates with model generalization, using activation-patching circuits rather than transcoder feature graphs and framing stability as an observable rather than a predictor of held-out causal validity. Our work complements this by providing a direct causal test—measuring held-out ablation effects—and by operating at the feature level in addition to the head level.

Transcoders and sparse autoencoders. Dunefsky et al. [2024] introduce transcoders as interpretable replacements for MLP sublayers, enabling feature-level decomposition of MLP computation. We use the pretrained GPT-2 small transcoders from `pchlenski/gpt2-transcoders` in Study 1.

3 Experimental Framework

3.1 Model, Task, and Performance Metric

We study GPT-2 small (12 layers, 768-dimensional residual stream, 12 attention heads per layer, 124M parameters) run on CPU for hardware-independent reproducibility. All runs use random seed 42.

The task is *Indirect Object Identification* (IOI): given a sentence of the form *"When Mary and John went to the store, John gave the bag to [BLANK]"*, the model should assign higher logit to the indirect object (Mary) than the repeated subject (John). Performance is measured by the *logit difference*

$$D = \text{logit}(\text{IO}) - \text{logit}(\text{S}) \tag{1}$$

at the final token position. For the unmodified model, $D \approx +4.15$ on our source prompts. A cross-task control uses the *Greater-Than* (GT) task [Hanna et al., 2023]: given *"The war lasted from the year 1815 to the year 18[BLANK]"*, the model should favor digits greater than 15.

3.2 Prompt Sets

We use three disjoint sets:

- **Source** ($n = 10$): IOI prompts with varied name pairs and contexts (store, park, library, etc.), used to build circuits.
- **Paraphrase** ($n = 5$): IOI prompts with fresh name pairs and the same grammatical structure, never seen during circuit construction, used to test transfer.
- **Control** ($n = 5$): Greater-Than prompts, used to test task specificity.

3.3 Attribution Method: Input×Gradient

For a model component with output vector $\mathbf{o} \in \mathbb{R}^{d_{\text{model}}}$ and a sensitivity direction $\hat{r} = \partial D / \partial \mathbf{r}_{11}$ (the gradient of D with respect to the post-block-11 residual stream, computed via autograd through the final LayerNorm), the IxG attribution score is

$$c = \mathbf{o} \cdot \hat{r}. \tag{2}$$

A positive c indicates the component's output pushes the model toward predicting IO; negative pushes toward S. We use the autograd gradient rather than the linear approximation $W_U[:, \text{IO}] - W_U[:, \text{S}]$ because it accounts for the final LayerNorm's input-dependent Jacobian.

3.4 Evaluation Metrics

Logit difference D. Defined in Eq. 1. Used as a continuous behavioral readout.

Sufficiency (Faith$_S$). For a frozen component set S,

$$\text{Faith}_S = \frac{D_{\text{full}} - D_S}{D_{\text{full}} - D_{\text{corrupt}}} \tag{3}$$

where D_S is the logit difference after mean-ablating S, D_{full} is the baseline logit difference in the intervention framework, and $D_{\text{corrupt}} = 0$ (model indifferent between IO and S). Faith$_S = 1$ means S is fully causally sufficient for the behavior; Faith$_S = 0$ means S is irrelevant; Faith$_S > 1$ means ablating S inverts the behavior.

Jaccard similarity. $J(A, B) = |A \cap B| / |A \cup B|$, used to measure overlap between circuits extracted from different prompts.

Spearman ρ. Rank correlation between predicted scores c and realized per-component ΔD under individual ablation.

3.5 Research Questions

We evaluate each study along five dimensions:

RQ1 (Direction). Does ablating the top-attributed component move D in the direction its score predicts?

RQ2 (Stability). Are attributed component sets consistent across the 10 source prompts?

RQ3 (Transfer). Does a frozen circuit extracted from source prompts remain causally sufficient on unseen paraphrase prompts?

RQ4 (Specificity). Is the IOI circuit causally inert on the unrelated Greater-Than task?

RQ5 (Calibration). Does the magnitude of attribution scores predict the magnitude of realized ablation effects?

4 Study 1: Transcoder Feature Attribution

4.1 Method

Transcoders. We use the 12 pretrained GPT-2 small transcoders from `pchlenski/gpt2-transcoders` [Dunefsky et al., 2024], each a sparse autoencoder with $d_{\text{sae}} = 24{,}576$ features trained on the post-LayerNorm2 input of its respective MLP layer. All 12 layers are replaced simultaneously during both attribution and ablation passes, keeping attribution and intervention in a consistent activation space.

IxG feature scores. For layer ℓ, feature n with decoder column $W_{\text{dec}}^{(\ell)}[n] \in \mathbb{R}^{d_{\text{model}}}$ and activation f_n at the final token, the score is

$$c_n = (W_{\text{dec}}^{(\ell)}[n] \cdot \hat{r}) \cdot f_n. \tag{4}$$

Ablation. Mean ablation replaces selected features' contributions with their dataset-mean activations:

$$\hat{x}_{\text{ablated}} = \hat{x} + \sum_{n \in S} (\mu_n - f_n)\, W_{\text{dec}}[n], \tag{5}$$

where μ_n is the mean activation of feature n over a 50-sentence neutral corpus at the final token position. Faith_S is computed relative to $D_{\text{full}}^{(\text{TC})}$, the logit difference under full-MLP replacement.

Circuit construction. For each source prompt, the top 20 features by c_n are extracted. The consensus set S comprises features appearing in ≥ 2 of the 10 circuits ($|S| = 45$).

4.2 Results

4.2.1 RQ1 — Causal Direction

The top feature on the first source prompt is (L11, F15690) with $c_n = 0.112$. Ablating it yields $\Delta D = +0.0086$, matching the predicted sign. The realized effect is 7.7% of c_n, a first indication of attenuation within the all-MLP transcoder regime (Section 4.3).

4.2.2 RQ2 — Cross-Prompt Stability

The structure is bimodal (Table 1): feature (L11, F15690) is the top-attributed feature on 8 of 10 prompts (top-1 Jaccard $= 1.0$ on the majority of pairs), while the surrounding top-20 set overlaps only $\approx 25\%$ across prompts. The method isolates one highly stable feature embedded in a diffuse, prompt-specific tail.

Table 1: Stability of top-20 transcoder feature circuits across 10 IOI source prompts.

Quantity	Value
Top-20 Jaccard (median)	0.250
Top-20 Jaccard (mean)	0.296
Top-20 Jaccard (min / max)	0.111 / 0.739
Top-1 Jaccard (median)	**1.000**
Consensus set $\|S\|$	45 features

Table 2: Transcoder transfer results on 5 paraphrase prompts. $D_{\text{full}}^{(\text{TC})}$ is negative on all paraphrases due to approximation error, invalidating Faith_S as a metric.

Prompt	$D_{\text{full}}^{(\text{TC})}$	Faith_S(attr)	Faith_S(rand)
Grace / Mike	-1.14	-0.065	0.000
Lily / Dan	-1.92	$+0.002$	0.000
Eva / Leo	-0.46	-0.127	0.000
Nina / Carl	-1.32	-0.001	0.000
Vera / Jack	-2.69	$+0.018$	0.000
Mean	-1.51	$\mathbf{-0.035}$	$\mathbf{0.000}$

4.2.3 RQ3 — Paraphrase Transfer

The transcoder baseline $D_{\text{full}}^{(\text{TC})}$ is negative on all five paraphrase prompts (Table 2). Because $D_{\text{full}} < D_{\text{corrupt}} = 0$, the denominator of Eq. 3 is negative and Faith_S no longer carries its intended meaning. Transfer cannot be assessed in this regime.

4.2.4 RQ4 — Cross-Task Specificity

Table 3: Causal effect of the IOI transcoder circuit on Greater-Than prompts. Mean $\text{Faith}_S \approx 0$ indicates task-specific selection.

GT prompt (years)	$D_{\text{full}}^{(\text{TC})}$	Faith_S
$1815 \to 18$	-0.280	-0.003
$1823 \to 18$	-0.256	$+0.081$
$1847 \to 18$	$+0.254$	$+0.006$
$1862 \to 18$	-0.147	-0.024
$1878 \to 18$	$+0.173$	$+0.180$
Mean	-0.051	$\mathbf{+0.048}$

Cross-task $\text{Faith}_S = 0.048 \approx 0$ (Table 3). Despite the metric degradation on IOI, the selected features are tied to IOI structure rather than globally high-activation directions that would perturb any task.

4.2.5 RQ5 — Quantitative Calibration

Ablating each of the 45 consensus features individually across all 10 source prompts ($n = 450$ pairs) and correlating predicted c_n with realized ΔD yields Spearman $\rho = -0.018$ ($p = 0.70$). There is no monotonic relationship between attribution magnitude and realized ablation magnitude, even in-sample.

4.3 Root Cause: Compounding Approximation Error

Each transcoder is trained independently on activations from the *original* model. When all 12 are chained, the residual stream drifts from the distribution each subsequent transcoder was trained on. This distributional shift compounds across depth: a small approximation error at layer 0 moves the

Table 4: Effect of transcoder replacement on the IOI logit difference.

Configuration	D	ΔD from original
Original GPT-2 small	$+4.15$	—
All 12 MLPs $\rightarrow$ TCs	-1.46	-5.61

layer-1 input off-manifold, amplifying layer-1 error, and so on. The result is a 5.6-logit displacement (Table 4)—larger than the 4.15-logit IOI signal.

With $D_{\text{full}}^{(\text{TC})} < 0 < D_{\text{corrupt}}$, the denominator of Eq. 3 is negative:

$$\text{Faith}_S = \frac{D_{\text{full}} - D_S}{D_{\text{full}} - 0}, \quad D_{\text{full}} < 0, \tag{6}$$

so any ablation nudging D toward zero registers as negative Faith_S, inverting the intended reading. The same mechanism attenuates and scrambles per-feature effects, explaining the 7.7% realization rate in RQ1 and the negligible ρ in RQ5.

5 Study 1b: Localized (Single-Layer) Transcoder Replacement

Study 1 cannot distinguish two very different explanations for its null result: MLP features might genuinely carry little causal weight for IOI, or the measurement itself might be broken. Section 4.3 shows the second explanation dominates. Rather than abandon transcoder attribution, we repair the precondition directly: restrict replacement to the single transcoder at layer 11—the layer containing the consistently top-attributed feature identified in Study 1 (Section 4)—and leave the remaining 11 MLP sublayers unmodified. This bounds approximation error to one layer instead of compounding it across twelve, and lets us ask the transfer question on a measurement that is verifiably valid.

5.1 Method

All aspects of Study 1's protocol are unchanged (same prompts, same IxG formula Eq. 4, same mean-ablation rule Eq. 5, same consensus construction) with one modification: only layer 11's MLP is replaced by its transcoder. Layers 0–10 run as the original, unmodified MLP. Attribution and ablation remain in a single consistent activation space, now restricted to layer 11's $24{,}576$ features.

5.2 Results

5.2.1 RQ1 — Causal Direction and Precondition Check

The transcoder baseline is $D_{\text{full}}^{(\text{L11})} = +4.61$, close to the original model's $D = +4.15$ and, critically, *positive*—the precondition that failed in Study 1 is restored. The top feature, again (L11, F15690), has $c_n = 0.063$. Ablating it yields $\Delta D = +0.0214$ in the predicted direction: a 33.7% realization rate, roughly $4\times$ higher than the 7.7% observed under full-MLP replacement, though still incomplete—even single-layer transcoder reconstruction is not a perfect approximation of the true MLP.

5.2.2 RQ2 — Cross-Prompt Stability

Table 5: Stability of top-20 layer-11-only feature circuits across 10 IOI source prompts.

Quantity	Value
Top-20 Jaccard (median)	**0.667**
Top-20 Jaccard (mean)	0.715
Top-20 Jaccard (min / max)	0.429 / 0.905
Top-1 Jaccard (median)	1.000
Consensus set $\lvert S\rvert$	25 features

Restricting attribution to a single layer sharply increases stability (Table 5): top-20 Jaccard rises from 0.250 (all-layer) to 0.667, and the consensus set shrinks from 45 to 25 features, all located in layer 11.

Confining the search space to one layer's $24{,}576$ features, rather than pooling noisy signal across twelve layers as in Study 1, produces a visibly purer circuit.

5.2.3 RQ3 — Paraphrase Transfer

Table 6: Layer-11-only transcoder transfer results on 5 paraphrase prompts. $D_{\text{full}}^{(\text{L11})} > 0$ on the source prompt (RQ1) confirms the precondition holds, so Faith_S is meaningful here.

Prompt	Faith_S(attr)	Faith_S(rand)
Grace / Mike	0.0034	0.000
Lily / Dan	0.0106	0.000
Eva / Leo	0.0039	0.000
Nina / Carl	0.0161	0.000
Vera / Jack	0.0024	0.000
Mean	**0.0073**	**0.000**

With a valid measurement in hand, the transfer result is now interpretable, and it is a genuine negative: mean $\text{Faith}_S = 0.007$ (Table 6), a small but non-zero gap over the random baseline. Ablating the 25-feature consensus set removes less than 1% of the behavior needed to explain IOI on held-out paraphrases. Unlike Study 1's -0.035, this number is not an artifact of a broken denominator—D_{full} is positive throughout—so it can be read at face value: layer-11 MLP features are directionally correct and reproducibly attributed, but causally minor contributors to IOI relative to the mechanism as a whole.

5.2.4 RQ4 — Cross-Task Specificity

Cross-task $\text{Faith}_S = 0.011$ on Greater-Than prompts (individual values 0.003–0.023 across the 5 GT prompts), near zero and smaller in magnitude than the already-small IOI transfer effect. The layer-11 feature set is task-specific but, consistent with RQ3, simply does not carry much causal weight for either task.

5.2.5 RQ5 — Quantitative Calibration

Ablating each of the 25 consensus features individually across all 10 source prompts ($n = 250$ pairs) yields Spearman $\rho = 0.483$ ($p = 5.2 \times 10^{-16}$)—the strongest, most significant calibration result of any regime in this paper, substantially exceeding both the all-layer transcoder result ($\rho = -0.018$) and the attention-head result reported in Section 6 ($\rho = 0.167$).

This is not a coincidence of layer choice. Layer 11 is GPT-2 small's *terminal* MLP sublayer: nothing downstream can nonlinearly recompute or compensate for an ablated layer-11 feature except the fixed final LayerNorm and unembedding, both of which are held constant across the comparison. This structurally rules out the multi-layer compensation mechanism identified in Section 6 (RQ5) for attention heads, where early-layer Name Mover Heads are partially reconstructed by later layers after ablation. With no downstream layer available to compensate, the first-order IxG approximation is much closer to exact, and c_n tracks realized ΔD far more reliably.

6 Study 2: Attention Head Attribution

Study 1 establishes that a valid causal evaluation requires the intervention framework to preserve task signal. Study 2 addresses this directly. We switch from transcoder-based MLP replacement to exact hook-based intervention on attention heads, eliminating approximation error by construction. IOI is known to be primarily attention-driven [Wang et al., 2022], making attention heads the mechanistically appropriate target.

6.1 Method

Activation caching. We run the unmodified model via TransformerLens [Nanda & Bloom, 2022] with `use_attn_result=True`, which computes per-head outputs $\mathbf{r}^{(\ell,h)} \in \mathbb{R}^{d_{\text{model}}}$ before they are

summed. The activation cache stores value vectors, attention patterns, and per-head outputs at all layers.

IxG head scores. For layer ℓ, head h, destination position d (final token), the score sums contributions from all source positions s (excluding BOS):

$$c_h = \sum_{s \geq 1} A_{d,s}^{(\ell,h)} \left(V_s^{(\ell,h)} W_O^{(\ell,h)} \right) \cdot \hat{r}, \tag{7}$$

where $A_{d,s}^{(\ell,h)}$ is the attention pattern weight, $V_s^{(\ell,h)}$ is the value vector at source s for head h at layer ℓ, and $W_O^{(\ell,h)}$ projects from d_{head} to d_{model} space. The score c_h is the total IxG contribution of head h at the final position toward the logit-difference direction $\hat{r}$.

Exact hook-based ablation. Mean ablation replaces the per-head output at the *final position only* with the head's mean output over a 50-sentence neutral corpus:

$$\mathbf{r}_{-1}^{(\ell,h)} \leftarrow \bar{\mathbf{r}}^{(\ell,h)}, \tag{8}$$

where $\bar{\mathbf{r}}^{(\ell,h)} \in \mathbb{R}^{d_{\text{model}}}$ is the mean of $\mathbf{r}_{-1}^{(\ell,h)}$ over neutral sentences. This is exact: no computation is replaced, only one position's activation is substituted. Consequently $D_{\text{full}} = +4.15$, identical to the original model, with zero approximation error.

We ablate only the final position because the causal question is specifically about the head's contribution to the final-token prediction. Ablating all positions would propagate residual-stream perturbations through subsequent attention layers, confounding the single-head causal estimate.

Circuit construction. For each source prompt, the top 20 heads by c_h are extracted. The consensus set S comprises heads appearing in ≥ 2 of the 10 circuits ($|S| = 37$).

6.2 Results

6.2.1 RQ1 — Causal Direction

The top attributed head on the first source prompt is (L9, H6) with $c_h = 2.08$—one of the three Name Mover Heads identified by Wang et al. [2022]. Individual ablation of single early-layer heads exhibits nonlinear compensation: layers 10–11 partially reconstruct the signal when only one upstream head is removed, so the single-head causal direction test is uninformative. We therefore assess causal direction on the full top-10 set, which is the quantity Faith_S measures: ablating the top-10 heads jointly drops D from $+4.15$ to $+0.72$ ($\Delta D = 3.43$ logits), unambiguously in the predicted direction.

6.2.2 RQ2 — Cross-Prompt Stability

Table 7: Stability of top-20 attention head circuits across 10 IOI source prompts.

Quantity	Value
Top-20 Jaccard (median)	**0.379**
Top-20 Jaccard (mean)	0.385
Top-20 Jaccard (min / max)	0.250 / 0.667
Top-1 Jaccard (median)	0.000
Consensus set $\lvert S \rvert$	37 heads

The top-20 Jaccard median is 0.379 (Table 7), well above the transcoder figure of 0.250 and above a minimum meaningful overlap threshold. The top-1 Jaccard is 0, reflecting that the top head alternates between L9H6 and L9H9—two near-equivalent Name Movers whose relative dominance varies with name embeddings—but the full top-20 sets are substantially shared. The 37-head consensus set includes the three Wang et al. Name Mover Heads and a range of supporting heads across layers 6–11.

Table 8: Attention head transfer results on 5 paraphrase prompts. $\text{Faith}_S > 1.0$ on all prompts indicates the consensus circuit is causally sufficient and more than accounts for the full IOI behavior.

Prompt	D_{full}	Faith_S(attr)	Faith_S(rand)
Grace / Mike	$+3.65$	$\mathbf{1.322}$	0.085
Lily / Dan	$+4.56$	$\mathbf{1.195}$	-0.098
Eva / Leo	$+4.24$	$\mathbf{1.020}$	0.316
Nina / Carl	$+4.42$	$\mathbf{1.009}$	-0.069
Vera / Jack	$+6.84$	$\mathbf{1.433}$	0.012
Mean	$+4.74$	$\mathbf{1.196}$	$\mathbf{0.049}$

6.2.3 RQ3 — Paraphrase Transfer

$\text{Faith}_S > 1.0$ on all five paraphrase prompts (Table 8). The mean is 1.196, indicating that ablating the consensus set not only eliminates the IOI preference but inverts it: after ablation $D < 0$ on all prompts (the model prefers S over IO). This super-sufficiency confirms that the 37-head set captures the full IOI mechanism plus enough supporting circuitry that the behavior collapses when they are removed. The mean attribution gap over random is $1.196 - 0.049 = 1.146$, demonstrating that the attribution-guided selection is essential.

6.2.4 RQ4 — Cross-Task Specificity

Table 9: Causal effect of the IOI attention head circuit on Greater-Than prompts. Mean $\text{Faith}_S = 0.313$ indicates partial cross-task overlap.

GT prompt (years)	D_{full}	Faith_S
$1815 \to 18$	$+5.32$	0.151
$1823 \to 18$	$+1.53$	0.706
$1847 \to 18$	$+5.59$	0.213
$1862 \to 18$	$+0.60$	0.237
$1878 \to 18$	$+4.44$	0.257
Mean	$+3.50$	$\mathbf{0.313}$

Cross-task $\text{Faith}_S = 0.313$ (Table 9), above zero but below the IOI transfer figure of 1.196. The 37-head consensus set is not purely IOI-specific: it contains early-layer heads (L1–L2) that participate in general sequence processing tasks and are also engaged by GT. The elevated cross-task effect is concentrated in one prompt ($1823 \to 18$, $\text{Faith}_S = 0.706$), suggesting sensitivity to specific GT prompt structures. A smaller consensus set built with a stricter min-hits threshold would likely reduce this cross-task leakage.

6.2.5 RQ5 — Quantitative Calibration

Ablating each of the 37 consensus heads individually across all 10 source prompts ($n = 370$ pairs) yields Spearman $\rho = 0.167$ ($p = 0.0013$). The correlation is statistically significant but modest. The weak calibration reflects the nonlinear compensation identified in RQ1: early-layer Name Mover Heads with large positive c_h show negative ΔD under individual ablation because layers 10–11 compensate, violating the linear independence assumption underlying IxG. Later-layer heads (L10–L11) show better individual calibration because fewer downstream compensating paths remain. The correlation is positive and significant, indicating that c_h carries real rank information despite imperfect magnitude prediction.

7 Comparative Analysis

Table 10 displays the full comparison. Several patterns stand out.

Table 10: Side-by-side comparison across all three regimes and the five research questions. †: all-layer TC transfer result reflects metric invalidation ($D_{\text{full}} < 0$), not a genuine generalization failure; the L11 and attention results are both measured under a verified-valid precondition ($D_{\text{full}} > 0$).

RQ	Dimension	Metric	TC all-layer	TC L11	Attn (Study 2)
		D_{full} vs. orig. D=4.15	-1.46	$+4.61$	$+4.15$
1	Direction	realization ratio	7.7%	33.7%	— (set-level, see §6)
2	Stability	top-20 Jaccard (median)	0.250	**0.667**	0.379
		top-1 Jaccard (median)	1.000	**1.000**	0.000
		consensus size $\lvert S\rvert$	45	25	37
3	Transfer	Faith$_S$ (mean)	$-0.035^{\dagger}$	0.007	**1.196**
		gap over random	$-0.035^{\dagger}$	0.007	**1.146**
4	Specificity	cross-task Faith$_S$	0.048	**0.011**	0.313
5	Calibration	Spearman ρ	-0.018	**0.483**	0.167
		p-value	0.70	$\mathbf{5 \times 10^{-16}}$	0.001

Task-signal preservation is the gating factor, and it is separable from whether the component class matters. The all-layer transcoder framework shifts D by -5.61 logits before any ablation is run; the single-layer transcoder framework shifts it by only $+0.46$ logits (and in the *positive* direction); the hook-based attention framework introduces zero displacement. Only the latter two produce an interpretable Faith$_S$. Comparing them directly isolates a question Study 1 alone could not answer: given a valid measurement, *does this component class carry the mechanism?* The answer is no for layer-11 MLP features (Faith$_S = 0.007$) and yes for attention heads (Faith$_S = 1.196$)—a $170\times$ difference in mean sufficiency between two measurements that are both methodologically sound. This is only visible because Study 1b repairs the precondition without changing the attributed component class, holding one variable fixed while Study 2 changes the other.

Attention circuits are stable and transferable; MLP features are stable but causally minor. With valid measurement frameworks in place, the two regimes diverge sharply on transfer despite both being methodologically sound. Attention heads exhibit Faith$_S > 1.0$ on all five held-out paraphrases—the circuit not only destroys the IOI preference but inverts it. Layer-11 MLP features, despite being highly stable across prompts (top-20 Jaccard 0.667, the highest of any regime) and correctly signed (RQ1), account for less than 1% of the same behavior. Stability and sufficiency are therefore dissociable properties: a feature can be reliably and reproducibly attributed without being causally important.

Calibration is best where compensation is structurally impossible. The ordering $\rho_{\text{TC-L11}}(0.483) \gg \rho_{\text{Attn}}(0.167) \gg \rho_{\text{TC-all}}(-0.018)$ is explained by a single structural fact: layer 11 is GPT-2 small's terminal MLP sublayer, so no downstream computation can compensate for an ablated layer-11 feature. Attention heads at layers 6–9 are compensated by heads at layers 10–11 (Section 6, RQ5); all-layer transcoder features are subject to both compensation *and* compounding reconstruction error across depth. Terminal-layer, single-layer attribution is the regime closest to satisfying IxG's implicit assumption of no downstream nonlinear interaction, and it is the regime where the linear approximation performs best.

Specificity does not simply track circuit size. The attention head circuit (37 components) shows higher cross-task Faith$_S$ (0.313) than either TC regime (0.048 all-layer, 0.011 L11)—but the L11 circuit (25 components) is more specific than the 45-component all-layer circuit despite being smaller, and the all-layer circuit's low cross-task effect is measured in the same broken regime that invalidates its transfer result, so it should not be read as a clean specificity signal. The most reliable specificity comparison is L11 (0.011) vs. attention (0.313), both measured under valid preconditions: a larger, mechanistically dominant circuit (attention heads, Faith$_S = 1.196$ on-task) shows more cross-task leakage than a smaller, causally marginal one (L11 features, Faith$_S = 0.007$ on-task). This suggests cross-task Faith$_S$ scales with how much of the actual mechanism a circuit captures, not merely with component count.

8 Discussion

8.1 When Do Attribution Circuits Generalize?

Our central finding is that attribution-based attention head circuits do generalize: the same 37 heads are causally responsible for IOI behavior across 10 training prompts and 5 held-out paraphrase prompts with different names, objects, and locations. This is consistent with the mechanistic hypothesis that these heads implement a generalizable name-position lookup algorithm rather than a lookup table of specific name pairs.

The two transcoder studies together clarify, rather than contradict, this conclusion. Study 1 cannot distinguish "the measurement is broken" from "MLP features don't generalize"—both produce a degenerate Faith$_S$. Study 1b resolves the ambiguity by repairing the measurement while keeping the attributed component class fixed: once D_{full} is restored to a positive, near-original value, the transfer result is a small but genuine positive number (Faith$_S = 0.007$), not an artifact. The attributed feature (L11, F15690) is therefore not a false lead—it is stably and correctly attributed, appears in 8 of 10 circuits, does not affect the unrelated GT task, and its ablation moves D in the predicted direction on every prompt we test. It is simply a minor contributor to a behavior that is overwhelmingly implemented elsewhere in the network. The three regimes together support a clean picture of IOI in GPT-2 small: the mechanism is concentrated in attention heads (Faith$_S = 1.196$), with layer-11 MLP features playing a small, reliable, but non-load-bearing supporting role (Faith$_S = 0.007$)—consistent with Wang et al. [2022]'s original head-level circuit and extending it with a quantitative bound on how little the MLP pathway contributes by comparison.

8.2 The Nonlinearity Problem in Head-Level Attribution

Study 2 reveals a practical limit of IxG attribution: individual head scores (c_h) predict the direction of individual causal effects but not the magnitude, particularly for early-layer heads. This is a consequence of multi-layer compensation: removing one Name Mover Head prompts later layers to partially reconstruct the missing signal. The compensation is strong enough to reverse the sign of single-head ablation effects for L9H6 and L9H9.

Importantly, this does not undermine the transfer result. Faith$_S$ is measured by ablating the *full* consensus set, eliminating the compensation pathway simultaneously with the original mechanism. Set-level causal validity and individual-component calibration are distinct properties, and the data show they can dissociate.

8.3 Design Principles for Causal Interpretability

The three regimes together suggest four design principles for experiments that aim to measure attribution-circuit generalization:

1. **Verify task-signal preservation before interpreting causal metrics.** Report $D_{\text{full}}^{(\text{framework})}$ vs. original D as a precondition check. If the intervention framework itself displaces the task signal, downstream Faith$_S$ values are uninterpretable—as in Study 1, where this check alone would have prevented misreading a broken measurement as a generalization failure.
2. **If full-sublayer replacement is required, localize it.** Study 1b shows that restricting transcoder replacement to a single, well-chosen layer is often sufficient to repair the precondition without abandoning the transcoder framework entirely, at the cost of only characterizing that layer's contribution rather than the full network's.
3. **Use hook-based ablation where possible.** Replacing whole computational sublayers (MLPs via transcoders) introduces distributional shift that compounds with depth. Hook-based substitution at specific activations, as used for attention heads in Study 2, leaves the forward pass intact and avoids this error class entirely.
4. **Evaluate circuits at the set level, not only per-component, and expect calibration to depend on network position.** Individual attribution scores are noisy predictors of individual causal effects when downstream compensation is possible (Study 2, all-layer Study 1). Calibration is best when the attributed component is structurally shielded from compensation, as in Study 1b's terminal-layer features ($\rho = 0.483$). Set-level Faith$_S$

remains the more robust measure of whether the attribution method has identified the right components; per-component calibration should be interpreted relative to the attributed component's position in the network.

9 Conclusion

We asked whether attribution circuits generalize to held-out prompts and answered the question under three methodologically distinct conditions. Transcoder-based MLP feature attribution, applied to all 12 layers simultaneously, is blocked from answering the question at all by a compounding approximation error that inverts the task signal; this is a finding about the measurement framework, not about the attribution method per se, and on its own it is consistent with either a real or an artifactual generalization failure. Localizing transcoder replacement to a single terminal layer repairs the measurement precondition and resolves the ambiguity directly: the resulting transfer result ($\text{Faith}_S = 0.007$) is small but genuine, and the resulting calibration result ($\rho = 0.483$) is the strongest observed in this paper, because a terminal layer has no downstream computation available to compensate for the intervention. Hook-based IxG attention head attribution, which introduces zero approximation error by construction, extracts a 37-head consensus circuit that transfers to held-out IOI paraphrase prompts with mean $\text{Faith}_S = 1.196$—a 1.146 gap over a random baseline—confirming that the attributed heads are causally sufficient and mechanistically general, and that the IOI mechanism in GPT-2 small is concentrated in attention computation rather than in the final MLP layer.

The broader implication is methodological. Causal generalization is a well-defined and measurable property of attribution circuits, but measuring it requires a measurement framework that preserves the behavior being explained, and a negative transfer result is only scientifically informative once that precondition has been verified. Absent verification, a broken measurement and a genuinely non-general circuit are empirically indistinguishable—as they were in Study 1 until Study 1b separated them. We recommend verifying task-signal preservation, preferring hook-based or localized interventions, and evaluating circuits at the set level as standard practice in any causal interpretability evaluation that aims to make out-of-sample claims.

Reproducibility

All code, results, and the LaTeX source for this paper are available at `https://github.com/dzhan111/ioi_tc_pilot`. The full pipeline runs on CPU in approximately 2 hours; transcoder checkpoints download automatically from `pchlenski/gpt2-transcoders` on HuggingFace. Environment: Python 3.11, PyTorch 2.2.0 (CPU), TransformerLens 1.11.0.

- `scripts/00_smoke.py`–`05_calibration.py`: Study 1 (all-layer transcoder) and Study 1b (layer-11-only transcoder) observations for RQ1–RQ5. Both studies use the same scripts; the layer set replaced is controlled by the `tc_layers` parameter in `src/pilot/intervene.py` and `src/pilot/features.py` (default: `[11]`, i.e. Study 1b; Study 1 sets this to all 12 layers).
- `scripts/00_attn_smoke.py`–`05_attn_calibration.py`: Study 2 (attention head) observations for RQ1–RQ5.
- `results/*.json`: raw numeric outputs for Study 1b and Study 2. Study 1 (all-layer) outputs are archived separately, as the current default configuration reproduces Study 1b.
- `src/pilot/`: shared metrics, model loading, prompts.

References

Ameisen, D., Lindsey, J., et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. `https://transformer-circuits.pub/2025/attribution-graphs/methods.html`

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable LLM feature circuits. *arXiv preprint arXiv:2406.11944*, 2024.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Hanna, M., Liu, O., and Variengien, A. How does GPT-2 compute greater-than? Interpreting mathematical abilities in a pre-trained language model. *arXiv preprint arXiv:2305.00586*, 2023.

Hernández, O., et al. Circuit stability characterizes language model generalization. *arXiv preprint arXiv:2505.24731*, 2025.

Lindsey, J., et al. On the biology of a large language model. *Transformer Circuits Thread*, 2025. `https://transformer-circuits.pub/2025/attribution-graphs/biology.html`

Nanda, N. and Bloom, J. TransformerLens. `https://github.com/neelnanda-io/TransformerLens`, 2022.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022.