CONDITIONAL CO-ABLATION: RECOVERING SELF-REPAIR BACKUPS IN TRANSFORMER CIRCUITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic interpretability seeks to explain transformer behavior through *circuits*: sets of internal components that causally support a behavior. However, *self-repair* creates a blind spot: ablating a *primary* component can activate a dormant *backup*, so a circuit that explains behavior in the intact model can become incomplete under the intervention used to test it. We formulate this gap as *conditional circuit completion*: given a primary set, identify components that become causally important after its removal. We introduce *conditional co-ablation* (COAX), which ranks candidates by growth in ablation effect after primary-set removal. We show that a perfectly dormant backup can be indistinguishable from an irrelevant component to per-unit intact-state scores, whereas its conditional effect change exactly aggregates all interaction orders linking it to the removed set. On GPT-2-small's Indirect Object Identification (IOI) circuit, COAX recovers the documented backup heads at 0.941 ROC-AUC, versus 0.665 for the strongest intact-state attribution baseline and 0.758 for the matched conditional-energy control. Recovery drops to 0.40 ± 0.13 AUC for alternative component sets matched in behavioral effect, output displacement, and depth, showing that recovery is specific to the removed circuit. Beyond recovery, the COAX-selected heads are causally load-bearing: freezing them after primary removal sharply reduces the IOI margin, while adding them to the incomplete circuit reduces incompleteness from 0.47 to 0.24. More broadly, conditional growth aligns with intervention-derived repair in $11/12$ held-out instances across 4 mechanism clusters, and COAX completions outperform matched random completions on all 8 non-GPT-2 models spanning 6 architecture families. Therefore, causal explanations of self-repairing transformers must account for backup circuitry when primary components fail.

1 INTRODUCTION

Mechanistic interpretability seeks to explain transformer behavior through internal mechanisms, enabling transparent and auditable accounts of model behavior (Bereska & Gavves, 2024). A central object in such explanations is a *circuit*: a set of internal components that causally contribute to a model's behavior (Olah et al., 2020; Elhage et al., 2021). To identify such circuits, prior work commonly intervenes on individual components, most directly through *ablation*, and measures how the model's output changes. Intervention and attribution methods use such effects to rank components or edges and assemble circuits (Wang et al., 2023; Conmy et al., 2023; Syed et al., 2024; Hanna et al., 2024; Kramár et al., 2024; Marks et al., 2025). These methods have been effective at identifying the circuitry that supports behavior in the intact, unablated model.

However, *self-repair* in transformers challenges this picture: ablating a circuit component can change which circuit carries the behavior. In the Hydra effect (McGrath et al., 2023; Rushing & Nanda, 2024), ablating a *primary* component can activate a dormant *backup* that takes over its role. Standard intact-state attribution can then underestimate both components: the primary because its loss is compensated, and the backup because it has not yet acted in the intact model.

The Indirect Object Identification (IOI) circuit in GPT-2-small provides a concrete example. Given "Mary and John went to the store; John gave a drink to __," three *primary* name-mover heads help write the correct name, "Mary," to the output (Wang et al., 2023). Ablating these three heads changes the correct-versus-competing-name logit margin by only 0.28 from its intact value of 2.57, because

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

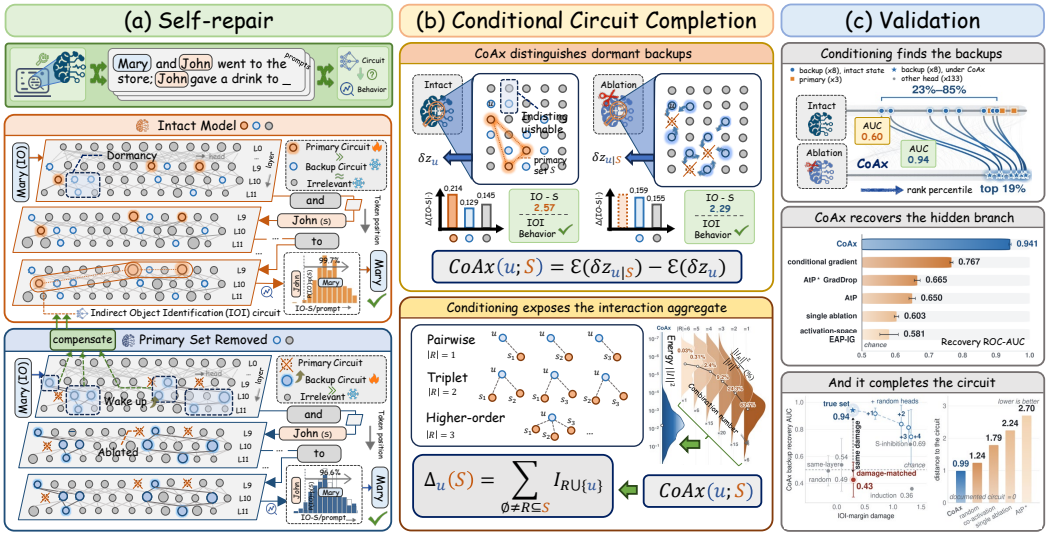

Figure 1: **From self-repair to conditional circuit completion. (a)** Ablating the primary IOI circuit recruits a dormant backup branch while largely preserving behavior. **(b)** CoAx compares each candidate's ablation effect with the primary set present and removed; this contrast distinguishes dormant backups and exposes the all-order effect-change aggregate linking each candidate to the removed set. **(c)** On GPT-2-small IOI, conditioning exposes the documented backups, improves their recovery over intact-state baselines, and substantially completes the missing backup branch.

eight documented *backup* name-movers compensate for their removal. Yet ablating the primaries together with backups causes $1.9\times$ as much damage as the sum of their individual effects. Thus, the backup branch is a causally important part of how the model sustains the behavior under ablation.

These observations expose a gap in circuit explanation: a circuit can faithfully explain the intact model yet become incomplete under the intervention used to test it. This raises a *conditional question*: which components become causally important only after part of the circuit is removed?

We formalize this problem as *conditional circuit completion*. Given a primary set $\mathcal{S}$ supplied by the analyst, the goal is to identify components whose causal importance emerges after $\mathcal{S}$ is removed. Our method, *conditional co-ablation* (CoAx), compares each component's ablation effect with $\mathcal{S}$ present and removed, and ranks candidates by how much that effect grows. A dormant backup therefore has little effect while $\mathcal{S}$ is present but a large effect after $\mathcal{S}$ is removed, whereas an irrelevant component has little effect in either condition.

This conditional comparison also has a precise theoretical interpretation. Per-unit information measured only in the intact model can be insufficient to distinguish a perfectly dormant backup from an irrelevant component (Proposition 1). By contrast, the change in a component's ablation effect when $\mathcal{S}$ is removed exactly aggregates all interaction orders linking that component to $\mathcal{S}$ (Proposition 2). Thus, conditioning on the primary set exposes the interactions through which dormant backups become causally important.

On the documented GPT-2-small IOI circuit, CoAx recovers the documented backup branch at 0.941 ROC-AUC, outperforming both the strongest intact-state attribution baseline (0.665) and the matched conditional-energy control (0.758). Replacing the primary set with matched alternative component sets drops backup recovery to about 0.40 AUC, confirming that recovery is specific to the removed primary set. Beyond recovery, freezing the recovered heads at their intact-state outputs after primary removal causes a 0.90 IOI-margin loss, versus 0.03 for conditional energy and 0.27 for size-matched random heads, showing that they are causally load-bearing. At the circuit level, the same raw CoAx ranking reduces circuit incompleteness from 0.47 to 0.24. Beyond IOI, conditional growth remains positively aligned with intervention-derived repair in $11/12$ held-out instances across four mechanism clusters, and CoAx-based circuit completion outperforms size-matched random completion on all eight non-GPT-2 models spanning six architecture families. Taken together, these results show that CoAx recovers primary-set-specific backups, identifies components that causally carry repair, and extends conditional circuit completion beyond a single annotated circuit.

Contributions. (i) Problem and method: we formulate *conditional circuit completion* and introduce CoAx, which ranks components by growth in ablation effect after a primary set is removed (Section 2.1). **(ii) Theory**: we show that per-unit intact-state information can fail to distinguish dormant backups from irrelevant components, and characterize the effect change after removal as aggregated interactions with the removed set (Section 2.2). **(iii) Evidence**: on IOI, CoAx recovers the documented backup branch, establishes recovery specificity to the removed primary set, verifies its causal role in self-repair, and substantially closes the circuit's incompleteness gap (Sections 3–4). **(iv) Generalization**: beyond the annotated IOI backups, we validate conditional growth against intervention-derived repair across four mechanism clusters and show conditional circuit completion on eight non-GPT-2 models spanning six architecture families (Section 5).

2 Conditional Co-Ablation: Importance Under Intervention

We consider frozen decoder-only transformers, where a *structured unit* is any component with an additive residual-stream contribution that can be zeroed by ablation (Elhage et al., 2021). Let $\mathcal{S}$ be the supplied primary set and $\mathcal{U}$ the candidate structured units for conditional completion, $\mathcal{U} \cap \mathcal{S} = \emptyset$; in our main experiments, these units are attention heads, matching the granularity of the documented IOI primary and backup heads (Wang et al., 2023). For an input–position pair (x, t), ablating a set M of units yields vocabulary logits $z_M(x, t)$; we suppress (x, t) when unambiguous. We write $z_\emptyset \equiv z_0$ for the clean logits and $p_0 = \mathrm{softmax}(z_0)$ for the corresponding clean output distribution.

2.1 The conditional co-ablation score

For a candidate $u \in \mathcal{U}$, ablating u on the intact model gives the effect $\delta z_u = z_0 - z_{\{u\}}$, the change from the clean logits z_0 to the logits after ablating u. With the primary set $\mathcal{S}$ removed, $z_\mathcal{S}$ becomes the reference logits, and ablating u in addition gives

$$\delta z_{u|\mathcal{S}} = z_\mathcal{S} - z_{\mathcal{S}\cup\{u\}}. \tag{1}$$

Thus, δz_u measures u with $\mathcal{S}$ present, whereas $\delta z_{u|\mathcal{S}}$ measures it with $\mathcal{S}$ removed. A dormant backup changes the output little while $\mathcal{S}$ is present but much more after $\mathcal{S}$ is removed, whereas an irrelevant unit has little effect in either condition. Let $\mathcal{E}(\cdot)$ denote the label-free measure of ablation-effect magnitude defined in Eq. 3. We define the *conditional growth* of this magnitude as:

Definition 1 (Conditional co-ablation score)**.**

$$\mathrm{CoAx}(u; \mathcal{S}) = \mathcal{E}\big(\delta z_{u|\mathcal{S}}\big) - \mathcal{E}\big(\delta z_u\big). \tag{2}$$

Thus, $\mathrm{CoAx}(u; \mathcal{S})$ measures how much the magnitude of u's ablation effect grows when the primary set is removed. The reference logits z_0 and $z_\mathcal{S}$ are shared across candidates; scoring all $|\mathcal{U}|$ candidates requires two shared reference evaluations and two candidate-specific evaluations per candidate, hence $O(|\mathcal{U}|)$ forward evaluations. No backward pass or task gradient is required.

If removing $\mathcal{S}$ leaves u's effect unchanged, then $\delta z_{u|\mathcal{S}} = \delta z_u$ and $\mathrm{CoAx}(u; \mathcal{S}) = 0$, even if u is important in both conditions. Taking the difference subtracts each candidate's intact-state magnitude, isolating the change in ablation-effect magnitude associated specifically with primary-set removal.

Because the score is signed, $\mathrm{CoAx}(u; \mathcal{S}) > 0$ indicates that removing $\mathcal{S}$ increases the magnitude of u's ablation effect, while $\mathrm{CoAx}(u; \mathcal{S}) < 0$ indicates a decrease. CoAx ranks positive growth, the pattern expected for a backup that becomes important after primary-set removal.

The output measure $\mathcal{E}$. We measure logit perturbations in the Fisher geometry induced by the clean output distribution (Amari, 1998), so the magnitude reflects local output-distribution sensitivity rather than raw Euclidean logit distance. For any vocabulary-logit perturbation $v = v(x, t)$, define

$$\widetilde{v} = \sqrt{p_0} \odot \big(v - \langle p_0, v\rangle \mathbf{1}\big), \qquad \mathcal{E}(v) = \mathbb{E}_{x,t}\left[\|\widetilde{v}\|_2^2\right]. \tag{3}$$

Here $\odot$ denotes elementwise multiplication, $\mathbf{1}$ is the all-ones vector, and $\sqrt{p_0}$ is taken elementwise; the expectation is over the evaluated input–position pairs. Centering by $\langle p_0, v\rangle$ removes the shared logit shift to which softmax is invariant, while the $\sqrt{p_0}$ weighting yields $\langle \widetilde{v}, \widetilde{w}\rangle = v^\top F w$ for perturbations v and w, where $F = \mathrm{diag}(p_0) - p_0 p_0^\top$ is the categorical Fisher matrix (Appendix A.1). Both terms in Eq. 2 use the same clean reference geometry, so their difference reflects a change in ablation-effect magnitude rather than a change in metric.

2.2 What Conditioning Reveals and Intact-State Scores Miss

Two facts motivate conditioning. First, per-unit information from the intact model can be insufficient to distinguish a perfectly dormant backup from an irrelevant unit, motivating measurement after the primary set $\mathcal{S}$ is removed (Proposition 1). Second, the resulting effect change $\delta z_{u|\mathcal{S}} - \delta z_u$ exactly decomposes into interactions linking the candidate u to $\mathcal{S}$ (Proposition 2). Proofs are in Appendix A.2.

Definition 2 (Intact-state score)**.** We call a per-unit score *intact-state* if it is determined only by the candidate's per-unit information measured on the intact model without removing $\mathcal{S}$. Here, this information consists of its forward activation, output derivative, and single-ablation effect δz_u. Single-ablation saliency and standard local attribution scores are examples of this class.

Proposition 1 (Non-identifiability from intact-state information)**.** *There exist a model, an input distribution, a primary set $\mathcal{S}$, and two candidates $b, n \in \mathcal{U}$ with identical intact-state information, yet $\mathcal{E}(\delta z_b) = \mathcal{E}(\delta z_n) = 0$. After $\mathcal{S}$ is removed, $\mathcal{E}(\delta z_{b|\mathcal{S}}) > 0$ while $\mathcal{E}(\delta z_{n|\mathcal{S}}) = 0$. Every intact-state score assigns b and n the same value, whereas $\mathrm{CoAx}(b; \mathcal{S}) > 0$ while $\mathrm{CoAx}(n; \mathcal{S}) = 0$.*

Proposition 1 formalizes this blind spot: a perfectly dormant backup can be indistinguishable from an irrelevant unit on the intact model, yet become important after primary-set removal.

To characterize this change exactly, we extend the single-unit effect to any ablated set M by defining $\delta z_M := z_0 - z_M$. Then $\delta z_\emptyset = 0$, $\delta z_{\{u\}} = \delta z_u$, and $\delta z_{u|\mathcal{S}} = \delta z_{\mathcal{S}\cup\{u\}} - \delta z_{\mathcal{S}} = z_{\mathcal{S}} - z_{\mathcal{S}\cup\{u\}}$. Thus, $M \mapsto \delta z_M$ is a vector-valued set function. Using the standard Möbius transform (Rota, 1964), associate each nonempty subset $T \subseteq M$ with the interaction $I_T := \sum_{A\subseteq T}(-1)^{|T|-|A|}\delta z_A$, where A ranges over the subsets of T. Möbius inversion then decomposes the joint ablation effect of M as

$$\delta z_M = \sum_{\emptyset \neq T \subseteq M} I_T. \tag{4}$$

The interaction order is $|T|$: $I_{\{u\}} = \delta z_u$ is the first-order effect visible from ablating u alone. For two units, $I_{\{u_1,u_2\}} = \delta z_{\{u_1,u_2\}} - \delta z_{u_1} - \delta z_{u_2}$ is their second-order non-additivity—the joint effect beyond the single-unit effects; higher orders extend this to larger sets. We now specialize this decomposition to the interactions linking a candidate u to the removed primary set $\mathcal{S}$.

Proposition 2 (Exact all-order interaction decomposition)**.** *For any candidate $u \in \mathcal{U}$,*

$$\delta z_{u|\mathcal{S}} = \sum_{R\subseteq\mathcal{S}} I_{R\cup\{u\}} = \delta z_u + \sum_{\{s\}\subseteq\mathcal{S}} I_{\{u,s\}} + \sum_{\{s,s'\}\subseteq\mathcal{S}} I_{\{u,s,s'\}} + \cdots. \tag{5}$$

Hence the effect change $\Delta_u(\mathcal{S}) := \delta z_{u|\mathcal{S}} - \delta z_u = \sum_{\emptyset\neq R\subseteq\mathcal{S}} I_{R\cup\{u\}}$ *is exactly the aggregate of all interaction orders linking u to the removed primary set $\mathcal{S}$ and supported on $\mathcal{S}\cup\{u\}$.*

Proposition 2 shows that the change in u's ablation effect under primary-set removal is exactly the aggregate of all interaction orders linking u to $\mathcal{S}$ and supported on $\mathcal{S}\cup\{u\}$; conditioning yields this aggregate without enumerating its terms. This effect change also gives a direct decomposition of the CoAx score. Since $\delta z_{u|\mathcal{S}} = \delta z_u + \Delta_u(\mathcal{S})$, Eq. 2 becomes

$$\mathrm{CoAx}(u; \mathcal{S}) = \mathbb{E}_{x,t}\left[2\left\langle \widetilde{\delta z}_u, \widetilde{\Delta}_u(\mathcal{S})\right\rangle + \left\|\widetilde{\Delta}_u(\mathcal{S})\right\|_2^2\right]. \tag{6}$$

Here $\widetilde{\Delta}_u(\mathcal{S})$ applies the transformation in Eq. 3 to $\Delta_u(\mathcal{S})$. For a perfectly dormant unit, $\mathcal{E}(\delta z_u) = 0$ makes the intact-state term vanish, so Eq. 6 reduces exactly to $\mathcal{E}(\Delta_u(\mathcal{S}))$.

Explicitly enumerating all interaction orders is combinatorial in $|\mathcal{S}|$. Even a pairwise truncation requires $O(|\mathcal{U}||\mathcal{S}|)$ evaluations and omits higher orders, whereas conditioning yields the full interaction aggregate with the $O(|\mathcal{U}|)$ cost of Section 2.1.

Why signed growth targets backup recruitment. The conditional effect change $\Delta_u(\mathcal{S})$ reveals whether removing $\mathcal{S}$ changes u's role, while backup discovery asks the more specific question of whether u becomes more important after the primary set is removed. Equation 6 makes this recruitment criterion explicit: $\mathrm{CoAx}(u; \mathcal{S}) > 0$ means that the magnitude of u's ablation effect grows after primary-set failure, whereas $\mathrm{CoAx}(u; \mathcal{S}) < 0$ means that it shrinks. For a perfectly dormant

backup, the intact-state term vanishes and the score reduces exactly to $\mathcal{E}(\Delta_u(\mathcal{S}))$, the newly exposed conditional effect. We therefore rank candidates directly by positive signed growth.

Together, these results establish why conditional measurement is necessary and what it reveals: dormant backups can be invisible in the intact model, while primary-set removal exposes the all-order interaction aggregate underlying their effect change. The experiments below first test the signed recruitment signal on a documented dormant backup circuit (Sections 3–4), then ask whether the same conditional measurement tracks repair beyond that annotated setting (Section 5).

3 Recovering the Dormant Backups of a Known Circuit

Primary-circuit failure exposes the dormant branch hidden from intact-state scoring. We test whether CoAx recovers this branch (Section 3.1), whether recovery is specific to the removed primary set (Section 3.2), and whether the recovered heads causally carry repair (Section 3.3).

Protocol. We use the well-characterized GPT-2-small IOI circuit (Radford et al., 2019; Wang et al., 2023), where both the primary path and its backup branch are independently documented: the three primary name-mover heads form $\mathcal{S}$, the eight documented backup name-movers are positives, and all other heads are candidates. The backup labels are used only for evaluation; Appendix C.1.7 reconstructs a primary-removal-free reference set from intact-state signals, confirming that recovery is not tied to the original labeling route. Full experimental setup is in Appendix B.

Table 1: **Backup-head discovery on GPT-2-small IOI.** Head-level ROC-AUC over eight documented backups (four prompt seeds). Conditional methods use the same supplied primary set; structural and historical readouts are analyzed separately in Appendix C.1.4–C.1.5.

Intact only				Removed only		Intact → removed
Single ablation	AtP	EAP-IG	AtP* GradDrop	Conditional gradient	Conditional energy	**CoAx**
0.603 ± 0.007	0.650 ± 0.010	0.581 ± 0.027	0.665 ± 0.009	0.767 ± 0.005	0.758 ± 0.004	**0.941** ± 0.004

3.1 Backup discovery

The two-state contrast recovers the backup branch. The central result is that CoAx recovers the documented backups at 0.941 ROC-AUC, improving by 0.276 over the strongest intact-state attribution baseline and by 0.174 over the strongest removed-state control; against the representation-matched conditional-energy control, the gap is 0.183. Average precision gives the same conclusion under the $8/141$ class imbalance (0.557 versus 0.098 for AtP*; Appendix C.1.2), confirming that the advantage extends to the top of the ranking. A derived two-state AtP* contrast reaches 0.938 AUC, showing that the conditional principle transfers across estimators; unlike CoAx, it is a task-gradient approximation that requires backward passes rather than a direct finite-intervention score over the full output distribution (Appendix C.1.3).

The reason is simple: the three settings answer different questions. An intact-state score asks which components already matter; a score read only after primary-set removal asks which components matter in the intervened model, but still mixes pre-existing importance with newly recruited importance. CoAx instead asks which candidates become more important *because* the primary set was removed. At the vector level, the corresponding effect change $\Delta_u(\mathcal{S}) = \delta z_{u|\mathcal{S}} - \delta z_u$ has the exact decomposition in Proposition 2, aggregating every interaction order linking u to the removed primary set. Consistent with this interaction view, the primary-plus-backup module shows 1.9× joint amplification over summed isolated effects, while a size-matched random set is approximately additive (Figure 2). The 0.941 recovery therefore reflects access to the conditional causal information needed to identify a backup branch, rather than simply a stronger importance score. A historical direct-logit change reaches 0.971 AUC but requires the IO−S output direction (Appendix C.1.5).

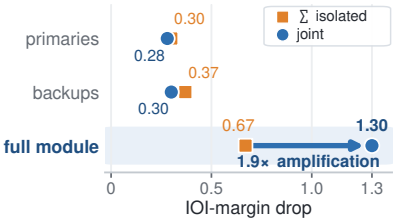

Figure 2: **Joint amplification** (1.9× vs. ∼1× random).

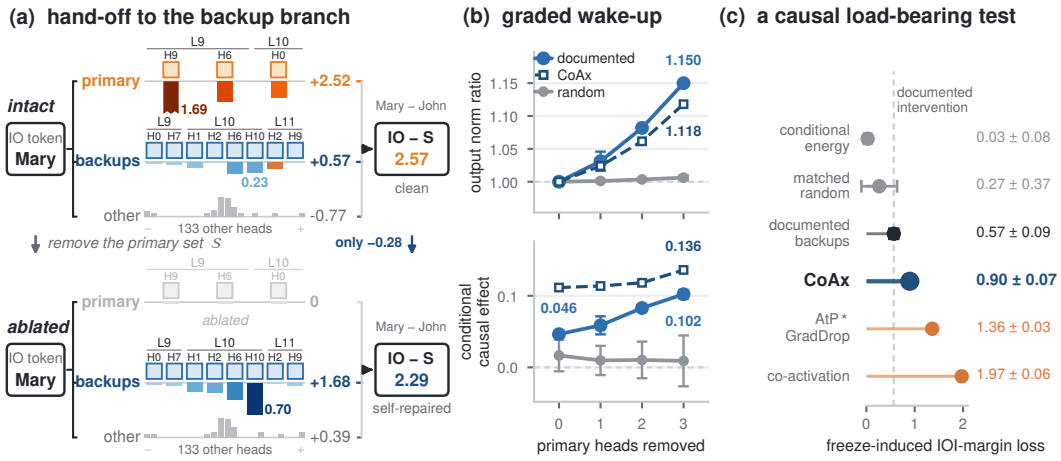

Figure 3: **Backup hand-off and causal repair** (GPT-2-small IOI; four prompt seeds). **(a)** Removing the primary name-movers shifts direct-logit attribution toward the documented backup branch while the IO−S margin remains largely preserved. **(b)** The documented backups and the CoAx-selected heads both increase on the wake-up readouts as more primaries are removed, while matched random heads remain flat. **(c)** Freezing each selector's top-8 at its intact-state outputs tests whether those heads carry repair; CoAx causes a 0.90-point IOI-margin loss.

Conditioning carries the main signal; Fisher geometry sharpens it. Holding the two-state contrast fixed, replacing the Fisher geometry with plain ℓ_2 still yields 0.905 AUC (Figure 5b); the full geometry raises this to 0.941. Thus the conditional comparison exposes most of the backup signal, while the Fisher geometry provides a principled refinement of its measurement.

3.2 Recovery tracks the identity of the removed circuit

To test whether recovery tracks the identity of the removed circuit, we first ask whether other primary sets with similar intervention severity recover the same backups, and then test whether LayerNorm rescaling could explain the signal.

Recovery follows circuit identity, not intervention severity. A damage-matched random set nearly matches the true primary intervention in IOI damage ($+0.295$ versus $+0.284$) but reaches only 0.426 backup AUC (Appendix C.2). We strengthen this control by searching 5,000 random three-head sets and retaining the 16 best matched in behavioral damage, Fisher displacement $\mathcal{E}(\delta z_S)$, and depth. On held-out prompts, these matched wrong sets reach only 0.40 ± 0.13 AUC, versus about 0.94 for the true primary set (Table 13). Recovery therefore tracks *which circuit* is removed, rather than merely how strongly the model is perturbed.

Freezing normalization sharpens the backup signal. Ablation can alter downstream LayerNorm statistics and rescale the remaining computation, a known component of transformer self-repair (Rushing & Nanda, 2024). We remove this rescaling channel by replaying ablated forwards with clean LayerNorm denominators fixed. Backup AUC rises from 0.941 to 0.983: the documented backups retain 80% of their conditional growth, while non-backup growth is suppressed enough for its mean to shift below zero (Appendix C.2.2). Thus LayerNorm rescaling does not create the backup signal; removing it instead suppresses background growth and sharpens the separation.

3.3 Causal validation of the backup hand-off

Backup-label recovery alone does not establish that the recovered branch actually carries self-repair. Figure 3 therefore addresses two stronger questions: does the backup response grow as more of the primary circuit is removed, and does preventing that response impair the repair?

The hand-off is graded and causally load-bearing. Removing the primary name-movers eliminates their direct-logit contribution, while the documented backup branch roughly triples its contri-

bution from $+0.57$ to $+1.68$ and the IO$-$S margin falls only from 2.57 to 2.29 (Figure 3a). As more primaries are removed, both documented backups and the CoAx-selected heads increase on the two wake-up readouts, while matched random heads remain flat (Figure 3b). Freezing the documented backups at their intact-state outputs causes a 0.57 ± 0.09 IOI-margin loss (Figure 3c). Together, these results show that backup recruitment grows with primary-circuit failure and is causally load-bearing.

CoAx's label-free selection is causally load-bearing. Applying the same freezing intervention to each selector's top-8, the CoAx set causes a 0.90 ± 0.07 IOI-margin loss, compared with 0.03 ± 0.08 for conditional energy alone and 0.27 ± 0.37 for a size-matched random set (Figure 3c). The freezing readout establishes that the selected set is causally load-bearing; circuit completion and knockout then test whether it matches the missing branch's causal role. Co-activation nearly matches CoAx on backup-label AUC (0.93 versus 0.941; Appendix C.1.4) but gives substantially worse backup-role knockout matching (Section 4). Thus, CoAx identifies a primary-set-specific branch that is recruited by circuit failure and causally carries repair, without using backup labels.

4 Completing the Circuit

Section 3 shows that CoAx recovers a branch that wakes up and carries self-repair. Conditional circuit completion asks for more: the recovered components should repair the causal account that was incomplete under intervention. We use the same raw headline CoAx ranking used for discovery, without backup labels, task gradient, or task-direction filtering, and take its top eight heads as the completion. Across the two circuit-level tests, this set reduces incompleteness from 0.47 to 0.24 and closely matches the documented backup knockout with little collateral change.

Table 2: **Circuit completion and backup-role validation** (GPT-2-small IOI; four prompt seeds). **Left:** the raw headline CoAx top-8 is added to the incomplete primary circuit, with no backup labels, task gradient, or task-direction filter. **Right:** each selector's raw top-8 knockout is compared with the documented backup intervention. Lower is better.

circuit	$D(C)\downarrow$
primary circuit	0.474 ± 0.030
documented circuit	0.300 ± 0.027
+conditional energy	$\mathbf{0.181\pm0.027}$
+CoAx	0.236 ± 0.037
+amplification ratio	0.258 ± 0.035
+random	0.499 ± 0.073

top-up	margin dist.↓	output KL↓	unrel. KL↓
documented backups	0.00	0.00	0.09
+CoAx	**0.99**	0.17	0.14
+amplification ratio	**0.99**	**0.07**	**0.09**
+conditional energy	1.71	0.46	0.47
+co-activation	1.79	0.47	0.15
+CoAx next-8	2.03	1.38	0.86
+AtP* GradDrop	2.70	1.02	0.47
+random	1.24	0.47	0.24

The recovered set substantially completes the circuit. Following Wang et al. (2023), completeness asks whether a circuit reproduces the full model's response to the same ablation. Let Ω denote the full head set, and let $C \subseteq \Omega$, with $\mathcal{S} \subseteq C$, denote the retained head support of a circuit for this completeness test. Let $\mathrm{LD}_C(x)$ be the IOI margin when only heads in C remain active (all others are mean-ablated). Writing $r(C) = \mathbb{E}_x[\mathrm{LD}_C(x) - \mathrm{LD}_{C\setminus\mathcal{S}}(x)]$ for the response of circuit C to primary-set removal, we define its incompleteness as $D(C) = |r(\Omega) - r(C)|$; lower is better. The incomplete primary circuit has $D = 0.47$, while the documented circuit reaches 0.30. Adding the raw CoAx top-8 reduces incompleteness to 0.24 (Table 2), using exactly the same label-free, task-direction-free ranking as discovery. Thus CoAx does more than localize backup components: it restores the missing counterfactual response and turns backup localization into circuit completion. Conditional energy reaches a lower scalar gap (0.18) by selecting generally large removed-state effects, but its knockout is much farther from the documented backup intervention; scalar response matching and branch identity are therefore complementary criteria.

The recovered set reproduces the backup knockout. We reverse the completion test and ask whether the selected knockout reproduces the documented backup effect. A large performance drop after removal does not by itself identify the backup, because disrupting unrelated computation can also hurt IOI. We therefore compare each primary-plus-selected removal with the documented

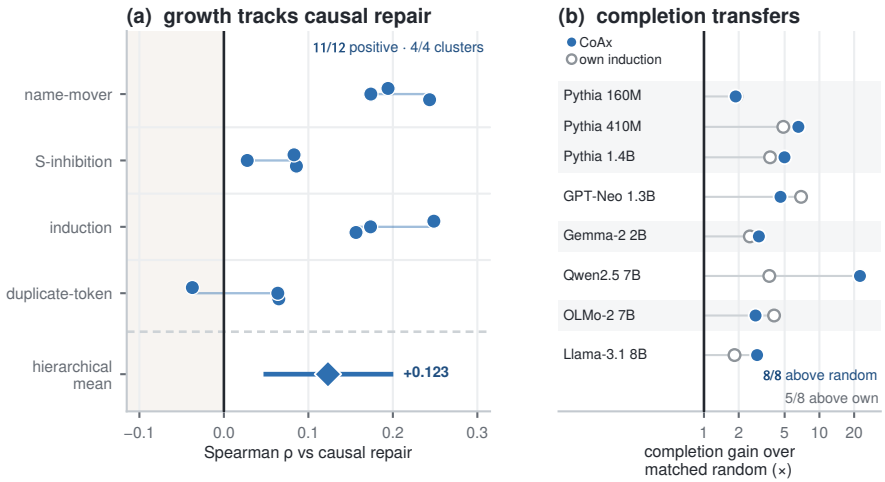

Figure 4: **Conditional repair generalizes across mechanisms and architectures.** **(a)** Signed growth has positive rank agreement with mechanism-matched held-out causal repair in $11/12$ instances and all four mechanism clusters; the bottom row shows the hierarchical mean and nested 95% CI. **(b)** CoAx-selected completions outperform matched random completions on all eight non-GPT-2 models spanning six architecture families; hollow circles show each model's next induction heads as a stronger role-matched control.

primary-plus-backup removal in both the IOI margin and the answer-position output distribution, while tracking collateral KL on unrelated text. The CoAx-selected set closely matches the documented intervention (0.99 margin distance and 0.17 KL), while keeping unrelated-text KL low at 0.14 (Table 2). A normalized amplification-ratio variant is statistically tied in margin distance and attains lower KL; signed growth remains our default because it is directly tied to the exact interaction decomposition in Proposition 2 and avoids division by small intact effects. Together with the completion result above, this closes the circuit-level loop: completion restores the missing counterfactual response, while knockout reproduces the causal role of the missing backup branch. Full completion protocol, knockout, and paired-uncertainty analyses are in Appendix D.

5 Conditional Repair Beyond Documented Backups

Sections 3–4 establish CoAx on a documented dormant backup branch. We now test whether the same conditional signal extends beyond that annotated setting in two ways: against independently measured repair across mechanisms and through held-out circuit completion across model families. Neither test requires a pre-identified backup branch.

Conditional growth aligns with repair across mechanisms. We construct an independently measured causal target: after ablating $\mathcal{S}$, we freeze each candidate u to its intact-state output, so any further behavioral loss measures repair carried by u. We quantify this by $\mathrm{repair}(u \mid \mathcal{S}) = m(\text{ablate } \mathcal{S}) - m(\text{ablate } \mathcal{S},\ \text{freeze } u)$. The primary sets follow documented IOI head roles (Wang et al., 2023); we evaluate name-mover and S-inhibition sets on the IOI margin, and induction and duplicate-token sets on repeated-token induction behavior (Olsson et al., 2022). Scores and repair targets use disjoint prompts, making their agreement out of sample. Across twelve held-out instances from four mechanism clusters, CoAx has positive rank agreement with mechanism-matched causal repair in $11/12$ instances and all four clusters, with a hierarchical mean of $+0.123$ and a nested 95% CI of $[+0.047, +0.201]$ (Figure 4a). Thus the signed-growth signal that recovers the documented IOI backups also tracks independently measured repair across distinct mechanisms. See Appendix E.1 for full details.

Conditional circuit completion transfers beyond GPT-2. We finally test set-level transfer on eight non-GPT-2 models spanning six architecture families. Under disjoint detection, calibration, and evaluation splits with equal selector budgets, adding CoAx-selected induction heads outper-

forms a size-matched random completion on all eight models (Figure 4b). A stronger role-matched control, each model's next-ranked induction heads, is also load-bearing; COAX gives the larger completion on five of eight models. Full held-out comparisons against single ablation, co-activation, AtP, and AtP* are in Appendix E.2. Thus conditional circuit completion transfers beyond GPT-2-small across diverse induction circuits. Taken together, cross-mechanism repair alignment and cross-model circuit completion show that primary-set removal is more than a diagnostic perturbation: it exposes how causal responsibility reorganizes after failure and provides the information needed to complete the circuit account that the intact model alone leaves hidden.

6 RELATED WORK

Circuit discovery and completeness. Mechanistic interpretability identifies circuits by intervening on components or edges and measuring how model behavior changes. Automated Circuit Discovery (ACDC) automates intervention-based pruning (Conmy et al., 2023); edge attribution patching and its integrated-gradient variant estimate edge effects efficiently (Syed et al., 2024; Hanna et al., 2024); and AtP* improves component localization with gradient-based corrections (Kramár et al., 2024). Subsequent work refines circuit representations, search procedures, or pruning objectives (Marks et al., 2025; Ferrando & Voita, 2024; Hsu et al., 2025; Bhaskar et al., 2024). These methods primarily seek the computation supporting a reference behavior. Complementary work on circuit completeness shows that redundant, OR-like paths can be omitted by standard discovery and combines noising and denoising interventions to recover them (Chen et al., 2025). COAX focuses on completeness relative to a specified failure state: given a supplied primary set, it asks which remaining components become causally important after that set is removed.

Self-repair and compensatory circuitry. Self-repair makes this state dependence concrete. The Hydra effect (McGrath et al., 2023), documented IOI backup name-movers (Wang et al., 2023), copy-suppression (McDougall et al., 2024), and LayerNorm and anti-erasure mechanisms (Rushing & Nanda, 2024) show how transformer computation can reorganize after intervention. AtP* GradDrop targets gradient-attribution failure modes, Gradient Interaction Modifications (GIM) introduces interaction-aware gradient corrections, and joint ablations expose redundancy hidden by isolated effects (Kramár et al., 2024; Edin et al., 2025; Ye, 2026). COAX complements these analyses by localizing compensating candidates given a supplied primary set, then testing whether they causally carry repair and complete the circuit.

Conditional and interaction-aware effects. State-dependent importance is relational: a candidate's effect depends on what else has been removed. Neuron Shapley averages marginal effects across many coalitions (Ghorbani & Zou, 2020), while interventional redundancy–synergy decompositions characterize dependence across removal contexts (Aghilar et al., 2026). Closest to our theory, causal-mediation analysis shows that component-wise patching estimands contain pairwise and higher-order interaction effects when one component's causal role depends on others (Vaidyanathan et al., 2026). The Möbius decomposition itself is standard (Rota, 1964); our contribution is to make the supplied failure set part of the query and use the resulting all-order aggregate as the estimand for conditional completion. Input-side co-activation exposes co-firing components but does not condition on a specified failure (Xu, 2026). Signed growth uses the intervention-induced increase in ablation-effect magnitude to target recruitment after the supplied failure.

7 CONCLUSION

Self-repair makes causal importance conditional on what remains available, so an intact-state circuit can cease to explain the model once that circuit is perturbed. COAX turns this difficulty into a probe: removing the supplied primary set reveals which previously weak components become causally important and allows the missing response to be incorporated back into the circuit account. Across documented backups, independent repair targets, and multiple architectures, this conditional view exposes intervention-dependent circuitry that intact-state localization alone cannot characterize. A causal explanation of a self-repairing model is therefore not a single intact-state circuit, but an account of how causal responsibility reorganizes when that circuit is disrupted.

Reproducibility statement. The paper is designed for direct reproduction: Section 2 and Appendix A specify the estimand, algorithm, implementation, and complexity; Appendix A.2 contains complete proofs; Table 6 consolidates protocol constants and random seeds; and Appendix B specifies the models, data, metrics, and baselines. Appendix F.4 reports compute and reproduction details. An anonymized supplementary archive provides the implementation, tested environment, machine-readable reference results, and scripts for reproducing the main experiments and figures, together with a run-all quickstart notebook and a concise three-minute tutorial. All model weights are frozen and no training is performed; the core GPT-2-small IOI experiments can also be reproduced on CPU.

Ethics statement. This work performs inference-only analysis of publicly released pretrained language models and uses no human subjects, private data, or personally identifying information. Model licenses, data sources, and broader-impact considerations are documented in Appendix F.

AI use statement. Generative AI tools were used to assist with language editing, presentation and organization, literature search, and feedback on the communication of experimental results. The research questions, methodology, theoretical analysis, implementation, experiments, and scientific conclusions were developed and verified by the authors. All AI-assisted material was reviewed and revised by the authors, who take full responsibility for the paper.

REFERENCES

Potito Aghilar, Sabino Roccotelli, Stanislao Fidanza, Vito Walter Anelli, Sebastiano Stramaglia, and Tommaso Di Noia. The representational limit of scalar interactions: An interventional decomposition, 2026. URL https://arxiv.org/abs/2606.19410.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=ePUVetPKu6.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 18506–18534, 2024. doi: 10.52202/079017-0587. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/20fdaf67581e6d7157376d1ed584040a-Abstract-Conference.html.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. Zenodo, 2021. URL https://doi.org/10.5281/zenodo.5297715.

Hang Chen, Jiaying Zhu, Xinyu Yang, and Wenya Wang. Rethinking circuit completeness in language models: AND, OR, and ADDER gates. In *Advances in Neural Information Processing Systems*, volume 38, pp. 150511–150540, 2025. doi: 10.52202/085713-5032. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/dd37fdb24a4e1cfa3ed5c247217a7394-Abstract-Conference.html.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Joakim Edin, Casper L. Christensen, Róbert Csordás, Tuukka Ruotsalo, Zhengxuan Wu, Maria Maistro, Jing Huang, and Lars Maaløe. Correcting gradient-based circuit localization via interaction-aware backpropagation, 2025. URL https://arxiv.org/abs/2505.17630.

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

Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17432–17445. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.965. URL https://aclanthology.org/2024.emnlp-main.965/.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Amirata Ghorbani and James Y Zou. Neuron shapley: Discovering the responsible neurons. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5922–5932, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/41c542dfe6e4fc3deb251d64cf6ed2e4-Abstract.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

Aliyah R. Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Y. Odisho, Peter R. Carroll, and Bin Yu. Efficient automated circuit discovery in transformers using contextual decomposition. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=41HlN8XYM5.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 337–363. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.22. URL https://aclanthology.org/2024.blackboxnlp-1.22/.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations, 2023. URL https://arxiv.org/abs/2307.15771.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zSf8PJyQb2.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

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

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019.

Gian-Carlo Rota. On the foundations of combinatorial theory i. theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368, 1964. doi: 10.1007/BF00531932.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855. PMLR, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint*, 2026. arXiv:2606.27510.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Yongzhong Xu. Closure-validated circuit discovery in attention heads: Co-activation proposes, ablation disposes. *arXiv preprint*, 2026. arXiv:2606.09607.

Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. Measuring affinity between attention-head weight subspaces via the projection kernel, 2026. URL https://arxiv.org/abs/2601.10266.

Donald Ye. Hidden heroes and gradient bloats: Layer-wise redundancy inverts attribution in transformers, 2026. URL https://arxiv.org/abs/2602.01442.

APPENDIX OVERVIEW

The appendix is organized into six parts, following the technical and empirical progression of the paper. Parts A–B specify the method and shared experimental protocol; Parts C–E provide the detailed evidence behind Sections 3–5; and Part F records scope, ethics, and reproducibility information.

Part A: Method Details. Appendix A provides the complete technical specification of COAX.

Appendix A.1: Notation and output geometry. Formal notation, terminology, Fisher geometry, and the finite-ablation interpretation.
Appendix A.2: Proofs. Complete proofs of Propositions 1 and 2, together with the signed-growth decomposition.
Appendix A.3: Algorithm, cost, and protocol. The full algorithm, pass complexity, measured efficiency, and scoring conventions.
Appendix A.4: Hyperparameters and robustness. Protocol constants, ablation-value robustness, calibration efficiency, output-geometry analysis, and task-alignment diagnostics.

Part B: Experimental Setup. Appendix B defines the models, circuit roles and supplied primary sets, evaluation data, metrics, and intact-state, conditional, and structural baselines used throughout the experiments.

Part C: Backup Recovery and Mechanistic Validation. Appendix C expands Section 3.

Appendix C.1: Backup recovery and statistical validation. Recovery across seeds, significance and imbalance-sensitive metrics, historical-readout and layer-depth controls, and a primary-removal-free reference construction.
Appendix C.2: Recovery specific to the removed primary set. Alternative and geometry-matched primary sets, contamination, and LayerNorm-freezing controls.
Appendix C.3: Causal validation of the recovered branch. Circuit context, graded recruitment and dose response, selector-specific freezing, structural corroboration, and a concrete backup head.
Appendix C.4: What drives conditional growth. The intact-to-removed contrast, output geometry, controlled dormancy and task-alignment tests, and empirical non-additivity.
Appendix C.5: Robustness. Primary-set size and prompt-template robustness.

Part D: Conditional Circuit Completion and Knockout. Appendix D expands Section 4.

Appendix D.1: Circuit-completion controls. Raw headline completion, matched controls, and prompt-level uncertainty.
Appendix D.2: Knockout against the documented backup intervention. Counterfactual matching in IOI margin and output distribution, together with collateral-change and paired-uncertainty controls.

Part E: Conditional Repair Beyond Documented Backups. Appendix E expands the two generalization tests of Section 5.

Appendix E.1: Repair prediction across mechanisms. The twelve held-out instances under mechanism-matched causal targets and hierarchical inference.
Appendix E.2: Conditional circuit completion across models. Held-out, equal-budget comparisons underlying Figure 4b.

Part F: Limitations, Broader Impact, Ethics, and Reproducibility. Appendix F records the empirical scope and granularity of the study, broader-impact and ethics considerations, model and data licenses, compute, and reproduction details.

A METHOD DETAILS

A.1 NOTATION AND OUTPUT GEOMETRY

Notation and conventions. Table 3 summarizes the symbols for the theoretical development.

Table 3: Core notation for the conditional-completion analysis.

Symbol	Meaning
	Problem objects
$\mathcal{S}$	supplied primary set whose removal defines the conditional intervention
$\mathcal{U}$	candidate units, with $\mathcal{U} \cap \mathcal{S} = \emptyset$
u	a candidate unit in $\mathcal{U}$
	Sets used in the interaction decomposition
M	generic ablated set, with logits z_M and joint effect $\delta z_M = z_0 - z_M$
$T \subseteq M$	interaction set contributing to the decomposition of δz_M
$A \subseteq T$	dummy subset index in the Möbius transform defining I_T
$R \subseteq \mathcal{S}$	primary-set subset in Proposition 2, with $T = R \cup \{u\}$
	Effects and scores
δz_u	ablation effect of u with $\mathcal{S}$ present
$\delta z_{u\|\mathcal{S}}$	ablation effect of u after $\mathcal{S}$ is removed
$\Delta_u(\mathcal{S})$	effect change $\delta z_{u\|\mathcal{S}} - \delta z_u$ induced by primary-set removal
I_T	Möbius interaction associated with the unit set T
$\mathcal{E}(v)$	magnitude of a logit perturbation in the fixed clean output geometry
$\mathrm{CoAx}(u;\mathcal{S})$	CoAx conditional-growth score

Output geometry. We first justify the output measure in Eq. 3. Fix an evaluated input–position pair (x, t) and suppress this dependence below; recall that $p_0 = \mathrm{softmax}(z_0)$ is the clean output distribution. For a logit perturbation v, the centered term $v - \langle p_0, v\rangle \mathbf{1}$, where $\mathbf{1}$ is the all-ones vector, removes the shared-logit-shift direction to which softmax is invariant, while the subsequent $\sqrt{p_0}$ weighting yields the transformed perturbation $\widetilde{v}$. As shown below, Euclidean inner products between such transformed perturbations coincide exactly with Fisher inner products at the clean output distribution p_0.

For vocabulary-logit perturbations v and w, let $\bar{v} = \langle p_0, v\rangle$ and $\bar{w} = \langle p_0, w\rangle$ denote their p_0-weighted means. Using the transformation in Eq. 3,

$$\begin{aligned}\langle \widetilde{v}, \widetilde{w}\rangle &= \sum_i (p_0)_i (v_i - \bar{v})(w_i - \bar{w}) \\ &= \sum_i (p_0)_i v_i w_i - \bar{v}\,\bar{w} \\ &= v^\top \left(\mathrm{diag}(p_0) - p_0 p_0^\top\right) w = v^\top F w.\end{aligned} \tag{7}$$

Thus Eq. 3 is exactly the squared Fisher-metric magnitude of the perturbation, averaged over the evaluated input–position pairs:

$$\mathcal{E}(v) = \mathbb{E}_{x,t}\left[v^\top F v\right]. \tag{8}$$

Because F is positive semidefinite, $\mathcal{E}(v) \geq 0$. Moreover, adding the same constant to every logit leaves both the softmax distribution and $\mathcal{E}(v)$ unchanged, so the measure discards the shared-shift direction that is invisible to softmax.

This Fisher metric also has a direct distributional interpretation. For a scalar path parameter α, define the straight-line logit path $z_\alpha = z_0 - \alpha v$ and the corresponding scalar function

$$\phi(\alpha) := \mathrm{KL}(p_0 \,\|\, \mathrm{softmax}(z_0 - \alpha v)). \tag{9}$$

Since $\nabla_z \mathrm{KL}(p_0 \,\|\, \mathrm{softmax}(z)) = \mathrm{softmax}(z) - p_0$, the clean point satisfies $\phi(0) = 0$ and $\phi'(0) = 0$. Moreover, the Hessian at z_0 is the categorical Fisher matrix $F = \mathrm{diag}(p_0) - p_0 p_0^\top$, so

$$\phi''(0) = v^\top F v. \tag{10}$$

A Taylor expansion around $\alpha = 0$ for fixed v therefore gives

$$\mathrm{KL}(p_0 \,\|\, \mathrm{softmax}(z_0 - \alpha v)) = \frac{\alpha^2}{2}\, v^\top F v + O(\alpha^3), \qquad \alpha \to 0. \tag{11}$$

Hence $v^\top F v$ characterizes the local quadratic sensitivity of the output distribution along the logit direction v. This gives the interpretation used throughout the paper: $\mathcal{E}(v)$ measures the magnitude of an observed logit displacement in the fixed clean output geometry.

Interpretation for finite ablations. The local KL expansion above motivates the geometry of the output measure, while COAX itself is applied to finite ablations. For each intervention, we directly observe the resulting logit displacement and measure its magnitude using the Fisher matrix F at the clean output distribution as a fixed reference geometry.

Eq. 11 therefore describes only the local behavior of the straight-line path $z_\alpha = z_0 - \alpha v$ in logit space. A full internal ablation need not follow this path: for example, if h_u denotes the residual-stream contribution of component u, gradually scaling it as $h_u \leftarrow (1-\alpha)h_u$ is followed by nonlinear transformer computation, so the resulting logits need not equal $z_0 - \alpha\delta z_u$. Accordingly, the method does not require the full-ablation KL to equal its quadratic approximation, nor does it assume a linear internal ablation trajectory. The Fisher expansion provides a local geometric interpretation of the metric used to compare the finite displacements actually produced by intervention.

A.2 PROOFS

We prove Proposition 1 by an explicit non-identifiability construction and Proposition 2 by Möbius inversion, and then derive the score decomposition in Eq. 6. Except where $\mathcal{E}$ is applied, the identities below hold pointwise for each evaluated input–position pair (x, t).

A.2.1 PROOF OF PROPOSITION 1

Proof. We construct two candidates with identical intact-state information in the sense of Definition 2, but different causal roles after the primary set is removed.

It suffices to consider an input distribution supported on one evaluated input–position pair. Choose a nonconstant logit direction w. Since the clean softmax distribution p_0 has strictly positive entries, Eq. 8 gives

$$\mathcal{E}(w) = w^\top F w > 0. \tag{12}$$

Let q denote a scalar state supplied by the primary path: $q = 1$ when $\mathcal{S}$ is present and $q = 0$ when it is removed. Let a_b and a_n denote the activations of a prospective backup b and an irrelevant candidate n, respectively. Take $a_b = a_n = 1$ on the original model, with ablation setting the corresponding activation to zero. Consider the downstream logit readout

$$z(q, a_b, a_n) = z_{\mathrm{base}} + \big[q + (1-q)a_b\big]\, w, \tag{13}$$

where z_{base} collects all remaining logit contributions. The candidate n does not enter this readout.

When $\mathcal{S}$ is present, $q = 1$, so the coefficient of w is one regardless of a_b. Consequently, ablating either candidate leaves the clean logits unchanged:

$$z_{\{b\}} = z_{\{n\}} = z_0 = z_{\mathrm{base}} + w, \qquad \delta z_b = \delta z_n = 0. \tag{14}$$

Moreover, at the intact state,

$$\frac{\partial z}{\partial a_b} = (1-q)w = 0, \qquad \frac{\partial z}{\partial a_n} = 0. \tag{15}$$

Thus b and n have the same forward activation, the same output derivative, and the same single-ablation effect on the intact model. By Definition 2, they therefore have identical intact-state information, with

$$\mathcal{E}(\delta z_b) = \mathcal{E}(\delta z_n) = 0. \tag{16}$$

Hence every intact-state score assigns the two candidates the same value.

Now remove the primary set, so $q = 0$. With b present, its activation supplies exactly the missing logit direction w:

$$z_{\mathcal{S}} = z_{\text{base}} + w = z_0. \tag{17}$$

Thus b perfectly repairs the primary-path contribution at the output. If b is also ablated, that contribution disappears, whereas ablating n still changes nothing:

$$z_{\mathcal{S}\cup\{b\}} = z_{\text{base}}, \qquad z_{\mathcal{S}\cup\{n\}} = z_{\mathcal{S}}. \tag{18}$$

Therefore the conditional ablation effects are

$$\delta z_{b|\mathcal{S}} = z_{\mathcal{S}} - z_{\mathcal{S}\cup\{b\}} = w, \qquad \delta z_{n|\mathcal{S}} = z_{\mathcal{S}} - z_{\mathcal{S}\cup\{n\}} = 0. \tag{19}$$

Using Eq. 12,

$$\mathcal{E}(\delta z_{b|\mathcal{S}}) > 0, \qquad \mathcal{E}(\delta z_{n|\mathcal{S}}) = 0. \tag{20}$$

Finally, because both intact effects vanish, Eq. 2 gives

$$\text{CoAx}(b;\mathcal{S}) = \mathcal{E}(w) > 0, \qquad \text{CoAx}(n;\mathcal{S}) = 0. \tag{21}$$

Thus the two candidates are indistinguishable from their intact-state records, while conditional co-ablation separates the dormant backup b from the irrelevant candidate n. □

Interpretation. Proposition 1 establishes a non-identifiability result for the intact-state information class in Definition 2. The same per-unit evidence on the original model can correspond to different causal roles after the primary set is removed: in the construction above, b exactly replaces the lost primary contribution w, while n remains irrelevant. The missing distinction is therefore genuinely conditional on primary-set removal. Perfect dormancy gives the sharpest case; the experiments examine whether this distinction remains recoverable when real backups retain some intact-state signal.

A.2.2 Proof of Proposition 2

Proof. Fix a candidate $u \in \mathcal{U}$, so $u \notin \mathcal{S}$. Recall that for any ablated set M, $\delta z_M = z_0 - z_M$, with $\delta z_\emptyset = 0$. For the vector-valued set function $M \mapsto \delta z_M$, the Möbius interaction associated with each nonempty set T is

$$I_T = \sum_{A\subseteq T} (-1)^{|T|-|A|} \delta z_A, \tag{22}$$

where A is a dummy subset of T. Möbius inversion on the subset lattice (Rota, 1964) applies coordinatewise to the logit vector and gives, for every ablated set M,

$$\delta z_M = \sum_{\emptyset\neq T\subseteq M} I_T. \tag{23}$$

The conditional ablation effect can be written as the difference between two joint ablation effects:

$$\delta z_{u|\mathcal{S}} = z_{\mathcal{S}} - z_{\mathcal{S}\cup\{u\}} = \delta z_{\mathcal{S}\cup\{u\}} - \delta z_{\mathcal{S}}. \tag{24}$$

Applying Eq. 23 to the two terms yields

$$\begin{aligned} \delta z_{u|\mathcal{S}} &= \sum_{\emptyset\neq T\subseteq \mathcal{S}\cup\{u\}} I_T - \sum_{\emptyset\neq T\subseteq \mathcal{S}} I_T \\ &= \sum_{\substack{T\subseteq \mathcal{S}\cup\{u\} \\ u\in T}} I_T. \end{aligned} \tag{25}$$

Indeed, every interaction whose set T does not contain u is a subset of $\mathcal{S}$ and therefore appears in both sums, so these terms cancel exactly.

Because $u \notin \mathcal{S}$, every surviving set containing u can be written uniquely as $T = R \cup \{u\}$ for a subset $R \subseteq \mathcal{S}$. Re-indexing the remaining terms therefore gives

$$\delta z_{u|\mathcal{S}} = \sum_{R\subseteq \mathcal{S}} I_{R\cup\{u\}}. \tag{26}$$

The term $R = \emptyset$ is $I_{\{u\}} = \delta z_u$. More generally, an index set R of size $|R|$ contributes an interaction of order $|R| + 1$: $|R| = 1$ gives the pairwise candidate–primary interactions, while $|R| \geq 2$ gives all higher orders.

Subtracting the first-order term gives the effect change induced by primary-set removal:

$$\Delta_u(\mathcal{S}) := \delta z_{u|\mathcal{S}} - \delta z_u = \sum_{\emptyset \neq R \subseteq \mathcal{S}} I_{R \cup \{u\}}. \tag{27}$$

Thus $\Delta_u(\mathcal{S})$ is exactly the aggregate of every interaction order that contains the candidate u and at least one unit from the removed primary set. No approximation, pairwise truncation, or assumption of linear model responses is involved; higher-order nonlinear joint effects are represented by the corresponding I_T terms. □

A.2.3 SCORE DECOMPOSITION

We now connect the all-order effect change in Proposition 2 to the scalar CoAx score. By definition,

$$\delta z_{u|\mathcal{S}} = \delta z_u + \Delta_u(\mathcal{S}), \tag{28}$$

where Eq. 27 shows that $\Delta_u(\mathcal{S})$ is exactly the aggregate of all interaction orders linking u to the removed primary set.

For each evaluated input–position pair (x, t), the clean distribution p_0 is fixed, so the transformation $v \mapsto \widetilde{v}$ in Eq. 3 is linear. Hence

$$\widetilde{\delta z}_{u|\mathcal{S}} = \widetilde{\delta z}_u + \widetilde{\Delta}_u(\mathcal{S}), \tag{29}$$

where $\widetilde{\Delta}_u(\mathcal{S})$ is the same centered, Fisher-weighted transformation applied to the effect change. Substituting into $\mathrm{CoAx}(u; \mathcal{S}) = \mathcal{E}(\delta z_{u|\mathcal{S}}) - \mathcal{E}(\delta z_u)$ gives

$$\begin{aligned} \mathrm{CoAx}(u; \mathcal{S}) &= \mathbb{E}_{x,t}\left[\left\|\widetilde{\delta z}_u + \widetilde{\Delta}_u(\mathcal{S})\right\|_2^2 - \left\|\widetilde{\delta z}_u\right\|_2^2\right] \\ &= \mathbb{E}_{x,t}\left[2\left\langle \widetilde{\delta z}_u, \widetilde{\Delta}_u(\mathcal{S})\right\rangle + \left\|\widetilde{\Delta}_u(\mathcal{S})\right\|_2^2\right]. \end{aligned} \tag{30}$$

This is Eq. 6. The squared-change term measures the magnitude of the *aggregate* interaction-induced effect change; because $\Delta_u(\mathcal{S})$ is a vector sum, different interaction orders may reinforce or cancel before this norm is taken. For a perfectly dormant unit, $\mathcal{E}(\delta z_u) = 0$, so the intact-state cross term vanishes and

$$\mathrm{CoAx}(u; \mathcal{S}) = \mathcal{E}(\Delta_u(\mathcal{S}))\,. \tag{31}$$

Thus the score directly measures the newly exposed conditional effect in the sharp dormant case, while retaining the signed change in ablation-effect magnitude away from that limit.

A.3 ALGORITHM, COST, AND PROTOCOL

A.3.1 ALGORITHM AND PASS COMPLEXITY

Algorithm 1 gives the direct finite-ablation computation of the CoAx score in Eq. 2. Let $\mathcal{X}$ denote the calibration set of evaluated input–position pairs (x, t); all expectations below are empirical averages over $\mathcal{X}$. The clean output distribution $p_0(x, t)$ is computed once and then held fixed across all interventions.

Pass complexity. The computation mirrors the two states compared by CoAx. The clean logits z_0 and the primary-ablated logits $z_{\mathcal{S}}$ are shared across all candidates, while each $u \in \mathcal{U}$ requires two candidate-specific evaluations: $z_{\{u\}}$ for its intact ablation effect and $z_{\mathcal{S}\cup\{u\}}$ for its conditional ablation effect. Thus scoring $|\mathcal{U}|$ candidates requires $2|\mathcal{U}|+2$ forward evaluations over the calibration set when both shared references are counted, or $2|\mathcal{U}| + 1$ additional evaluations when the clean reference is already available. In either convention, the pass complexity is $O(|\mathcal{U}|)$, with no backward pass or task gradient. The same clean distribution p_0 is used for both the intact and conditional effects, so their magnitudes are compared in a common output geometry.

Algorithm 1 CoAx conditional co-ablation

Input: frozen model; calibration set $\mathcal{X}$ of input–position pairs (x, t); primary set $\mathcal{S}$; candidate set $\mathcal{U}$ with $\mathcal{U} \cap \mathcal{S} = \emptyset$
Output: $\mathrm{CoAx}(u; \mathcal{S})$ for all $u \in \mathcal{U}$
compute clean logits $z_0(x, t)$ and $p_0(x, t) \leftarrow \mathrm{softmax}(z_0(x, t))$ for all $(x, t) \in \mathcal{X}$
for $u \in \mathcal{U}$ **do**
    ablate u and compute $\delta z_u \leftarrow z_0 - z_{\{u\}}$
    compute $\widetilde{\delta z}_u \leftarrow \sqrt{p_0} \odot \big(\delta z_u - \langle p_0, \delta z_u \rangle \mathbf{1}\big)$
    compute and store $\mathcal{E}(\delta z_u) \leftarrow \mathbb{E}_{(x,t)\in\mathcal{X}} \|\widetilde{\delta z}_u\|_2^2$
end for
ablate $\mathcal{S}$ jointly and compute $z_{\mathcal{S}}$ {shared conditional reference}
for $u \in \mathcal{U}$ **do**
    ablate $\mathcal{S} \cup \{u\}$ and compute $\delta z_{u|\mathcal{S}} \leftarrow z_{\mathcal{S}} - z_{\mathcal{S}\cup\{u\}}$
    compute $\widetilde{\delta z}_{u|\mathcal{S}} \leftarrow \sqrt{p_0} \odot \big(\delta z_{u|\mathcal{S}} - \langle p_0, \delta z_{u|\mathcal{S}} \rangle \mathbf{1}\big)$
    $\mathrm{CoAx}(u; \mathcal{S}) \leftarrow \mathbb{E}_{(x,t)\in\mathcal{X}} \|\widetilde{\delta z}_{u|\mathcal{S}}\|_2^2 - \mathcal{E}(\delta z_u)$ {conditional growth}
end for
return $\mathrm{CoAx}(u; \mathcal{S})$ for ranking

Direct access to the interaction aggregate. Proposition 2 identifies the effect change $\Delta_u(\mathcal{S})$ as the exact aggregate of all interaction orders linking candidate u to the removed primary set. CoAx does not recover those constituent terms individually: doing so would, in general, require subset evaluations, with even a pairwise truncation costing $O(|\mathcal{U}||\mathcal{S}|)$ and omitting higher orders. Instead, the two-state scan measures the complete vector aggregate $\Delta_u(\mathcal{S})$ directly in $O(|\mathcal{U}|)$ evaluations.

Approximation and reuse. For the optional top-r large-vocabulary approximation, the top-r vocabulary support is selected once from the clean distribution for each evaluated input–position pair and then held fixed across all interventions. The clean probabilities are renormalized on that support before applying Eq. 3. If circuit completion is repeated after enlarging $\mathcal{S}$, the intact effects δz_u, their transformed representations, and $\mathcal{E}(\delta z_u)$ can be reused for candidates that remain outside the primary set. Only the new shared conditional reference $z_{\mathcal{S}}$ and the corresponding conditional candidate evaluations $z_{\mathcal{S}\cup\{u\}}$ need to be recomputed.

A.3.2 Cost

The $O(|\mathcal{U}|)$ pass complexity translates to modest wall-clock overhead in practice. Table 4 reports measured runtime and peak GPU memory for one intact single-unit scan and for the complete two-state CoAx scoring pass, each using 48 calibration prompts. Across architectures from GPT-2-small to Llama-3.1-8B, scoring takes seconds to minutes on a single GPU and requires no backward pass. Runtime reflects both the cost of a model forward pass and the number of scored units, while peak memory is dominated by the frozen model.

Table 4: Measured cost of CoAx on a single GPU with 48 calibration prompts. *Single* denotes one intact candidate scan; CoAx *total* denotes the full two-state computation, including the intact and primary-conditional candidate scans. Both are linear in the number of scored units. *Units* counts the scored structured units, one per attention head (layers × query heads; on grouped-query models the query heads, not the key–value groups). Peak memory includes the frozen model.

model	params	units	single (s)	CoAx total (s)	peak GB
GPT-2-small	124M	144	2.9	5.7	0.8
GPT-2-medium	355M	384	19	39	1.7
Pythia-410m	410M	384	18	37	1.9
GPT-Neo-1.3b	1.3B	384	64	128	5.6
Gemma-2-2b	2B	208	18	35	6.0
OLMo-2-7B	7B	1024	182	368	14.9
Llama-3.1-8B	8B	1024	180	360	16.4

A.3.3 EFFICIENCY

COAX uses two forward-only candidate scans, and each scan has a distinct role: the conditional scan measures candidate effects after the primary set is removed, while the intact scan removes pre-existing importance to isolate the growth caused by that intervention. This two-state design provides three complementary efficiency advantages.

One additional scan buys the conditional contrast. The closest matched control is conditional energy alone: it uses the same supplied primary set, prompts, finite-ablation machinery, and Fisher geometry as COAX, but ranks candidates only by $\mathcal{E}(\delta z_{u|\mathcal{S}})$. It requires a single conditional candidate scan and reaches 0.758 ± 0.004 backup AUC. Adding the intact scan enables the subtraction in Eq. 2, raising recovery to 0.941 ± 0.004. Thus the additional $O(|\mathcal{U}|)$ scan is spent specifically on separating intervention-induced growth from importance that was already present before the intervention.

All-order information without interaction enumeration. By Proposition 2, the effect change $\Delta_u(\mathcal{S})$ obtained from these two states is exactly the aggregate of all interaction orders linking candidate u to the removed primary set. Recovering those orders individually is exponential in $|\mathcal{S}|$. Even the natural pairwise truncation requires $O(|\mathcal{U}||\mathcal{S}|)$ candidate–primary joint evaluations and omits every order above two, whereas COAX obtains the complete aggregate in $O(|\mathcal{U}|)$ forward evaluations. Conditioning therefore obtains the complete effect-change aggregate directly, without decomposing it into constituent interaction orders.

Forward-only and data-efficient. Unlike the gradient baselines, COAX requires neither a backward pass nor a task gradient; all candidate interventions are independent and can be batched or parallelized. On GPT-2-small, the strongest intact-state gradient baseline reaches 0.665 ± 0.009 backup AUC, compared with 0.941 ± 0.004 for COAX. The calibration requirement is also small: 16 unlabeled prompts already reach 0.931 AUC, 32 reach 0.937, and the 96-prompt main setting reaches 0.941 (Figure 5a).

Table 5: Compute–recovery trade-off on GPT-2-small IOI. The three supplied primary heads are excluded from the backup ranking, leaving $|\mathcal{U}| = 141$ candidates. Forward/backward counts omit the shared clean reference. L denotes the number of transformer blocks. † denotes methods given the same supplied primary set as COAX.

method	forwards	backwards	task grad.	backup AUC
Intact-state scores				
single ablation	$\|\mathcal{U}\|$	0	—	$0.603 \pm .007$
AtP	1	1	✓	$0.650 \pm .010$
activation-space EAP-IG	$5L$	$5L$	✓	$0.581 \pm .027$
AtP* GradDrop	L	L	✓	$0.665 \pm .009$
Given the same primary set				
conditional energy only†	$\|\mathcal{U}\| + 1$	0	—	$0.758 \pm .004$
pairwise-truncated growth†	$\|\mathcal{U}\| + \|\mathcal{S}\| + \|\mathcal{U}\|\|\mathcal{S}\|$	0	—	$0.927 \pm .001$
COAX †	$2\|\mathcal{U}\| + 1$	0	—	$\mathbf{0.941} \pm .004$

Notes: For IOI, $|\mathcal{U}| = 141$ and $|\mathcal{S}| = 3$, giving 283 forward evaluations for COAX versus 567 for the pairwise-truncated construction. The latter retains only $\Delta_u^{(2)}(\mathcal{S}) = \sum_{s \in \mathcal{S}} I_{\{u,s\}}$, whereas Proposition 2 shows that COAX obtains the complete all-order aggregate $\Delta_u(\mathcal{S})$.

A.3.4 SCORING CONVENTIONS

The following conventions make the score and unit definition precise.

- **Aggregation.** For each prompt, $\mathcal{E}(\cdot)$ is first averaged over its evaluated token positions and then averaged across prompts, giving every prompt equal weight. In the main IOI discovery setting there is a single evaluated position—the answer position—as specified in Table 6.
- **Conditional intervention.** All units in the supplied primary set $\mathcal{S}$ are ablated jointly, rather than one at a time. The resulting logits $z_{\mathcal{S}}$ form a single conditional reference shared by all candidates, so every $\delta z_{u|\mathcal{S}}$ is measured relative to the same primary-ablated state.

- **Ranking.** Candidates are ranked directly by the signed growth score $\mathrm{CoAx}(u; \mathcal{S})$, with no ReLU or post-hoc threshold. Positive values indicate amplification after primary-set removal, whereas negative values indicate attenuation and therefore naturally fall lower in a backup ranking.
- **Grouped-query attention.** Each query head is treated as one unit through its contribution to the attention output. Shared key–value heads in grouped-query attention do not change this unit definition, so Table 4 counts layers $\times$ query heads for every model.

A.4 Hyperparameters and Robustness

A.4.1 Hyperparameters

CoAx has few tunable quantities; most settings specify the intervention and evaluation protocol rather than learned or task-optimized hyperparameters. The main IOI discovery setting uses 96 calibration prompts, reads logits at the *answer position*, and evaluates the *full vocabulary*. Zero ablation is the default intervention.

The supplied primary set follows the documented circuit definition: three name-mover heads for IOI and four induction heads for the induction experiments. Robustness to smaller supplied primary sets is reported in Appendix C.5.1. For the largest-vocabulary models in Appendix E.2, we use the fixed-support top-4096 approximation described above.

Table 6: Main protocol and auxiliary experimental settings. Task- or model-specific deviations are stated in the corresponding sections.

setting	value
Main discovery protocol	
output position	answer position (last prompt position)
vocabulary	full (50,257 for GPT-2-small); optional top-r approximation
output geometry	centered Fisher with fixed clean p_0 (Eq. 3)
discovery prompts	96 IOI calibration prompts
ablation value	zero; alternatives in Table 7
primary set $\lvert\mathcal{S}\rvert$	documented circuit definition (3 name-movers; 4 induction heads)
prompt seeds	$\{1, 15, 22, 8\}$
Baseline settings	
EAP-IG integration steps	5
Auxiliary analyses	
random-control draws	20 (attribution / knockout)
permutation shuffles	10,000
synthetic trials	40

A.4.2 Robustness to the ablation value

Faithfulness estimates can depend on the choice of ablation operator (Miller et al., 2024). We therefore recompute the same CoAx ranking under four qualitatively different replacement rules: zero ablation, mean replacement, empirical resampling, and moment-matched Gaussian replacement.

Replacement statistics are estimated from head activations on the same 96 IOI calibration prompts, pooled over token positions: mean replacement uses the empirical mean, resampling draws uniformly from this pool, and Gaussian replacement matches its per-dimension mean and standard deviation. Each stochastic draw is held fixed across the intact and conditional states, preventing resampling noise from entering their contrast.

Table 7: Robustness to the ablation operator on GPT-2-small IOI. Values are backup-discovery AUC averaged over the four prompt seeds $\{1, 15, 22, 8\}$.

method	zero	mean	resample	Gaussian
single ablation	$0.603 \pm .007$	$0.753 \pm .008$	$0.691 \pm .011$	$0.699 \pm .018$
CoAx (ours)	$\mathbf{0.941} \pm .004$	$\mathbf{0.974} \pm .002$	$\mathbf{0.961} \pm .008$	$\mathbf{0.963} \pm .003$

As shown in Table 7, CoAx achieves 0.941–0.974 across the four operators, while single-ablation ranking remains at 0.603–0.753. The conditional ranking therefore remains stable across deterministic, empirical, and stochastic replacement rules rather than depending on a particular ablation value.

A.4.3 Method analysis: data efficiency, output geometry, and task alignment

Figure 5 isolates three design properties of CoAx: its calibration requirement, the contribution of the output geometry, and its dependence on alignment with the task direction.

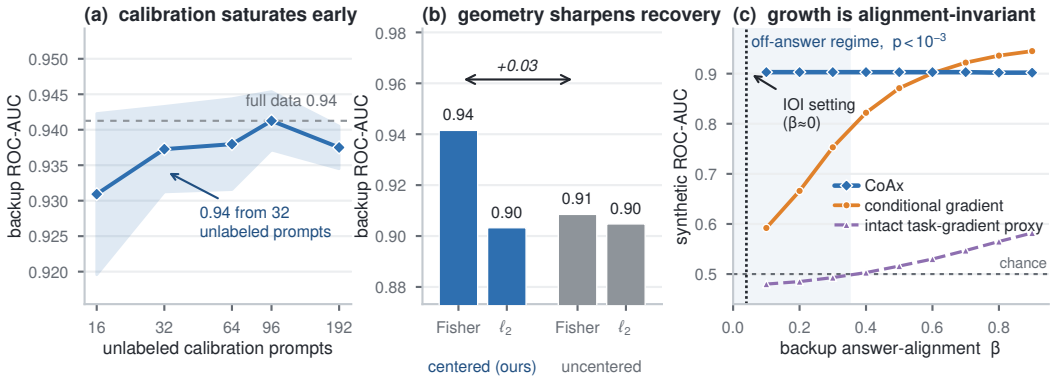

Figure 5: **Diagnostics of the CoAx score. (a)** Backup AUC versus unlabeled calibration-set size (GPT-2-small IOI, four prompt seeds); recovery is already near the 96-prompt result with 16–32 prompts. **(b)** Output-geometry ablation under the same conditional contrast: the full centered Fisher geometry outperforms plain ℓ_2, centering alone, and probability weighting alone. **(c)** Synthetic backup AUC as answer alignment β varies. Full-distribution CoAx remains stable as backups move off the task direction, whereas the conditional gradient and an intact task-gradient proxy deteriorate; the latter remains near chance. The measured IOI setting lies in the low-answer-alignment regime of the sweep.

Data efficiency. CoAx requires only unlabeled calibration prompts. On GPT-2-small IOI, 16 prompts already reach 0.931 backup AUC, 32 reach 0.937, and the 96-prompt main setting reaches 0.941 (Figure 5a). Thus most of the final recovery is already obtained from a small calibration set.

Output geometry. Conditioning provides the main recovery signal (Table 17); Figure 5b isolates the additional contribution of the output geometry. Using the same conditional contrast, a plain ℓ_2 magnitude reaches 0.905, while the full centered Fisher geometry reaches 0.941. Centering alone (0.903) and probability weighting alone (0.909) do not recover this gain. This is consistent with Eq. 7: centering and $\sqrt{p_0}$ weighting jointly realize the Fisher metric $F = \mathrm{diag}(p_0) - p_0 p_0^\top$.

Full-distribution scoring. CoAx measures the complete output displacement rather than projecting it onto the answer direction. In the synthetic sweep of Figure 5c, β controls how strongly the planted backup effect aligns with that task direction. The CoAx AUC remains essentially unchanged as β decreases, whereas the conditional gradient deteriorates as the backup moves off the answer direction; intact-state scores remain near chance. The documented IOI setting also lies in the low-answer-alignment region of this controlled sweep ($\beta \approx 0$; Appendix C.4.2). Thus the controlled experiment shows that full-distribution scoring remains stable as a planted conditional effect rotates away from the task direction.

B EXPERIMENTAL SETUP

B.1 MODELS

We evaluate GPT-2 (Radford et al., 2019) (small 124M, medium 355M, large 774M), Pythia (Biderman et al., 2023) (160M, 410M, 1.4B), GPT-Neo-1.3B (Black et al., 2021), Gemma-2 (Gemma Team et al., 2024) (2B, 9B), OLMo-2-7B (Team OLMo et al., 2025), Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen2.5-7B-Instruct (Qwen et al., 2024). All checkpoints remain frozen and no method updates model weights. The cross-model panel uses `Qwen/Qwen2.5-7B-Instruct` and `meta-llama/Meta-Llama-3.1-8B-Instruct`; exact checkpoint identifiers and software revisions are provided in the anonymized supplementary archive.

B.2 CIRCUIT ROLES AND PRIMARY SETS

For GPT-2-small IOI, we use the documented head roles of Wang et al. (2023): name-mover, backup name-mover, S-inhibition, induction, duplicate-token, and previous-token heads. The main backup experiment takes the three primary name-movers as $\mathcal{S}$, the eight documented backups as positives, and all remaining heads as candidates. For cross-model induction analyses, induction, duplicate-token, and previous-token roles are detected from attention offsets on repeated-random sequences (Olsson et al., 2022): previous-token heads attend to $p-1$, duplicate-token heads to $p-T$, and induction heads to $p-T+1$ for repeated-sequence length T. These detected roles define primary sets when documented cross-model labels are unavailable.

B.3 DATA

IOI prompts instantiate "When {A} and {B} went to the {place}, {B} gave a {object} to" using the single-token pools of Wang et al. (2023); the indirect object (IO) appears once and the subject (S) twice. The main discovery protocol uses 96 sampled fillings, and the structural attention summary in Figure 8 is averaged over the same prompts. Two additional surface templates test prompt-form robustness in Appendix C.5.2. Induction experiments use uniformly random token sequences repeated once and score repeated-token log-probability on the second copy. WikiText-2 windows (Merity et al., 2017) provide unlabeled calibration data for cross-scale geometry analyses. Protocol constants and prompt seeds are in Table 6.

B.4 METRICS

For IOI, the behavioral readout is the answer margin $m = z_{\mathrm{IO}} - z_{\mathrm{S}}$. Backup recovery uses head-level ROC-AUC over the eight documented backups, excluding the supplied primary heads; average precision and precision at small k provide imbalance-sensitive checks. Section 5 reports Spearman rank agreement between each score and held-out freezing-based repair. Circuit completion uses the incompleteness gap from Section 4; knockout additionally measures IOI-margin distance and answer-position KL to the documented primary-plus-backup intervention, with unrelated-text KL tracking collateral change.

B.5 BASELINES

Intact-state scores. We compare single-ablation saliency, AtP (Syed et al., 2024) (gradient times activation), activation-space EAP-IG (Hanna et al., 2024; Sundararajan et al., 2017) with a zero head-output baseline and five integration steps, and AtP* GradDrop (Kramár et al., 2024) adapted to the same zero-ablation target. Following the respective published aggregation conventions, AtP averages per-example attribution magnitudes, whereas activation-space EAP-IG accumulates signed dataset scores and ranks by the magnitude of the aggregate. EAP-IG interpolates head outputs one attention layer at a time. For GradDrop, we follow the published blockwise residual-path decomposition and $L-1$ normalization, aggregating absolute per-example attributions. The QK correction instead targets attention saturation and is not applicable to the post-attention head-output nodes scored here. Task-gradient baselines use the IOI margin.

Conditional and structural controls. Conditional energy ranks by $\mathcal{E}(\delta z_{u|\mathcal{S}})$ using the same primary set, ablation machinery, and Fisher geometry as COAX, but omits the intact effect. Conditional gradient recomputes attribution after primary-set ablation, testing whether the intervened state alone exposes backups. We also compare co-activation and the weight-space projection kernel (Yamagiwa et al., 2026). A task-specific historical control ranks heads by the change in direct IO$-$S logit attribution after primary removal; because this mirrors the readout used to characterize the IOI backup branch, we report it separately in Appendix C.1.5. These controls isolate conditioning, the intact-to-removed contrast, and the scoring representation.

C BACKUP RECOVERY AND MECHANISTIC VALIDATION

This appendix expands the evidence behind Section 3 in the same order as the main text. We first report full backup-recovery results, then test whether recovery is specific to the removed primary set and whether the recovered branch causally carries repair. We close with diagnostics of conditional growth and robustness to reasonable protocol variation.

C.1 BACKUP RECOVERY AND STATISTICAL VALIDATION

Figure 6 gives the distributional view of the main recovery result. The documented backups lie near the bottom of the intact-state single-ablation ranking but move sharply upward under conditional growth, producing the inversion summarized by Table 1.

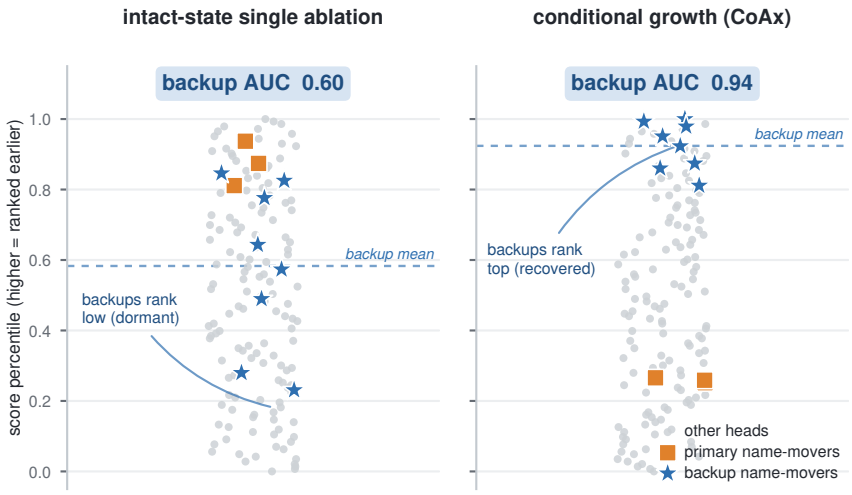

Figure 6: **The same backups under intact-state and conditional scoring.** Each point is one of the 144 GPT-2-small heads; the 141 non-primary heads are ranked for evaluation and the three primary name-movers are shown only for context. The eight documented backups (blue stars) rank near the bottom under intact-state single-ablation saliency (AUC 0.60) but near the top under CoAx conditional growth (AUC 0.94). Primary heads are excluded from the backup ROC-AUC.

C.1.1 RECOVERY ACROSS PROMPT SEEDS

The headline result is stable across the four declared prompt seeds rather than being driven by one prompt sample. CoAx leads every baseline on all four seeds and reaches 0.941 ± 0.004 backup ROC-AUC (Table 8). AtP* GradDrop remains the strongest intact-state attribution baseline and conditional gradient the strongest removed-only control on every seed.

Table 8: **Backup ROC-AUC across prompt seeds** (GPT-2-small IOI; 96 prompts per seed). CoAx leads on all four seeds and has the smallest spread.

score	s1	s15	s22	s8	mean ± std
single ablation	0.599	0.614	0.604	0.597	0.603 ± 0.007
AtP	0.633	0.656	0.656	0.655	0.650 ± 0.010
conditional gradient	0.759	0.770	0.768	0.772	$\underline{0.767 \pm 0.005}$
activation-space EAP-IG	0.623	0.573	0.582	0.547	0.581 ± 0.027
AtP* GradDrop	0.652	0.677	0.666	0.665	0.665 ± 0.009
conditional energy	0.757	0.764	0.758	0.755	0.758 ± 0.004
CoAx (ours)	0.945	0.944	0.934	0.943	$\mathbf{0.941}_{\pm 0.004}$

ROC-AUC measures the full ranking; top-k recovery gives the corresponding small-set view. The CoAx ranking contains $4/8$ documented backups in its top 8, $4/8$ in the top 10, $5/8$ in the top

15, and $6/8$ in the top 20. Because only $8/141$ candidates are positive, average precision gives a stricter top-of-ranking check: COAX reaches 0.557 ± 0.011 versus 0.098 ± 0.002 for AtP* GradDrop, with precision@8 of 0.50 versus 0.00. Across prompt seeds, the full COAX rankings have pairwise Spearman correlation 0.913 ± 0.032, with Jaccard overlap 0.852 ± 0.105 at top-8 and 0.806 ± 0.056 at top-20. The corresponding hypergeometric probabilities under random ranking are 2.8×10^{-4}, 8.1×10^{-4}, 3.2×10^{-4}, and 9.2×10^{-5}.

C.1.2 MATCHED CONTROLS AND STATISTICAL TESTS

Against the strongest intact-state attribution baseline, AtP* GradDrop, COAX improves backup AUC from 0.665 to 0.941. A DeLong test on seed-averaged head scores gives $p = 7.8\times10^{-5}$; bootstrap 95% CIs are $[0.885, 0.986]$ for COAX and $[0.471, 0.828]$ for AtP*. Treating each prompt seed as one paired observation gives a mean AUC gap of $+0.276\pm0.012$, $t_3 = 46.90$, $p = 2.1\times10^{-5}$. These tests quantify variation over prompts within this fixed model and benchmark rather than variation over models or circuits.

The more diagnostic comparisons hold the supplied primary set fixed. Conditional energy uses the same prompts, primary set, finite ablations, and Fisher geometry as COAX but omits the intact-state term; it reaches 0.758 ± 0.004, with DeLong $p = 3.8\times10^{-4}$ against COAX. Conditional gradient also scores the primary-ablated model but replaces the finite full-distribution effect with task-gradient attribution; it reaches 0.767 ± 0.005, with paired DeLong $p = 0.015$ against COAX. Thus reading the intervened state helps, but the intact-to-removed contrast supplies the larger gain.

Two exact non-parametric checks lead to the same conclusion. A $10{,}000$-shuffle label-permutation test places the observed AUC entirely outside the null distribution (null mean 0.50, 95th percentile 0.67, $p < 10^{-4}$), and the top-k hypergeometric tests above reject chance at every reported cutoff.

C.1.3 TWO-STATE GRADIENT CONTRASTS RECOVER THE CONDITIONAL SIGNAL

A natural alternative is to re-run an existing attribution method after removing the supplied primary set, or to take the same intact-to-removed difference of that attribution score. Table 9 tests both possibilities on the same 96 IOI prompts, four prompt seeds, primary set, and candidate pool used for backup recovery. Moving AtP or AtP* GradDrop to the primary-ablated model improves AUC from 0.650 to 0.767 and from 0.665 to 0.779, respectively. Taking the corresponding intact-to-removed differences raises recovery further, to 0.911 for AtP and 0.938 for AtP*, close to finite-displacement COAX at 0.941. This supports the central mechanistic claim: the informative signal is the change in causal responsibility across states, rather than importance in the removed state alone. The estimands remain distinct. The gradient contrasts are task-direction, local-linear approximations that require backward passes, whereas COAX directly scores the finite zero-ablation output displacement with forward passes, and its underlying effect change admits the exact interaction decomposition in Proposition 2.

Table 9: **Two-state gradient contrasts recover much of the same conditional signal.** Documented-backup ROC-AUC on GPT-2-small IOI (mean±std over four prompt seeds). "Conditional" scores the primary-ablated model; "two-state" subtracts the corresponding intact-state score.

score	intact	conditional	two-state
AtP (grad×act)	0.650 ± 0.010	0.767 ± 0.005	0.911 ± 0.002
AtP* GradDrop	0.665 ± 0.009	0.779 ± 0.004	0.938 ± 0.004
finite ablation displacement	0.603 ± 0.007	0.758 ± 0.004	$\mathbf{0.941}\pm0.004$

C.1.4 CO-ACTIVATION AS AN INPUT-SIDE CONTROL

Co-activation gives a complementary intact-state view. Ranking heads by the mean absolute correlation between their attention-output norms and those of the primary name-movers reaches 0.93 backup AUC, close to COAX's 0.941. This shows that the documented backups belong to a coherent co-firing neighborhood even before intervention.

The two scores nevertheless identify different properties. Co-activation records intact-state co-firing, whereas COAX asks whether a candidate's causal effect grows specifically after the supplied primary

set is removed. Freezing co-activation's selected set causes a larger raw IOI-margin loss than freezing the CoAx set (1.97 ± 0.06 versus 0.90 ± 0.07; Appendix C.3.4), but it gives substantially worse backup-role knockout matching (Appendix D.2). High label overlap or large raw freezing impact therefore does not by itself establish that a selector has isolated the missing causal role.

C.1.5 Historical direct-logit readout

The documented IOI backup branch was originally characterized using direct effects after name-mover knockout (Wang et al., 2023), so we separately test the corresponding task-specific ranking: the change in direct attribution to the IO$-$S logit direction after removing the primary set. As expected for a readout closely aligned with how the reference labels were characterized, it recovers those labels slightly better than CoAx (0.971 ± 0.003 versus 0.941 ± 0.004 ROC-AUC; average precision 0.620 ± 0.023 versus 0.557 ± 0.011). Its selected top-8 does not, however, reproduce the same causal role: it overshoots the freezing readout (1.62 versus 0.90 for CoAx), is farther from the documented knockout in IOI-margin distance (1.33 versus 0.99; paired difference $+0.344$, 95% CI $[+0.235, +0.450]$), and repeatedly includes the documented negative name-mover $[10, 7]$. Thus reproducing the historical label readout is not identical to recovering the missing branch's causal role.

C.1.6 Candidate-depth control

All eight documented backups lie in layers 9–11, making depth a strong structural prior. We therefore repeat the comparison under the *headline* last-position/full-vocabulary protocol while restricting the candidate pool to that late-layer band. CoAx remains at 0.909 ± 0.002 AUC within layers 9–11, compared with 0.691 ± 0.011 for intact single-ablation saliency; its score–layer correlation is also only $+0.114\pm0.011$, smaller in magnitude than for the other high-scoring structural and attribution selectors (Table 10). Thus the recovery is not explained by a preference for late-layer heads.

Table 10: **Candidate-depth control under the headline protocol.** The first column restricts candidates to layers 9–11; the second reports Spearman correlation between score and layer over the full candidate pool.

selector	L9–L11 AUC ↑	ρ(score, layer)
CoAx	0.909 ± 0.002	$+0.114\pm0.011$
conditional direct-logit change	0.887 ± 0.016	$+0.175\pm0.017$
co-activation	0.776 ± 0.010	$+0.436\pm0.003$
IO-attention structural read	0.788 ± 0.004	$+0.483\pm0.002$
AtP* GradDrop	0.739 ± 0.004	-0.309 ± 0.011
single ablation	0.691 ± 0.011	-0.328 ± 0.008

C.1.7 A primary-removal-free reference construction

The eight documented backups are a useful external reference, but their original characterization involved name-mover knockout (Wang et al., 2023). To verify that the intact-state blind spot is not an artifact of that labeling route, we construct a second GPT-2-small reference set without observing any primary-removal response. We first identify a name-mover family using two intact structural criteria: final-position attention to the IO name and an OV copy score that tests whether the head writes a name representation back toward the same token. Within this family, the three largest intact direct-logit contributors define the primaries; candidates within a 0.75 guard band of the weakest primary are excluded, and the remaining five weak contributors form the dormant reference set. This construction recovers exactly the three documented primary name-movers; three of its five dormant members are documented backups, while the remaining two are $[11, 1]$ and $[11, 6]$. Varying the family size from 12 to 28 and the guard band from 0.70 to 0.90 yields 20 constructions that all pass the independent causal check; across them, CoAx remains at 0.970–0.992 ROC-AUC while intact single ablation remains at 0.394–0.483.

The resulting labels also pass an intervention-based check that is not used in their construction. Against controls matched exactly in layer and closely in intact ablation energy, the dormant set has larger output-norm wake-up (1.159 versus 1.091), direct-logit hand-off ($+0.187$ versus $+0.047$),

Table 11: **Recovery of a primary-removal-free dormant reference set** (GPT-2-small; five positives, four prompt seeds). The reference labels use only intact structural reads and intact direct contribution; no quantity measured after primary removal enters their construction.

score	ROC-AUC ↑	average precision ↑
single ablation	0.444 ± 0.004	0.036 ± 0.000
co-activation	0.955 ± 0.001	0.312 ± 0.004
conditional direct-logit change	0.985 ± 0.003	0.599 ± 0.050
CoAx	0.983 ± 0.001	0.586 ± 0.006

conditional causal drop ($+0.197$ versus -0.062), and freeze-carried repair ($+0.514$ versus $+0.134$). Thus the conditional recovery signal persists when the positive set is defined independently of primary removal, while the two independently surfaced non-documented heads $[11, 1]$ and $[11, 6]$ receive convergent structural and intervention evidence.

C.2 Recovery Is Specific to the Removed Primary Set

Section 3.2 asks whether the recovered ranking follows the identity of the removed circuit rather than generic intervention severity. We test this with alternative documented and matched primary sets, then remove LayerNorm rescaling as a potential source of background growth.

C.2.1 Alternative and geometry-matched primary sets

Table 12 varies the supplied primary set while keeping the backup labels and evaluation protocol fixed. Matching behavioral damage, depth, or task relevance does not reproduce the ranking induced by the documented name-movers. The correct set gives 0.941 ± 0.004 backup AUC, whereas a damage-matched random set gives 0.426 ± 0.129, a same-layer control 0.536 ± 0.196, and documented S-inhibition and induction sets 0.688 ± 0.010 and 0.359 ± 0.017. Adding off-circuit heads to the correct set degrades recovery gradually from 0.92 with one contaminating head to 0.74 with four.

Table 12: **Backup recovery follows primary-set identity.** GPT-2-small IOI backup AUC under alternative supplied primary sets (mean±std over four prompt seeds; contamination entries are means). Matching behavioral damage or depth does not reproduce the documented backup ranking, while contamination of the correct set degrades recovery gradually.

primary set $\mathcal{S}$	joint damage	backup AUC
documented name-movers	$+0.284$	$\mathbf{0.941}$ ± 0.004
damage-matched random	$+0.295$	0.426 ± 0.129
random	-0.152	0.492 ± 0.113
same-layer, different heads	$+0.085$	0.536 ± 0.196
documented S-inhibition set	$+1.309$	0.688 ± 0.010
documented induction set	$+1.336$	0.359 ± 0.017
true set $+1/+2/+3/+4$ off-circuit heads	—	$0.92/0.84/0.81/0.74$

We make the negative control harder by matching wrong three-head sets in the score's own output geometry. From $5{,}000$ random triples scored on calibration prompts, we retain the 16 closest to the documented primary set in behavioral damage, answer-position Fisher displacement $\mathcal{E}(\delta z_{\mathcal{S}})$, and mean layer index, then evaluate them on held-out prompts. The matched wrong sets remain below chance on average: 0.397 ± 0.133 and 0.405 ± 0.129 AUC on two held-out prompt sets, versus 0.944 and 0.934 for the documented primaries. Even the closest behavioral-damage matches score only 0.319 and 0.413 on the two evaluations. Matching both behavioral damage and the output displacement used by the score therefore does not reproduce the backup ranking.

C.2.2 LayerNorm freezing

Primary-set ablation also changes downstream normalization statistics and can rescale the remaining computation. We isolate this channel by recording the intact-state centering mean and normalization

Table 13: **Geometry-matched primary-set control.** Backup AUC on two held-out prompt sets after selecting 16 wrong triples from a 5,000-set search, matched to the documented primary set in behavioral damage, Fisher displacement, and depth. For the matched controls, mean±std is across the selected triples.

primary set	held-out A	held-out B
documented name-movers	0.944	0.934
geometry-matched wrong sets	0.397±0.133	0.405±0.129

denominator for every GPT-2-small LayerNorm and replaying ablated forwards with these statistics partially or fully fixed. A *scale freeze* fixes the denominator, while a *full freeze* also fixes the centering mean. Replaying the recorded statistics on an otherwise intact forward reproduces the original logits to within 10^{-3} before either control is used.

Backup AUC rises from 0.941 ± 0.004 in the unmodified model to 0.983 ± 0.002 under scale freeze and 0.978 ± 0.002 under full freeze. Under scale freeze, the documented backups retain 0.81 ± 0.01 of their conditional growth ($0.076\pm0.003 \rightarrow 0.062\pm0.003$), while mean non-backup growth shifts from approximately zero to negative ($-0.001\pm0.001 \rightarrow -0.030\pm0.002$). Normalization rescaling therefore does not create the recovered backup signal; removing it suppresses background growth and sharpens the separation.

C.3 Causal Validation of the Recovered Branch

Backup-label recovery establishes ranking agreement, but the mechanistic claim requires the recovered branch to respond to primary-set failure and causally carry the resulting repair. We first situate that branch in the documented IOI circuit, then give the full measurements behind Figure 3 and an independent structural check.

C.3.1 Circuit context: the documented fallback branch

Figure 7 places the primary and backup name-movers inside the documented IOI information flow. The earlier duplicate-token, induction, and S-inhibition heads route the relevant name information toward the final position, after which primary and backup name-movers form parallel late write paths to the answer. The diagram provides circuit context; the intervention measurements below establish when the backup branch becomes causally active.

C.3.2 Hand-off and graded wake-up

Direct-logit hand-off. For Figure 3a, direct-logit attribution projects each head's residual-stream write onto $W_U[\mathrm{IO}] - W_U[\mathrm{S}]$ at the answer position using the final-LayerNorm statistics from that forward. We report branch totals because the figure's edge width represents the contribution of the whole branch. In the intact model, the three primary name-movers contribute $+2.519\pm0.169$, the eight documented backups $+0.567\pm0.048$, and the remaining 133 heads -0.773 ± 0.151 to the IO$-$S logit. After primary-set removal, the primary contribution is zero by construction, the documented backup branch rises to $+1.680\pm0.149$, and the remaining heads shift to $+0.392\pm0.141$. The IOI margin nevertheless changes only from 2.574 ± 0.083 to 2.290 ± 0.110. The redistribution inside the circuit is therefore much larger than the behavioral loss.

Graded wake-up. Figure 3b removes the three primary name-movers progressively in descending order of their single-head IOI effect. For each $k \in \{0, 1, 2, 3\}$, we measure the output-norm ratio and the additional IOI-margin drop caused by ablating a candidate after the first k primaries have already been removed. The documented backups grow monotonically on both readouts, whereas matched random heads remain approximately flat.

For the CoAx curve, each prompt seed selects its top-8 once using the full three-head primary set and holds that set fixed while k varies. The response therefore cannot be explained by reselecting heads at each intervention level. The recovered set follows the same monotone direction as the documented branch.

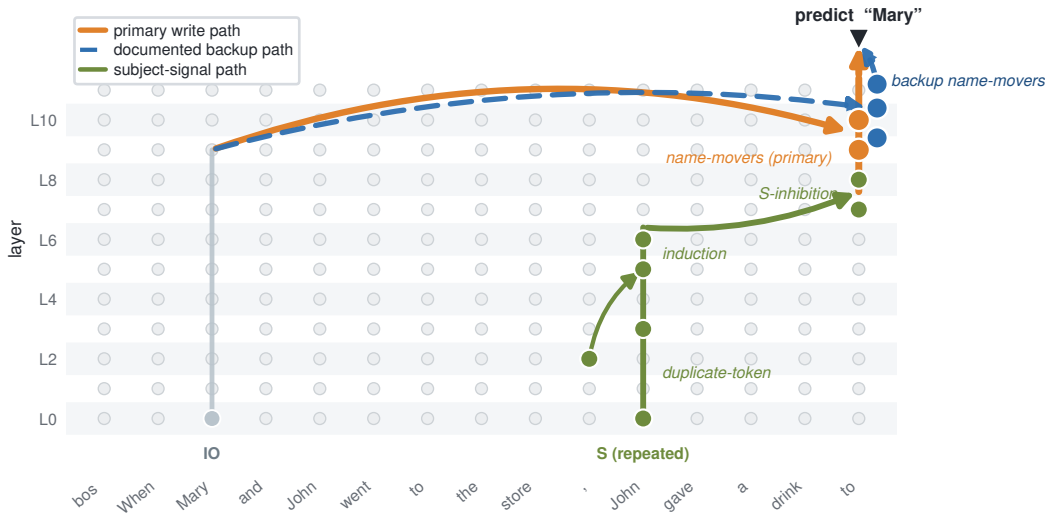

Figure 7: **The documented IOI fallback branch in token×layer context.** Each column is a prompt token and each row a layer; gray nodes denote the residual stream. Duplicate-token, induction, and S-inhibition heads carry the earlier subject-related signal (olive), while the primary name-movers (orange) and documented backup name-movers (blue, dashed) form parallel late write paths from the IO name to the answer position. The routes follow the documented IOI circuit (Wang et al., 2023); they are circuit structure rather than per-edge CoAx attributions. Figure 3 establishes that the backup branch is recruited after primary-set removal.

Table 14: **Graded wake-up behind Figure 3b.** Mean±std over four prompt seeds. The CoAx top-8 is selected once with the full primary set removed and held fixed across k; one matched-random set per prompt seed is likewise held fixed across the sweep.

	output-norm ratio			conditional causal effect		
k	documented	CoAx	random	documented	CoAx	random
0	1.000±.000	1.000±.000	1.000±.000	0.046±.001	0.112±.020	0.017±.022
1	1.031±.015	1.024±.012	1.001±.002	0.059±.013	0.114±.029	0.010±.020
2	1.082±.001	1.062±.005	1.004±.002	0.083±.003	0.118±.021	0.010±.026
3	1.150±.002	1.118±.008	1.006±.004	0.102±.006	0.136±.020	0.009±.036

C.3.3 Dose response to partial primary removal

The main experiment uses complete zero ablation of the primary set. As a continuous-strength check, we scale every primary head output by $1-\alpha$ for $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. The documented backups increase monotonically on both wake-up readouts as the intervention strengthens, while matched random heads remain essentially flat or move in the opposite direction (Table 15). This shows that recruitment is graded in intervention strength rather than appearing only at the endpoint ablation.

Table 15: **Dose response to partial primary removal** (GPT-2-small IOI; four prompt seeds). Wake-up is the output-norm ratio relative to the intact state; conditional drop is the additional IOI-margin loss from ablating a candidate after the partial primary intervention.

α	IOI margin	backup wake	random wake	backup cond. drop	random cond. drop
0.00	2.574	1.000	1.000	+0.046	+0.017
0.25	2.484	1.038	1.002	+0.061	+0.014
0.50	2.410	1.077	1.003	+0.075	+0.013
0.75	2.349	1.114	1.005	+0.088	+0.011
1.00	2.290	1.150	1.006	+0.102	+0.009

C.3.4 SELECTOR-SPECIFIC CAUSAL FREEZING

The graded curves establish intervention responsiveness; freezing tests whether that response is causally load-bearing. With the primary set already ablated, we freeze each selector's top-8 heads at their intact-state final-position outputs so that those heads cannot respond to the intervention. For IOI margin m, we report the freeze-induced loss

$$L_{\text{freeze}}(X) = m(\text{ablate } \mathcal{S}) - m(\text{ablate } \mathcal{S},\ \text{freeze } X). \tag{32}$$

We report this absolute quantity so every selector is compared under the same freezing intervention; it tests whether a selected set is load-bearing, not whether a larger value is a better circuit match.

Freezing the COAX-selected set causes a 0.90 ± 0.07 IOI-margin loss, compared with 0.03 ± 0.08 for conditional energy and 0.27 ± 0.37 for a size-matched random set. The documented backups themselves give 0.57 ± 0.09 under the same freezing intervention. Co-activation and AtP* produce larger raw losses, 1.97 ± 0.06 and 1.36 ± 0.03, respectively; this shows why freezing magnitude is a causal load-bearing diagnostic rather than a partition of total repair or a standalone circuit-matching criterion. Circuit completion and knockout in Appendix D provide the complementary specificity test.

Table 16: **Freeze-induced IOI-margin loss for each selector's top-**8 (GPT-2-small; mean$\pm$std over four prompt seeds). All rows use the same intervention: the selected heads are fixed at their intact-state outputs after the primary set has been ablated. "Overlap" is the mean number of documented backups in each selected set.

frozen set	IOI-margin loss	overlap
documented backups *(oracle reference)*	0.57 ± 0.09	$8.0/8$
COAX growth	0.90 ± 0.07	$4.0/8$
amplification-ratio variant	0.69 ± 0.13	$5.75/8$
conditional energy	0.03 ± 0.08	$1.0/8$
co-activation	1.97 ± 0.06	$3.0/8$
AtP* GradDrop	1.36 ± 0.03	$0.0/8$
random	0.27 ± 0.37	$0.25/8$

The overlap column is descriptive rather than an exhaustive precision measure: the documented eight define the known backup branch but do not imply that no other heads participate in compensation. The freezing readout therefore asks whether a selected set is causally load-bearing under primary failure; the completion and knockout tests ask the stricter question of whether it matches the missing branch's causal role.

C.3.5 INDEPENDENT STRUCTURAL CORROBORATION

The causal readouts above are intervention-based, so we additionally use a structural signal that COAX never observes. A name-mover's defining attention pattern is to attend from the answer position back to the indirect-object name (Wang et al., 2023). Reading this final-position QK attention without ablation, the documented backups attend to the IO token $5.8\times$ more strongly than other heads on average (0.19 versus 0.03). This structural score alone gives 0.96 backup ROC-AUC while correlating only weakly with COAX ($\rho = 0.09$ Spearman).

Figure 8 makes the structural check explicit. The primary and backup name-movers preferentially read the IO name rather than the repeated subject or filler tokens, while unrelated controls do not show the same pattern. Because this read uses the known name-mover token relation, we use it as independent post-hoc corroboration rather than as a general discovery score. Agreement between this ablation-free structural signature and the intervention-based evidence strengthens the mechanistic identification of the recovered branch.

C.3.6 A CONCRETE DORMANT BACKUP

The aggregate pattern above can be seen directly in the documented backup name-mover $[10, 6]$. In the intact model, its direct IO$-$S contribution is small ($+0.07$), yet from the answer position it attends strongly to the IO name (0.29), matching the structural role of a name-mover. After the

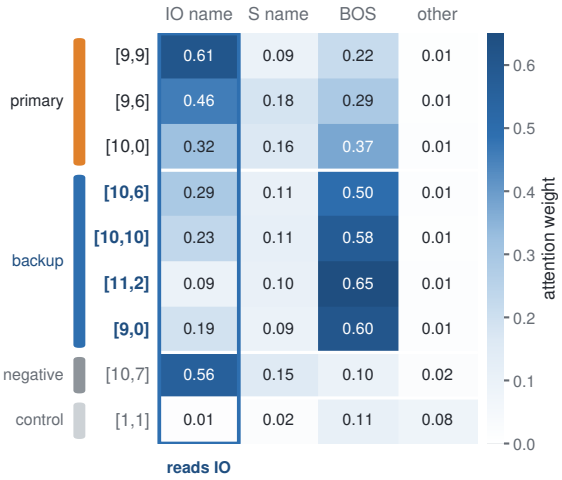

Figure 8: **Independent structural read of the name-mover branch.** Final-position attention to token roles, averaged over 96 IOI prompts. Documented primary and backup name-movers attend preferentially to the IO name, while control heads do not show the same pattern. This ablation-free structural read gives 0.96 backup ROC-AUC and correlates only weakly with the CoAx score ($\rho = 0.09$), providing complementary evidence for the recovered role.

primary name-movers are removed, its direct IO−S contribution rises to $+0.21$, while a random comparison head remains near zero ($+0.01$). This head-level view connects the structural role to intervention-induced recruitment; the set-level freezing experiment in Appendix C.3.4 provides the corresponding causal validation of repair.

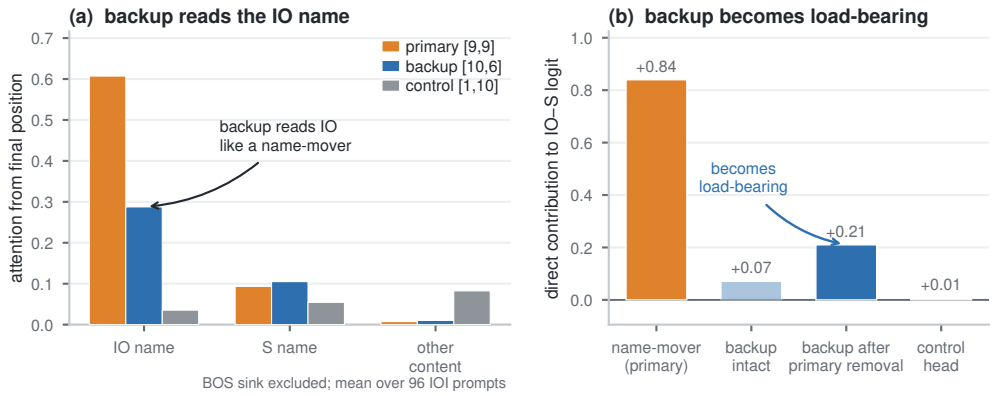

Figure 9: **A concrete dormant backup in GPT-2-small IOI. (a) Backup reads the IO name.** The documented backup $[10, 6]$ attends strongly to the IO name from the answer position, matching the structural signature of a name-mover. **(b) Its contribution grows after primary-set failure.** The backup's direct IO−S contribution increases from $+0.07$ in the intact state to $+0.21$ after the primary name-movers are removed, while a random comparison head remains near zero.

C.4 What Drives Conditional Growth

C.4.1 The Intact-to-Removed Contrast and Output Geometry

Table 17 separates two design choices: which states are compared and how output displacement is measured. The dominant gain comes from the state comparison. Intact-state single-ablation energy reaches 0.603 backup AUC; conditional energy in the primary-ablated model raises this to 0.758;

contrasting the two states reaches 0.941. Conditional energy alone retains components that are large in either state, whereas subtraction removes this pre-existing importance and isolates intervention-induced growth. A normalized alternative, $\mathcal{E}(\delta z_{u|\mathcal{S}})/\mathcal{E}(\delta z_u)$, reaches 0.984 AUC and retrieves all eight documented backups by rank 20. Its seed-averaged ranking is not statistically distinguishable from signed growth under paired DeLong ($p = 0.064$). We retain signed growth as the default because Proposition 2 applies directly to its underlying effect difference, it avoids unstable division by near-zero intact effects, and it transfers unchanged to the mechanism panel.

Holding the two-state comparison fixed, plain ℓ_2 displacement already reaches 0.905 AUC. Centering alone gives 0.903, probability weighting alone 0.909, and the full centered Fisher geometry 0.941. The remaining gain therefore belongs to the combined output geometry rather than either operation in isolation.

Table 17: **Component analysis of CoAx** (GPT-2-small; means over four prompt seeds). The top block varies the state comparison; the bottom block varies the output geometry while retaining signed growth.

score	backup AUC	recall@8	recall@20
state comparison			
intact-state energy, $\mathcal{E}(\delta z_u)$	0.603	0/8	0/8
conditional energy, $\mathcal{E}(\delta z_{u\|\mathcal{S}})$	0.758	1/8	3.0/8
amplification ratio (auxiliary variant), $\mathcal{E}(\delta z_{u\|\mathcal{S}})/\mathcal{E}(\delta z_u)$	0.984	5.75/8	8.0/8
signed growth (default), $\mathcal{E}(\delta z_{u\|\mathcal{S}}) - \mathcal{E}(\delta z_u)$	0.941	4.0/8	6.2/8
output geometry (signed growth throughout)			
plain ℓ_2	0.905	—	—
centering only	0.903	—	—
probability weighting only	0.909	—	—
centered Fisher (full)	**0.941**	—	—

C.4.2 Controlled test of dormancy and task alignment

To isolate two properties that the IOI benchmark does not vary independently, we use a controlled construction: whether an intact-state score can identify a near-dormant backup, and whether a task-direction gradient remains informative when repair is not aligned with that direction. This complements the empirical circuit experiments by separately controlling dormancy and task alignment.

Construction. We work in an $r = 192$ output space with unit answer direction e and Gaussian observation noise $\sigma = 0.05$. Four primaries have large intact-state effects and directions partially aligned with e. Each of 100 backups shadows a primary direction but is near dormant in the intact state ($a_b^{\text{intact}} \approx 0.04$) and becomes active after primary removal ($a_b^{\text{cond}} \approx 0.42$). A further 100 inert units have random directions, small intact-state effects, and no conditional change. Primaries are excluded from the classification task.

We compare intact-state energy, an intact-state answer-gradient proxy, the corresponding conditional gradient, and CoAx signed growth. The two intact-state scores sit at or below chance, whereas conditioning makes the planted backups identifiable (Table 18). CoAx reaches 0.90 ± 0.02 AUC and the conditional gradient 0.85 ± 0.04 in the default answer-aligned construction. Sweeping observation noise from $\sigma = 0.03$ to 0.16 reduces CoAx from 0.95 to 0.63, while the intact-state scores remain near chance.

Table 18: **Dormancy isolates the value of conditioning.** Controlled benchmark with 100 conditionally active backups and 100 inert units over 40 independent trials. Intact-state scores remain at or below chance, while both conditional scores recover the planted backups.

score	ROC-AUC ↑	DeLong vs. CoAx
CoAx growth	**0.90** $\pm$ .02	—
intact-state energy	$0.42 \pm .05$	$p < 10^{-15}$
intact-state gradient proxy	$0.51 \pm .04$	$p < 10^{-15}$
conditional gradient	$0.85 \pm .04$	n.s. (default alignment)

Task alignment. Let β denote the fraction of a backup's effect energy aligned with the answer direction. Holding the total conditional effect fixed while varying β leaves COAX at approximately 0.90 AUC, whereas the answer-direction conditional gradient falls from 0.95 at $\beta = 0.9$ to 0.59 at $\beta = 0.1$ (Table 19). The two conditional scores therefore answer different questions: conditional growth reads the full output displacement, while the gradient reads its task-aligned projection.

The documented IOI backups also lie in this low-alignment regime. Across the eight heads, the measured IO−S energy fraction is below 0.01 for every backup, with mean 0.002 and across-head standard deviation 0.003. We use this measurement only to place the empirical IOI setting on the synthetic sweep, not as a discriminator of backup identity; it is consistent with the empirical gap between conditional gradient and full-distribution conditional growth on IOI.

Table 19: **Full-distribution growth is insensitive to answer alignment.** Synthetic sweep over backup answer-alignment β (100 backups, 40 independent trials). At fixed effect norm, COAX remains stable while the answer-direction conditional gradient degrades as repair moves away from the task direction.

backup alignment β	COAX AUC	cond. grad. AUC	intact grad. AUC	DeLong p
0.9	0.90	0.95	0.58	cond. grad. higher (0.03)
0.5	0.90	0.87	0.52	n.s. (0.21)
0.3	0.90	0.75	0.49	$< \mathbf{10^{-3}}$
0.1	**0.90**	0.59	0.48	$< \mathbf{10^{-3}}$

C.4.3 INTERACTION STRUCTURE AND EMPIRICAL NON-ADDITIVITY

The conditional effect change has an exact all-order interaction interpretation, but COAX does not enumerate those terms. As a diagnostic, Figure 10 visualizes only the pairwise slice of the IOI circuit. The primary and backup name-movers form a strong interaction block, consistent with each candidate's effect depending on whether the primary branch remains available.

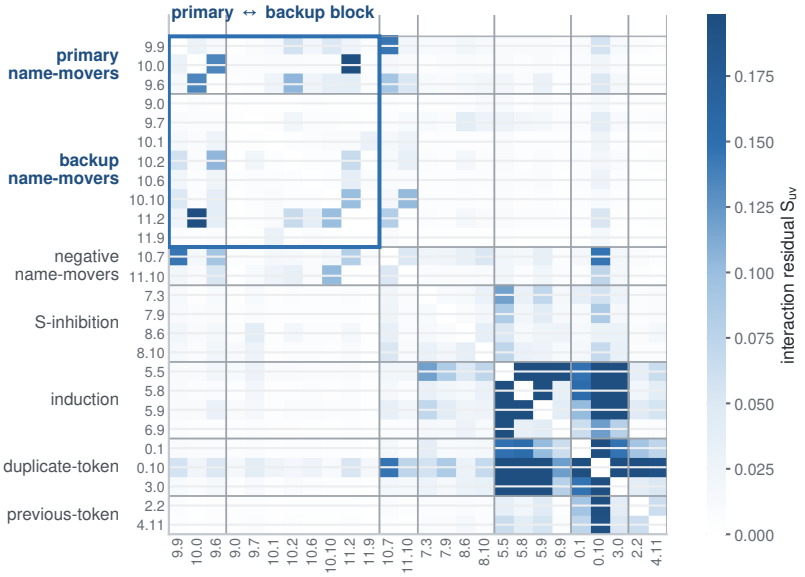

Figure 10: **Pairwise slice of the IOI interaction structure.** The magnitude of the pairwise co-ablation residual is shown for documented IOI circuit heads grouped by role. Primary and backup name-movers form a strong block (blue outline). This is only the order-two slice of the all-order aggregate in Proposition 2; COAX does not compute this matrix.

The same dependence appears at the module level. Jointly ablating the three primaries changes the IOI margin by only 0.28 because the backup branch remains available. Ablating the full primary-plus-backup module changes it by 1.30, almost twice the 0.67 predicted by summing the eleven

single-head effects and $4.6\times$ the primary-only effect. A size-matched random set is approximately additive. The redundant module therefore cannot be reconstructed from isolated intact-state effects, matching the conditional dependence characterized by Proposition 2.

Table 20: **Joint versus summed ablation** on GPT-2-small IOI (IOI-margin drop; intact margin 2.57; means over four prompt seeds). The primary-plus-backup module is strongly non-additive, whereas a size-matched random set is approximately additive.

ablated set	$\sum$ single	joint	ratio
3 primary name-movers	0.30	0.28	$0.9\times$
8 documented backups	0.37	0.30	$0.8\times$
primaries + backups	0.67	**1.30**	$\mathbf{1.9}\times$
random, size-matched	approximately additive		$\approx 1\times$

C.5 Robustness

C.5.1 Primary-set size

CoAx is conditioned on a supplied primary set, so we test whether strong recovery requires all three documented name-movers. We order the primary heads by their own ablation energy and evaluate the first one, two, or all three. Backup AUC is already 0.94 with one primary, rises to 0.97 with two, and is 0.94 with the full set. Together with the contamination experiment in Appendix C.2, this shows that recovery tolerates partial specification while remaining sensitive to off-circuit contamination.

Table 21: **Backup AUC versus supplied primary-set size** (mean over four prompt seeds). A partial documented primary set already supports strong recovery.

primary set	$\lvert\mathcal{S}\rvert{=}1$	$\lvert\mathcal{S}\rvert{=}2$	$\lvert\mathcal{S}\rvert{=}3$
documented primaries	0.94	**0.97**	0.94

C.5.2 Prompt-template robustness

The main IOI experiment uses one surface template. Recomputing backup recovery on two alternative forms that exercise the same indirect-object name-moving behavior ("After … arrived at the …, …handed a … to" and "While … were at the …, …passed a … to") gives AUC 0.93 and 0.92, versus 0.94 for the standard template. The recovery therefore persists across prompt forms rather than depending on one surface realization.

D Conditional Circuit Completion and Knockout

This appendix expands Section 4 at the circuit level. We first report completion and matched controls for the headline CoAx ranking, then give the full knockout comparison against the documented backup intervention. Recovery, freezing, completion, and knockout all use this same ranking; no backup labels, task gradients, or task-direction filters enter the CoAx selection.

D.1 Circuit-Completion Controls

The incomplete primary circuit has incompleteness 0.474 ± 0.030, while the documented circuit reaches 0.300 ± 0.027. Using the same headline ranking as discovery, the CoAx top-8 reduces incompleteness to 0.236 ± 0.037; a size-matched random top-up gives 0.499 ± 0.073 (Table 22). Conditional energy attains the smallest scalar gap (0.181 ± 0.027), but its poor backup recovery and knockout match show that response magnitude alone need not identify the missing branch. The headline CoAx ranking therefore provides a direct completion result without introducing any task-direction selection step, while the knockout analysis below tests branch identity.

Prompt-level paired-bootstrap CIs for the CoAx reduction relative to the primary-only and random completions are $[-0.266, -0.211]$ and $[-0.300, -0.226]$, respectively. This uses the discovery rank-

Table 22: **Circuit completion with the headline ranking** (GPT-2-small; mean±std over four prompt seeds; lower is better). The CoAx row uses the same last-position/full-vocabulary ranking as backup discovery, with no task-direction filter.

circuit	incompleteness ↓
primary circuit *(no completion)*	0.474±0.030
+random	0.499±0.073
+conditional energy	**0.181±0.027**
+CoAx	0.236±0.037
+amplification-ratio variant	0.258±0.035
+AtP⋆ GradDrop	0.309±0.031
+CoAx next-8	0.342±0.040
+co-activation	0.403±0.038
documented circuit *(reference)*	0.300±0.027

ing itself: aside from the supplied primary set, selection uses neither backup labels nor task-direction information.

D.2 Knockout Against the Documented Backup Intervention

We next reverse the completion test. Each selector contributes eight heads, which are removed together with the primary name-movers and compared with removal of the documented primaries plus documented backups. The random row averages 20 independently sampled size-matched sets per prompt seed. We measure the mean absolute difference in IOI margin, KL divergence from the documented knockout at the answer position, and KL from the intact model on unrelated text. Task accuracy alone would give the wrong ordering: the documented knockout has IOI accuracy 0.714 and CoAx 0.779, whereas the CoAx next-8 and AtP⋆ selections drive accuracy much lower, to 0.552 and 0.180, despite lying farthest from the documented intervention. This inversion is why we judge knockout by reproducing the documented causal intervention rather than by maximizing task damage.

Signed CoAx and the amplification-ratio variant are tied in IOI-margin distance in Table 23 (0.99 each); their paired difference is 0.003 with 95% CI $[-0.087, 0.095]$. The ratio variant is closer in output KL, whereas signed growth remains well defined at near-zero intact effects, is directly tied to the exact interaction decomposition, and is used uniformly across the broader mechanism panel. Conditional energy, the CoAx next-8, AtP⋆, co-activation, and random removal are all farther from the documented intervention in margin distance. Thus the completion and knockout tests distinguish branch matching from merely maximizing response magnitude or task damage.

Table 23: **Knockout relative to the documented backup intervention** (mean over four prompt seeds). Each row removes the primary set plus eight selected heads. Lower values indicate closer reproduction of the documented knockout; unrelated-text KL measures collateral change.

top-up	margin distance ↓	output KL ↓	unrelated KL ↓
documented backups *(oracle)*	0.00	0.00	0.09
CoAx	**0.99**	0.17	0.14
amplification-ratio variant	**0.99**	**0.07**	**0.09**
conditional energy	1.71	0.46	0.47
co-activation	1.79	0.47	0.15
CoAx next-8	2.03	1.38	0.86
AtP⋆ GradDrop	2.70	1.02	0.47
random	1.24	0.47	0.24

Paired bootstrap intervals over the 384 prompt-level observations support the margin-distance ordering: relative to CoAx, co-activation is $+0.799$ farther from the documented knockout (95% CI $[+0.655, +0.941]$), AtP⋆ is $+1.712$ farther ($[+1.545, +1.883]$), and random is $+0.248$ farther ($[+0.158, +0.336]$).

E CONDITIONAL REPAIR BEYOND DOCUMENTED BACKUPS

This appendix expands the two generalization tests in Section 5: repair alignment across mechanism clusters and held-out circuit completion across model families.

E.1 REPAIR PREDICTION ACROSS MECHANISMS

The intervention panel contains twelve instances formed by three surface templates crossed with four primary sets: name-mover, S-inhibition, induction, and duplicate-token. Name-mover and S-inhibition instances use 96 IOI prompts for scoring and a disjoint set of 96 prompts for the freezing-based target. Induction and duplicate-token instances analogously use disjoint sets of 32 repeated-token sequences. The former use the IOI margin and the latter repeated-token induction behavior, so each causal target matches the behavior disrupted by its primary set.

The causal target is the continuous quantity used in Section 5, $\mathrm{repair}(u \mid \mathcal{S}) = m(\text{ablate } \mathcal{S}) - m(\text{ablate } \mathcal{S}, \text{ freeze } u)$. Agreement is measured by Spearman correlation across candidate heads. Signed conditional growth is positive in $11/12$ held-out instances and all four mechanism clusters.

E.1.1 HIERARCHICAL INFERENCE

The three templates within each primary set share the same model, head inventory, and mechanism, so we resample primary sets first and templates within them. The mechanism means are $+0.030$ for duplicate-token, $+0.193$ for induction, $+0.204$ for name-mover, and $+0.066$ for S-inhibition. Their equally weighted macro mean is $+0.123$, with nested 95% CI $[+0.047, +0.201]$. Dropping any one mechanism leaves the macro estimate between $+0.096$ and $+0.154$, so the positive agreement is not carried by a single cluster.

Table 24: **Agreement of CoAx with mechanism-matched causal repair** (Spearman ρ). Each mechanism value averages its three templates; the macro column weights the four mechanisms equally, and the interval resamples mechanisms first and templates within them.

score	dup.-token	induction	name-mover	S-inhib.	macro	nested 95% CI
CoAx	$+0.030$	$+0.193$	$+0.204$	$+0.066$	$\mathbf{+0.123}$	$[+0.047, +0.201]$

E.2 CONDITIONAL CIRCUIT COMPLETION ACROSS MODELS

We test the set-level claim behind Figure 4b under a held-out, equal-budget protocol. For each non-GPT-2 model, detection, selector calibration, and completion evaluation use disjoint repeated-token sequence sets: 32 sequences identify the top four induction heads as the primary set, every selector receives the same 16 calibration sequences, and completion is measured on 64 held-out sequences. No backup labels are used. Each selector contributes its top 10 remaining heads. Fisher scoring uses the full vocabulary through Gemma-2-2B and the fixed clean top-4096 support for Qwen2.5-7B, OLMo-2-7B, and Llama-3.1-8B. If $\Delta_{\mathrm{comp}}(C) = m(\varnothing) - m(\mathcal{S} \cup C)$ is the held-out induction-log-probability drop after adding completion C to the primary set, the reported completion gain is $\Delta_{\mathrm{comp}}(C)/\mathbb{E}_{C_{\mathrm{rnd}}}\Delta_{\mathrm{comp}}(C_{\mathrm{rnd}})$, where the expectation is over five independently sampled, size-matched random completions evaluated on the same held-out sequences.

Table 25 compares CoAx with intact-state single ablation, co-activation, AtP, AtP* GradDrop, and the role-matched next induction heads ("own"). CoAx exceeds matched random on all eight models and has a family-macro gain of 6.70; it also gives the larger completion than the role-matched own-head control on five of eight models. The $8/8$ above-random pattern is unchanged on an absolute rather than normalized scale: the smallest absolute gain is $+1.97$ log-probability, $16.5\times$ the corresponding random control's draw-to-draw standard deviation. Excluding the largest-gain model still leaves a $3.55\times$ family-macro gain over matched random, so the transfer result is not carried by that model alone.

Table 25: **Held-out, equal-budget cross-model conditional circuit completion** ($k = 10$). Entries are completion gain over a size-matched random completion; 1 therefore corresponds to the random reference. Detection, calibration, and evaluation use disjoint sequence sets, and every selector receives the same 16 calibration sequences.

model	COAX	single	co-act	AtP	AtP*	own
Pythia-160M	1.88	2.04	1.95	1.80	1.89	1.91
Pythia-410M	6.57	6.56	4.57	4.41	4.41	4.87
Pythia-1.4B	4.98	3.94	3.49	2.62	3.76	3.74
GPT-Neo-1.3B	4.60	8.76	3.99	7.68	7.68	6.94
Gemma-2-2B	2.99	2.74	0.82	2.12	2.12	2.50
Qwen2.5-7B-Instruct	22.46	3.15	1.46	22.36	22.36	3.68
OLMo-2-7B	2.80	2.55	2.35	2.31	3.52	4.05
Llama-3.1-8B-Instruct	2.89	3.82	3.82	2.11	2.11	1.84
family macro	6.70	4.20	2.63	6.59	6.86	3.75

F LIMITATIONS, BROADER IMPACT, ETHICS, AND REPRODUCIBILITY

F.1 LIMITATIONS

Documented backup labels are available only for GPT-2-small IOI (Wang et al., 2023), so direct backup-label recovery is evaluated in that setting; generalization beyond it is assessed through label-free causal targets across mechanisms and architectures. COAX conditions on a supplied primary set; automatically identifying that set is a separate problem, so the present work evaluates conditional circuit completion rather than joint primary-and-backup discovery. Our empirical validation is primarily at attention-head granularity, matching the documented backup circuit; extending conditional completion to MLP or feature-level units remains future work.

F.2 BROADER IMPACT

COAX is an analysis method for frozen models and does not train or modify model weights. Its intended benefit is more faithful mechanistic auditing: an intact-state analysis can appear to localize or remove a behavior while self-repair reroutes computation around the intervention. Exposing such conditional routes can reduce erroneous conclusions about circuit localization or removal. The method does not itself add model capabilities. The "x" in COAX is intentional: it reflects this open-ended downstream role, viewing mechanistic interpretability not only as an explanatory endpoint but also as an interface to auditing, editing, and other intervention-based analyses.

F.3 ASSETS, LICENSES, AND ETHICS

All experiments use publicly released pretrained checkpoints under their respective licenses: GPT-2 and GPT-Neo (MIT), Pythia, OLMo-2, and Qwen2.5-7B-Instruct (Apache-2.0), Gemma-2 (Gemma Terms of Use), and Meta-Llama-3.1-8B-Instruct (Llama 3.1 Community License). All checkpoints are used unmodified for inference-only analysis. Evaluation inputs comprise templated IOI prompts from Wang et al. (2023), procedurally generated repeated-token sequences for induction, and WikiText-2 windows (Merity et al., 2017) for unlabeled calibration. No human subjects, private data, or personally identifying information are involved; the IOI templates contain only common given names.

F.4 REPRODUCIBILITY

All model weights remain frozen and no training is performed at any stage. Appendix A specifies the complete COAX computation and protocol, Table 6 records the fixed protocol constants and random seeds, and Appendix B defines the checkpoints, data, metrics, and baseline settings. Measured wall-clock time and peak memory are reported in Table 4; the main experiments require only a single GPU and run from seconds to minutes per scan. The core GPT-2-small IOI discovery pipeline and its baselines also run on CPU in minutes, allowing the headline recovery result to be reproduced without GPU hardware.