CONTROLS ARE INTERVENTIONS: WHAT COMPLEMENT ABLATIONS ERASE, AND HOW TO AUDIT THEM

Anonymous authors
Paper under double-blind review

ABSTRACT

Circuit analyses often replace everything outside a circuit with mean, zero, or resampled activations, then knock out parts of it. This complement ablation is treated as a control but is an intervention, erasing each input's own activations and the replaced components' reactions to knockouts. We audit it against the full model on knockout effects and the head choices they induce. In three published GPT-2 circuits, for indirect object identification (IOI), acronyms, and greater-than, mean-ablating the non-circuit heads keeps clean accuracy within five points. Yet on every fresh input set, its per-input choice of the most damaging head pair misses more full-model effect than our tolerance allows. Freezing these heads at each input's clean activations keeps clean outputs exact and blocks only their reactions; it removes most of this error in IOI, about half elsewhere. Unfreezing the 48 non-circuit heads that attribution patching ranks highest passes in IOI and greater-than, not acronyms. In the original completeness test of the published IOI circuit, knockout sets found by greedy search have large gaps under mean ablation but small ones under freezing; a frozen-complement search finds the reverse. In IOI, freezing attention to the first token, the attention sink, removes over a third of the self-repair after zeroing the name-mover heads. Auditing the decisions a complement ablation supports, not its clean behavior, makes its faithfulness testable.

1 INTRODUCTION

A common way to explain a behavior of a neural network is to name a circuit, a small set of components claimed to carry the computation. The claim is tested by a complement ablation, which replaces everything outside the circuit with mean, zero, or resampled activations (Conmy et al., 2023; Hanna et al., 2024; Marks et al., 2025; Heimersheim & Nanda, 2024). Minimality and completeness tests then knock out parts of the circuit to see which components matter (Wang et al., 2023; Shi et al., 2024). One such test mean-ablates every head outside a circuit before knocking out its name-mover heads, which write the answer name to the output (Gong et al., 2026). In each case the complement ablation is treated as a control: when its knockout effects are read as the full model's, how the ablated components would have reacted is assumed not to matter.

A complement ablation is itself an intervention, and it does two things. It discards each input's own activations, a loss that clean metrics partly show. It also blocks the indirect effects of a knockout that run through the ablated components: whatever those components would have done in reaction, compensating for the knockout or amplifying it, is gone. No clean metric can reveal this, because nothing is knocked out on clean inputs. In mediation terms, the ablated model measures a direct effect of the knockout, with the ablated components held fixed as mediators, and the full model measures the total effect (Pearl, 2001; Vig et al., 2020; Mueller et al., 2026). The blocked part is the indirect effect through those components. Self-repair is the classic case where the direct and total effects differ (McGrath et al., 2023). Circuit completeness already compares a circuit with the full model under corresponding knockouts, and causal abstraction formalizes related requirements (Wang et al., 2023; Geiger et al., 2025); neither scores the component choices an analyst makes from the ablated model. We ask when the blocked indirect effects change those choices.

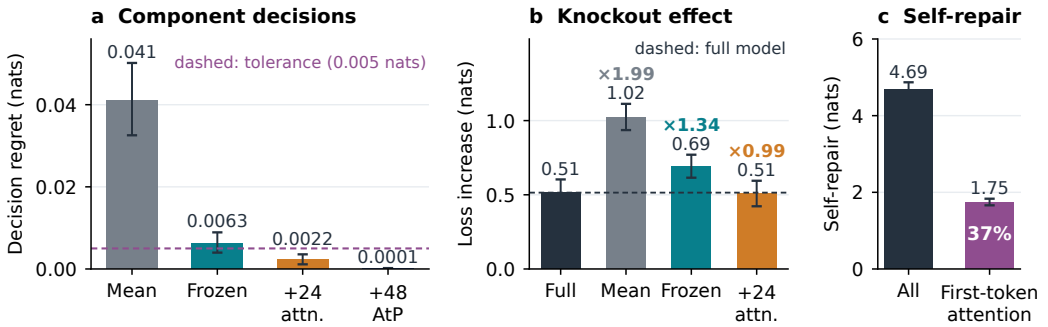
Both effects change which components an analyst would pick. In the GPT-2 circuit for indirect object identification (IOI), mean-ablating the 118 heads outside the 26-head circuit costs at most 4.7 points

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.

of clean accuracy across four held-out input sets. Yet on individual inputs its most damaging pair of eight circuit heads differs from the full model's about a quarter of the time. The full-model knockout effect these choices miss, the decision regret, is about eight times our tolerance of 0.005 nats (Figure 1a). Freezing the non-circuit heads at each input's clean activations removes most of this error. Still, about one choice in ten changes because of the blocked indirect effects alone. Unfreezing the 48 non-circuit heads with the highest attribution-patching (AtP) scores removes nearly all of the rest on two fresh sets. The 24 heads whose attention to the first token changes most pass on the first set but not on the second, so each choice of unfrozen heads must itself pass the audit. In two other published GPT-2 circuits the audit also fails mean ablation, and there freezing removes only about half of the error. In greater-than, mean ablation even changes the pair that is most damaging on average under zero knockouts (Section 3.4).

In the original completeness test of the IOI circuit (Wang et al., 2023), the knockout sets found by greedy search have large gaps under mean ablation but small ones under freezing, and a frozen-complement search finds other sets (Section 3.5). A mean ablation fitted per position, as in that study, also distorts a completeness-style measurement: in IOI, zeroing the name movers looks about twice as harmful as in the full model, and with the non-circuit heads frozen at their clean activations still about a third more harmful (Figure 1b). Unfreezing 24 non-circuit heads, selected on separate inputs by the change in their first-token attention or by AtP, brings the point estimate to 0.97–0.99 of the full model's. In the two other circuits, zeroing a group of circuit heads looks more harmful where the non-circuit heads' reactions offset the knockout and less harmful where they amplify it (Section 3.5).

To see what is blocked, we follow one channel in IOI that every attention head has: the change in its attention to the first token, the position known as the attention sink (Xiao et al., 2024; Gu et al., 2025). After the name movers are zeroed, freezing this change removes more than a third of the full model's self-repair (Figure 1c; Section 4.3).



GPT-2 IOI circuit. a: 256 inputs, 95% intervals. b, c: 128 inputs each, 97.5% intervals.

Figure 1: Mean ablation with similar clean accuracy still changes head choices and a completeness measurement, and the change in first-token attention carries part of the self-repair (GPT-2 IOI circuit). (a) Decision regret on the first fresh set (Tables 2–3): mean ablation leaves it about eight times the tolerance and freezing removes most of it; unfreezing the 24 heads whose first-token attention changes most, or the 48 with the highest attribution-patching (AtP) scores, brings it below the tolerance, but only the 48 also pass on the second set. (b) Loss increase from zeroing the name movers on 128 later inputs (Section 3.5; ×: ratio to the full model, dashed): mean ablation doubles it, freezing inflates it by about a third, and unfreezing the 24 heads of (a), ten of them downstream, brings its point estimate to 0.99 of the full model's. (c) Total self-repair (All) and the part removed by freezing the first-token attention of the 23 heads downstream of the zeroed name movers, 15 of them non-circuit (37%; Section 4.3).

Our contributions are:

- **An audit of complement ablations** on the knockout effects and component choices they induce, with freezing to separate their two effects. Mean ablation fails it in three published circuits, and unfreezing the 48 non-circuit heads ranked highest by attribution patching passes in two (Sections 2 and 3.1–3.4).
- **The standard completeness test, re-run.** We re-run the completeness test of Wang et al. (2023) as published and separate its signed gaps into replacement and blocked-reaction terms. The complement changes which sets greedy search finds; comparisons over random subsets and head classes still favor the published circuit over the naive one (Section 3.5).
- **One traced channel.** In IOI, freezing the change in first-token attention removes more than a third of the self-repair after the name movers are zeroed. In the non-circuit heads, this change is linked to part of the mean ablation's regret (Section 4).

2 Auditing a complement ablation

2.1 Knockout effects and their error

We compare an ablated model b, defined by which internal activations are replaced and by what values, against the **full model** N; a **complement ablation** replaces the head outputs outside a circuit. A **knockout** ablates a head subset S with one of two ablation types o: zero ablation or resample ablation. Resample ablation substitutes the head's full-model output on another input, rescaled, unlike standard resampling, to the norm of the output it replaces in the same forward pass. In the circuit example, b ablates the heads outside the circuit and S selects circuit heads to knock out. The full and ablated models receive the same knockout and, for resampling, the same donor input; parameters stay fixed, so any change in other components is downstream recomputation, not learning.

Let $L_b(x, S, o)$ be the answer-token negative log-likelihood (NLL) for input x, and $L_b(x, \varnothing)$ the same model without the knockout. The knockout effect is

$$e_b(x, S, o) = L_b(x, S, o) - L_b(x, \varnothing).$$

For a fixed collection $\mathcal{U}_D$ of knocked-out subsets (in Section 3, the candidate pairs), the mean absolute error (MAE) of the knockout effects is

$$D_b = \mathbb{E}_{x,o}\left[\frac{1}{|\mathcal{U}_D|}\sum_{S\in\mathcal{U}_D} |e_b(x, S, o) - e_N(x, S, o)|\right].$$

Subtracting each model's own clean loss separates a mismatch in knockout effects from a constant difference in clean loss; a lower D_b means closer agreement with the full model's knockout effects, not better task performance. D_b is a per-input form of the completeness comparison of Wang et al. (2023), which compares a circuit with the full model under the same knockouts; regret, defined next, scores the decisions these effects induce.

2.2 Decisions and regret

An analyst uses the knockout effects to choose components. For each input and ablation type, model b selects the pair with the largest effect from a fixed collection $\mathcal{U}_P$,

$$S_b^*(x, o) = \arg\max_{S\in\mathcal{U}_P} e_b(x, S, o),$$

and we score that choice in the full model by its **decision regret**, the full-model knockout effect the chosen pair misses:

$$R_b = \mathbb{E}_{x,o}\left[\max_{S\in\mathcal{U}_P} e_N(x, S, o) - e_N(x, S_b^*(x, o), o)\right].$$

The MAE covers all specified knockouts; regret covers only the chosen pair, and agreement records whether that pair matches the full model's choice. Per-input choices of this kind are what per-prompt analyses, such as attribution graphs, make (Ameisen et al., 2025). We say an ablated model is within

tolerance when the 95% interval of its regret lies at or below 0.005 nats. This is about a 0.5% relative change in the answer's probability. Sections 3.3 and 3.4 show which verdicts depend on it (details in Appendix A).

2.3 Complement ablations and setup

The complement ablations we compare keep progressively less of what the non-circuit heads do (Table 1). Mean ablation substitutes means fitted per layer, position, and head on 32 clean inputs. **Freezing** substitutes each non-circuit head's own clean output on the same input, as path patching does for the heads off a patched path (Wang et al., 2023); it is therefore exact on clean inputs and blocks only the indirect effects. Heads left **unfrozen** recompute under the knockout while the rest stay frozen. We choose them on separate fitting inputs, so no ablation in Table 1 uses the knocked-out full model on the evaluation inputs. With the non-circuit heads as mediators, the last column of Table 1 gives the effect each ablation measures (Pearl, 2001): a natural direct effect holds the mediators at their clean values on the same input, a controlled direct effect at fixed values (Appendix C covers resample knockouts). Any complement ablation b that substitutes values computed without the knockout, whether means from any distribution, zeros, optimal constants, or activations from other inputs, blocks the same reactions. Its error splits as $e_b - e_N = (e_b - e_{\text{frozen}}) + (e_{\text{frozen}} - e_N)$; only the first term depends on the substituted values, so they can offset the second, blocked term but not remove it. The total effect is the reference whenever an ablated model stands in for the full one; a direct effect is the target only for questions about direct paths, as in path patching. Claims about the circuit alone, such as its sufficiency, concern the ablated model itself and need no audit.

Table 1: Each complement ablation keeps less of what the non-circuit heads do and so measures a different effect; "+ unfrozen heads" freezes the other non-circuit heads and lets the selected ones recompute; its path-specific effect avoids the frozen heads.

Model	Clean activations	React to the knockout	Effect measured
Full model	input's own	all	total effect
+ unfrozen heads	input's own	unfrozen heads only	path-specific effect
Frozen	input's own	none	natural direct effect
Mean	fitted means	none	controlled direct effect

We study GPT-2 small in float32 (Radford et al., 2019) with the 26-head IOI circuit of arXiv v1 (Wang et al., 2023), ablating the non-circuit heads at all layers and positions; heads are indexed from zero (L9H9 is layer 9, head 9). In Sections 3.1–3.3 the knockouts cover all 28 pairs of eight candidate circuit heads under both ablation types, and each input contributes one choice per ablation type. Pairs are the smallest knockouts in which removing one head can change another's effect, as with backup heads. The eight are the name movers L9H9, L9H6, and L10H0, the S-inhibition heads L8H6 and L7H9, the induction head L5H5, the duplicate-token head L3H0, and the negative name mover L10H7. Apart from the original procedures re-run in Appendices O and P, every reported result uses inputs disjoint from those used to fit means or rank heads; the numbers of unfrozen heads were compared on the first set and re-tested on the second (Section 3.3). Means and head rankings are fitted on 32 inputs. The first evaluation uses 128 inputs (Sections 3.1–3.2), the audit tables use two fresh sets of 256 per circuit, the test of Section 4.2 uses 1,536 inputs, and the measurements shown in Figure 1b–c use sets of 128. Intervals are 95% bootstraps over evaluation inputs, or 97.5% for paired contrasts and the name-mover measurements of Sections 3.5 and 4.3 (Figure 1b–c), with the fitted means, head rankings, resample donors, and input generators held fixed (Appendices A–C). Ratios to the full model are point estimates.

3 Mean ablation changes component decisions and measurements

3.1 Similar accuracy, different decisions

Mean ablation of the 118 non-circuit heads, fitted on 32 clean inputs, keeps clean accuracy close to the full model (93.0% versus 94.5% on 128 evaluation inputs). Its clean loss rises from 0.89 to 1.12

nats, however, and its knockout effects change. The pair it selects matches the full model's choice in only 77% of cases (one per input and ablation type), and its regret is about eight times the tolerance. Two fresh sets of 256 inputs give the same picture, with 75% and 73% agreement and clean-accuracy gaps of 2.7 and 4.7 points (Table 2, Figure 1a). So do the 1,536 inputs of Section 4.2, with 76% agreement and a 3.3-point gap. Under zero or resample knockouts alone, mean ablation's regret on the two fresh sets is 0.035–0.049 nats (Appendix N). Zero and resample ablation of the complement are worse on clean accuracy (0.0% and 43.0% on the 128 evaluation inputs); our point concerns mean ablation as a proxy for full-model effects, not the original circuit's findings.

3.2 Freezing at each input's clean activations removes most of the IOI error

Freezing the non-circuit heads at each input's own clean outputs separates mean ablation's two effects: the clean computation stays exact, and only the indirect effects are blocked. Freezing removes most of the error: on the 128 evaluation inputs, the MAE falls by 71% and regret by 80%, and agreement rises from 77% to 91% (Appendix D.4). What remains comes from the blocked indirect effects alone. On 1,536 fresh inputs the frozen model's regret, 0.0079 nats, has its interval above the tolerance, and the chosen pair differs from the full model's in 11% of cases. This residual partly measures how incomplete the 26-head circuit is for these knockouts.

Table 2: Of these four IOI ablations, only unfreezing the 48 AtP-ranked heads keeps regret within tolerance on both fresh sets of 256 inputs (regret in nats with 95% intervals; agreement with the full model's pair choice and whether regret is within tolerance, each set 1 / set 2; both ablation types averaged within input). "+" rows leave the named non-circuit heads unfrozen and freeze the rest. "Within tolerance": yes if the interval lies at or below 0.005 nats, no if above, "?" if it crosses.

Complement ablation	Regret, set 1	Regret, set 2	Agreement (%)	Within tolerance
Mean	0.041 [0.033, 0.050]	0.045 [0.037, 0.053]	75 / 73	no / no
Frozen	0.0063 [0.0040, 0.0089]	0.0064 [0.0041, 0.0091]	90 / 90	? / ?
+24 by attention change	0.0022 [0.0011, 0.0036]	0.0038 [0.0020, 0.0059]	95 / 93	yes / ?
+48 by AtP	0.0001 [0.0000, 0.0002]	0.0003 [0.0000, 0.0007]	99 / 98	yes / yes

3.3 Unfreezing the right heads removes the rest in IOI

We rank the non-circuit heads by attribution patching (AtP; Syed et al., 2024; Kramár et al., 2024), a first-order estimate of how much freezing each head changes the answer loss, computed on the fitting inputs with each of the eight candidate heads zeroed in turn. We then unfreeze the top k heads, which recompute under the knockout. For comparison we also rank the heads by how much their first-token attention changes under the same knockouts (Section 4), and we test 8, 16, 24, and 48 heads.

Table 3: Regret is not monotone in the number k of unfrozen heads (first set, nats). * marks a 95% interval at or below 0.005 nats; details are in Appendix L.

Unfrozen heads k	0 (frozen)	8	16	24	48
Ranked by AtP	0.0063	0.0024*	0.0162	0.0065	0.0001*
Ranked by attention change	0.0063	0.0036	0.0028*	0.0022*	0.0237

Regret depends on k (Table 3). Ranked by AtP, the unfrozen heads bring regret within tolerance at 8 and 48 heads, but 16 heads more than double the frozen model's regret. Ranked by attention change, they lower it at every size up to 24 heads and nearly quadruple it at 48. On both sets, regret stays within tolerance from 32 AtP-ranked heads on and does not rise as heads are added. At 24 heads the attention ranking also has lower regret than random groups of 24 drawn from the 74 heads that can

react: all 50 on the first set and 46 of 50 on the second (Appendix N). The 15 non-circuit heads of layers 10–11 pass on neither input set (Appendix L).

The second input set re-tested the 24 attention-ranked heads, the 16 ranked by attention change, and the 48 ranked by AtP, with the layer 10–11 heads as a control, but not the top 8 by AtP. Of these groups, only the 48 AtP heads stay within tolerance there; the 24 attention-ranked heads cut the frozen model's regret by 41%, but their interval crosses the tolerance (Table 2; Appendix L). Each choice of unfrozen heads must therefore pass the audit itself. The failure of mean ablation and the pass of the 48 AtP heads hold for any tolerance between 0.001 and 0.025 nats, whereas the verdicts for freezing alone and for the 24 attention-ranked heads depend on the tolerance. Every selected head lies downstream of the knockouts, where 74 non-circuit heads can react, and unfreezing all 74 would reproduce the full model exactly. The 48 of them ranked highest by AtP suffice, whereas the 48 ranked highest by attention change do not. Random groups of 48 of them land near the tolerance, with 14 of 50 passing on each set, and the 48 AtP heads have lower regret than all 100 (Appendix N). Because frozen activations carry information from outside the circuit, a pass shows only that these heads' reactions suffice for these knockouts; it does not show that they are needed or that the circuit works on its own.

3.4 Mean ablation also fails in two more circuits

We repeated the audit in two other published GPT-2 circuits (Table 4; Appendix M, Figure 5). The acronym circuit of García-Carrasco et al. (2024) has eight heads, all of them candidates, four of which move the letters; its authors evaluated it by mean-ablating the heads outside it. The greater-than circuit described in the text of Hanna et al. (2023) has ten heads; the seven that feed MLPs 8 and 9 are the candidates. Each circuit has two fresh sets of 256 evaluation inputs, and the endpoint is the NLL of the answer, or of the set of valid years.

Table 4: Mean ablation fails the audit in all three circuits, and unfreezing the 48 AtP-ranked heads passes in IOI and greater-than. Decision regret (nats; set 1 / set 2 of 256 fresh inputs; * marks a 95% interval at or below the 0.005-nat tolerance, † one above it, and unmarked intervals cross it; clean accuracy: full model / mean ablation, set 1; intervals in Appendices L and M).

Circuit (heads)	Clean accuracy (%)	Mean	Frozen	+48 by AtP
IOI (26)	95.7 / 93.0	0.041† / 0.045†	0.0063 / 0.0064	0.0001* / 0.0003*
Acronyms (8)	91.4 / 94.5	0.070† / 0.092†	0.041† / 0.048†	0.0035 / 0.0036
Greater-than (10)	99.2 / 99.6	0.018† / 0.022†	0.0091† / 0.0093†	<0.0001* / <0.0001*

Mean ablation raises clean accuracy by 3.1 and 1.2 points for acronyms and changes it by less than a point for greater-than, yet it fails the audit on all four sets. It changes 17–20% of the acronym choices and 55–57% of the greater-than choices, where near-ties are common; for greater-than it also changes the pair that is most damaging on average under zero knockouts. Freezing removes only 42–58% of the regret, and the blocked indirect effects alone exceed the tolerance on every set. Unfreezing the 48 non-circuit heads with the highest AtP scores brings greater-than within tolerance on both sets; for acronyms it removes over 90% of the frozen model's regret, but the interval still crosses the tolerance. In both circuits these 48 heads have lower regret than each of 20 random groups of 48 downstream heads; in greater-than 18 of those groups also pass, and for acronyms none does. The 24 heads ranked by attention change fall within the range of random groups of 24 downstream heads. These verdicts depend more on the tolerance than in IOI: the 48 AtP heads would pass for acronyms at 0.0062 nats, and mean ablation fails greater-than only for tolerances up to 0.014 nats. The failure does not come from our choice of means: means fitted on up to 2,048 inputs, 17 further independent 32-input fits, and each original study's own replacement (node-level for greater-than) fail the audit on every set (Appendix N).

3.5 A DISTORTED COMPLETENESS MEASUREMENT AND ITS REPAIR

The standard completeness test of Wang et al. (2023) mean-ablates everything outside the circuit and greedily searches for the node set K with the largest gap $|F(C \setminus K) - F(M \setminus K)|$ in logit difference between the circuit and the full model. We re-ran it as published, changing only the complement: frozen at clean activations, it blocks only the complement's reactions (Appendix P). On two fresh sets of 250 inputs, the set the original search finds least complete scores 3.5–3.6, as large as the full model's logit difference, but 0.10–0.14 with the complement frozen; the frozen search's least complete set scores 1.3–1.4 there and 0.3 in the original test (Figure 6 in Appendix P). The study's naive circuit, which lacks the backup name movers, stays incomplete under freezing; the comparison holds for random subsets and head classes. A recent test with pooled means (Gong et al., 2026) also keeps its main conclusion under freezing (Appendix O).

With our circuit and per-position means, zeroing the name movers raises the answer loss about twice as much as in the full model, on each of four sets of 128 new inputs (Figure 1b shows one). Means fitted as Wang et al. (2023) did give 1.9 to 2.0 times in this measurement on the input sets of Table 2 (Appendix N). When the name movers are resample-ablated instead, the inflation is smaller, 1.3 to 1.4 times (Appendix K). The backup heads documented by Wang et al. (2023) become more important once the name movers are knocked out. Under mean ablation this added importance deviates from the full model's by 0.15 nats per head on average, and L10H1, whose full-model interval includes zero, gets one that lies above zero. Freezing the non-circuit heads at clean activations shrinks both distortions by half or more, though L10H1 still appears to gain importance. On the input sets of Table 2, mean ablation also scales the effects of zeroing the other six IOI head classes by 1.15 to 2.2, or for the previous-token heads flips the sign (Appendix N).

Only the 15 non-circuit heads of layers 10–11 lie downstream of the name movers, so unfreezing all of them reproduces the full-model measurement by construction. The test is whether heads selected on other inputs, from single-head knockouts rather than from this group knockout, also bring it close to the full model's. The 24 heads of Section 3.3 ranked highest by the change in first-token attention, scored with each of the eight candidate heads zeroed alone (name movers included), include ten of these 15. We ran them on the third and fourth sets and on a fifth with L9H9 alone zeroed. On all three they bring the knockout effect to 0.99 of the full model's, cut this deviation to about 0.007 nats per head or less, and recover the full model's backup heads. Twenty-four heads chosen by AtP, seven of them among the 15, bring it to 0.97–0.98 of the full model on the same three sets and also recover the full model's backup heads (Appendix K). On the two input sets of Table 2, the 48 AtP heads that pass the audit bring it to 1.02–1.03 of the full model (Appendix N).

In the two other circuits the same measurement is distorted in opposite directions. Under mean ablation, zeroing the four acronym letter-mover heads looks 1.5 times as harmful as in the full model, and zeroing the seven greater-than candidate heads only three quarters as harmful. Freezing, which blocks only the non-circuit heads' reactions, moves each measurement the same way (1.28–1.30 and 0.84–0.88 times): those reactions offset the acronym knockout and amplify the greater-than one. Unfreezing the 48 AtP heads brings both to 0.98–1.01 of the full model's (Appendix M).

4 ONE INDIRECT EFFECT TRACED: THE ATTENTION SINK

Which of the blocked indirect effects matter? In IOI, we follow one channel that every attention head has and that a single number measures: the change in the head's attention to the first token when upstream heads are knocked out.

4.1 FIRST-TOKEN ATTENTION AS A GATE

In the GPT-2 IOI inputs, position 0 is a prefixed end-of-text token. For one attention head at a later query, with $0 < a < 1$, write

$$z = av_0 + (1-a)u, \qquad u = \sum_{j>0} \frac{w_j}{1-a} v_j, \quad a = w_0.$$

Here w_j is the attention probability on visible position j, and v_j its value vector. The head's **first-token attention** a (Xiao et al., 2024; Gu et al., 2025; Guo et al., 2024) gates the other positions'

contribution by $1 - a$. Upstream knockouts can change this gate; fixing a head's output also fixes it. We isolate the gate by **freezing first-token attention** a at each input's clean value, at every query after position 0, while values and relative attention over other positions recompute.

4.2 The change in first-token attention is linked to part of the regret

On 1,536 fresh IOI inputs, where the mean ablation's regret is 0.040 nats, we patch the mean-ablated model (Figure 3 in Appendix H). To each non-circuit head's output we add the change that the shift in first-token attention produces in the full model under the knockout, $(a^d - a^c)(v_0^d - u^d)$, where c and d mark the full model's clean and knocked-out values. Regret falls by 14%, more than for three random patches that keep this patch's norm and its angle to the mean-ablated output but otherwise point in random directions. Because the patch uses the knocked-out full model, it diagnoses the error rather than repairing it. On queries at position 2 or later, where it can be matched in size with a patch built from the attention change at position 1, it barely lowers regret and cannot be distinguished from that patch. These tests link the change in first-token attention to part of the regret without showing that the first position is what matters. With the non-circuit heads frozen, letting only their first-token attention recompute does not detectably change regret (first set of 256; Appendix L).

4.3 Freezing first-token attention removes over a third of a known self-repair

IOI has a known self-repair mechanism: when the main **name-mover heads** are knocked out, documented backup heads become more important (Wang et al., 2023; Gong et al., 2026). On 128 new inputs we zero the three name movers, L9H9, L9H6, and L10H0, in the full model. Self-repair is the harm that downstream components absorb: the extra loss that appears when they are frozen at their clean values. Freezing the downstream components (the other 23 attention heads in layers 10–11, the layer 9–11 MLPs, and the final LayerNorm) raises the loss increase from 0.504 to 5.196 nats, so these components offset 4.692 nats. Freezing the downstream heads' whole outputs removes 95% of this self-repair; freezing only their first-token attention removes 1.748 of the 4.692 nats, or 37% (Figure 1c), while freezing their attention to position 1 instead has a small effect of opposite sign. The rest runs through the heads' values and their attention over other positions; single-channel shares need not add.

Eight of the 23 heads belong to the circuit, and an ablation of only the non-circuit heads leaves them free to react. The channel then keeps working: with every non-circuit head frozen at its clean outputs, freezing the circuit heads' first-token attention removes 36–37% of the self-repair measured in that model (Appendix K). This share comes from a different model, so it does not split the full model's 37% between circuit and non-circuit heads. Replacement models that freeze attention patterns in every head, such as the local replacement model of Circuit Tracing (Ameisen et al., 2025), block the channel in circuit heads as well.

5 Related work

Interventional evaluation. Ablation and activation-patching studies show that component-importance conclusions depend on the ablation, corruption, and metric (Miller et al., 2024; Zhang & Nanda, 2024). Li & Janson (2024) attribute the excess loss of zero, mean, and resample ablation over a loss-minimizing constant to spoofing. These studies compare ablations with one another; we hold clean behavior nearly fixed and compare an ablated model with the full model on the component choices it induces. Path patching freezes every head off a patched path at its clean value (Wang et al., 2023; Goldowsky-Dill et al., 2023), as our freezing does, and causal scrubbing resamples activations wherever a hypothesis says they should not matter (Chan et al., 2022). In both, the ablation is part of what is measured; we ask instead when an ablated model is a faithful proxy for the full one. Grant et al. (2026) show that intervention-induced divergence can be harmless or pernicious, and certified interventional fidelity adds anytime-valid confidence sequences (Asiaee, 2026). The error we audit needs no divergence: freezing substitutes each input's own clean activations yet still fails the audit in the acronym and greater-than circuits. Circuit Tracing validates feature influence against output effects and tests internal responses with attention frozen in both the local replacement and underlying models (Ameisen et al., 2025); we measure per-input regret against full-model knockout effects with attention recomputed.

Attention sinks and gating. Streaming attention exploits persistent attention to initial tokens (Xiao et al., 2024), and accounts of active and dormant heads, key biases, outlier rescaling, and sink-based gating explain why large attention mass need not mean a large direct value contribution (Guo et al., 2024; Gu et al., 2025; Qiu et al., 2026; Fesser et al., 2026; Fu et al., 2026). Head-specific output gating can train models without sinks (Qiu et al., 2025); we study how attention to the sink changes in a pretrained model that has one.

Self-repair and backup heads. The Hydra effect and later analyses establish downstream self-repair and examine its mechanisms (McGrath et al., 2023; Rushing & Nanda, 2024). Copy suppression attributes 39% of the self-repair after name-mover ablation, measured on heads' direct contributions to the logit difference, to the negative heads L10H7 and L11H10 (McDougall et al., 2024). Our share splits self-repair by attention to the first token, not by head, and is measured on the answer loss. Conditional co-ablation detects backup components missed by single ablation (Gong et al., 2026); we show that a completeness-style mean ablation distorts both the knockout effect and the backup heads' importance.

6 Implications and limitations

To infer full-model component importance from an ablated model, state what the ablation keeps, compare its knockout effects with the full model's, and check the component choices those effects induce. The audit is diagnostic, not a cheaper substitute for full-model knockouts: it is needed wherever effects measured in an ablated model are reported as the full model's, as in completeness tests and backup-head searches. When the audit fails, unfreeze the non-circuit heads downstream of the knockouts to be interpreted, starting with those AtP ranks highest, and audit those knockouts again, or else use full-model effects. For completeness tests, report a clean-frozen reference beside the original score: it separates gaps caused by substituted values from missing reactions, though freezing alone cannot show that a circuit suffices. Unfreezing can also raise regret, and for acronyms 48 heads did not suffice. A pass need not carry over to other knockouts: the 48 AtP heads also pass for single candidate heads and their triples in IOI and greater-than, but in IOI not for knockouts of the other circuit heads (Appendix N).

The self-repair role of the channel we traced depends on the knockout, and in Qwen2.5-0.5B (Qwen et al., 2025) it is not established (Appendices E and F). In two other GPT-2 IOI knockouts and in the acronym and greater-than circuits, the net reactions of all downstream components, MLPs included, amplify the knockout rather than offset it (Appendices F and M). Ranking heads by the change in first-token attention does not reliably beat random downstream heads for unfreezing outside IOI (Section 3.4), and Appendices I and J give its other limits.

These failures are existence results: they show that clean metrics cannot certify a complement ablation, not how often complement ablations fail in other models. They condition on GPT-2 small, three published circuits with our own input generators and fixed candidate heads, and NLL endpoints; under each study's own metric, mean ablation still changes a fifth or more of the choices in every circuit (Appendix N). Optimal ablation (Li & Janson, 2024) remains to be audited.

7 Conclusion

A control is an intervention. An ablation with similar clean accuracy can change which components look most damaging on individual inputs, how harmful their removal looks, which backup heads appear to take over, and which knockout sets look least complete. Freezing at each input's own activations removes part of the change, but not the part due to blocked indirect effects, which no clean metric can detect, and any choice of heads unfrozen to restore them must itself pass the audit. Explanations of the full model should audit the decisions they rest on, not clean behavior alone.

AI USE STATEMENT

We used generative AI tools for the following tasks with required disclosure: generating the synthetic induction, IOI, acronym, and greater-than inputs (Appendices B and M–O); developing the audit framework and the attention-sink analysis; formulating the definitions and estimands in Sections 2 and 4.1 and Appendices A, C, I, and J; proposing and refining hypotheses; designing the experimental protocols; implementing the ablations and knockouts and the evaluation pipeline; preparing and reformatting the natural-text inputs; interpreting results; and translating our non-English research notes and draft text into English. Proof-related tasks and qualitative or thematic data analysis are not applicable: this work contains no proofs and no qualitative data. Additionally, we used generative AI tools for tasks with recommended disclosure: figures, experimental parameters, software code and analysis artifacts, drafting and editing the manuscript and supplement, identifying, summarizing, and formatting references, brainstorming, and the paper's structure and title. AI tools carried out most implementation, experiment execution, analysis, and drafting under our direction; we set the research questions and each experiment's inputs and conditions, decided what to report, and reviewed every claim. We have reviewed all AI-assisted work: AI-written code was checked by the recorded numerical identity checks (Appendices G and P); the main estimates of the experiments in Appendices H–J were reproduced by independent re-analysis from saved raw rows; reported numbers were traced to the analysis files that produced them; and every reference was checked against its primary record. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix A defines the estimands, statistical units, bootstrap procedures, and acceptance rule; Appendix B lists the checkpoints and weight-file identifiers, input generators, natural-text preparation, and the fitting, calibration, and evaluation splits; Appendix C specifies each ablation and knockout, magnitude matching, probability clipping, and the adjusted-attention implementation; Appendix D gives the comparisons behind Sections 3.1–3.2 (Appendix D.4) and comparisons of local first-token attention ablations (Appendices D.1–D.3); Appendices E and F give the channel and scope tests cited in Section 6; Appendix G describes the numerical identity checks, software versions, and execution cost; Appendices H–J give the protocols, contrasts, and re-analysis checks for the test of Section 4.2 and for the prediction and pair-selection experiments cited in Section 6; Appendices K and L give the protocols, ablations, and contrasts for the self-repair measurement under complement ablation and for unfreezing selected heads; Appendix M gives the design, inputs, and results for the acronym and greater-than circuits; Appendix N gives the robustness checks of the audit, its transfer to other knockouts, and the decisions under each study's own metric; Appendix O re-implements a recent completeness test from its released code; and Appendix P re-runs the original completeness test of Wang et al. (2023). Where an experiment reuses evaluation inputs, Appendix B or its own appendix says so; Appendix N reuses the evaluation inputs of Tables 2 and 4, and Appendices O and P fit their means on the evaluated prompts, as the original tests do. We will release the code and the generated inputs with the camera-ready version.

REFERENCES

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Amir Asiaee. Certified interventional fidelity: Anytime-valid, adaptive evaluation of causal claims in mechanistic interpretability. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 210–227. PMLR, 2026. URL https://proceedings.mlr.press/v337/asiaee26d.html.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. AI Alignment Forum, 2022. URL https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing. Posted 3 December 2022.

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, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Lukas Fesser, Mozes Jacobs, Thomas Fel, Andy Keller, and Sham Kakade. A unifying view of attention sinks: Two algorithms, two solutions. *arXiv preprint arXiv:2606.08105*, 2026. URL https://arxiv.org/abs/2606.08105v1. Version 1.

Zizhuo Fu, Wenxuan Zeng, Runsheng Wang, and Meng Li. Attention sink forges native MoE in attention layers: Sink-aware training to address head collapse. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.01203v3. Version 3.

Jorge García-Carrasco, Alejandro Maté, and Juan Carlos Trujillo. How does GPT-2 predict acronyms? extracting and understanding a circuit via mechanistic interpretability. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3322–3330, 2024. URL https://proceedings.mlr.press/v238/garcia-carrasco24a.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023. URL https://arxiv.org/abs/2304.05969v2. Version 2.

Zhiren Gong, He Lu, Tiantong Wang, Yichi Zhang, Yixin Wang, Zihao Zeng, Ming Xiao, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits. *arXiv preprint arXiv:2607.01940*, 2026. URL https://arxiv.org/abs/2607.01940v3. Version 3.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/133e588e1429f9f1e25b215da145580e-Abstract-Conference.html.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f1b04face60081b689ba740d39ea8f37-Abstract-Conference.html.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs. *arXiv preprint arXiv:2410.13835*, 2024. URL https://arxiv.org/abs/2410.13835v2. Version 2.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL

https://proceedings.neurips.cc/paper_files/paper/2023/hash/efbba7719cc5172d175240f24be11280-Abstract-Conference.html.

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

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255v1. Version 1.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024. URL https://arxiv.org/abs/2403.00745v1. Version 1.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c55e6792923cc16fd6ed5c3f672420a5-Abstract-Conference.html.

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 *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3ba4d47a83e498c2b1a0868cba20f6de-Abstract-Conference.html.

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. *arXiv preprint arXiv:2307.15771*, 2023. URL https://arxiv.org/abs/2307.15771v1. Version 1.

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

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit evaluation metrics are not robust. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zSf8PJyQb2. The paper PDF and arXiv:2407.08734 title it "Transformer Circuit Faithfulness Metrics Are Not Robust".

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis. *Computational Linguistics*, 52(1):331–378, 2026. doi: 10.1162/coli.a.572. URL https://aclanthology.org/2026.cl-1.10/.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, 2001. URL https://arxiv.org/abs/1301.2300.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/904e89bb4e632e75fb47f093b620b257-Abstract-Conference.html.

Zihan Qiu, Zeyu Huang, Kaiyue Wen, Peng Jin, Bo Zheng, Yuxin Zhou, Haofeng Huang, Zekun Wang, Xiao Li, Huaqing Zhang, Yang Xu, Haoran Lian, Siqi Zhang, Rui Men, Jianwei Zhang, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. A unified view of attention and residual sinks: Outlier-driven rescaling is essential for transformer training. *arXiv preprint arXiv:2601.22966*, 2026. URL https://arxiv.org/abs/2601.22966v1. Version 1.

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. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115v2. Version 2.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

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.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/abccb8a90b30d45b948360ba41f5a20f-Abstract-Conference.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/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin 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 *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul. The 26-head circuit used here follows arXiv:2211.00593v1 (2022).

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5e5fd18f863cbe6d8ae392a93fd271c9-Abstract-Conference.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

A EVALUATION TARGET AND STATISTICAL UNITS

This appendix defines the evaluation target and the statistical rules used in Sections 2–4. Our target is the knockout effect of a specified full model, with its weights fixed. A complement ablation is evaluated as a surrogate for that effect. This target does not identify a unique human-interpretable circuit, and an ablation intended to answer a different causal question need not preserve it. In particular, retaining the clean outputs of attention heads outside a circuit provides additional input-specific information; improved faithfulness to full-model knockout effects under that ablation is not a demonstration that the isolated circuit is sufficient.

Let $L_b(x, S, o)$ be target negative log-likelihood (NLL, in nats) for input x, complement ablation b, ablation type o, and knocked-out attention-head set S. Each ablation has its own reference without the knockout; the ablation b remains applied:

$$e_b(x, S, o) = L_b(x, S, o) - L_b(x, \varnothing).$$

For a fixed collection $\mathcal{U}_D$, the MAE for input x and ablation type o is

$$D_b(x, o) = |\mathcal{U}_D|^{-1} \sum_{S \in \mathcal{U}_D} |e_b(x, S, o) - e_N(x, S, o)|,$$

where N denotes the full model. For a fixed set of candidate pairs $\mathcal{U}_P$, the surrogate chooses $S_b^*(x, o) = \arg\max_{S \in \mathcal{U}_P} e_b(x, S, o)$. Its decision regret is

$$R_b(x, o) = \max_{S \in \mathcal{U}_P} e_N(x, S, o) - e_N(x, S_b^*(x, o), o).$$

This regret measures the full-model loss increase missed by the selected head pair. It is not a generation failure rate. Pair agreement compares the surrogate and full-model argmax, with deterministic candidate ordering for ties. In the 28-pair evaluations, choices are made **separately for zero and resample knockouts**, and D, R, and agreement are then averaged over the two ablation types within each input. They are not obtained by first averaging two knockout-effect vectors and choosing one pair. Four-head experiments use all 15 nonempty head subsets for D, and their six size-two subsets for R. The larger effect-and-decision experiments use all 28 pairs of eight candidate heads for both quantities. The prediction experiment (Appendix I) instead evaluates local and joint prediction, and the selection experiment (Appendix J) fixes a single selected pair for evaluation.

Unless stated otherwise, synthetic induction and indirect-object identification (IOI) use the NLL of the designated next token; exceptions include the indirect-object–subject logit difference of Appendix D.5, the secondary margin noted in Appendix E, and the eight-token average of Appendix I. Natural-text experiments average next-token NLL over 111 targets at positions 17–127 of a 128-token window (zero-based indexing); positions 16–126 supply the corresponding logits. Clean accuracy, clean NLL, MAE, and decision regret are distinct measurements.

Unless a table specifies otherwise, intervals are paired percentile bootstrap intervals over 8,000 resamples of inputs or source documents. All branches, head subsets, and ablation types for a sampled input remain together. Primary contrasts generally use 97.5% intervals; branch estimates and tolerance evaluation generally use 95%. Exceptions include the three selection tests in Appendix J ($98\frac{1}{3}$% intervals) and Appendix I, which uses 95% intervals resampling both inputs and source-head groups, plus two Holm-corrected causal tests. Thus the many condition rows are not independent observations. Intervals condition on the chosen model, candidate heads, ablation family, fitted parameters, and task generator. They do not quantify uncertainty over possible circuits or architectures. Numerous auxiliary contrasts are descriptive and not jointly corrected. The endpoint check of Appendix D.5 uses its own family: two primary contrasts per endpoint at 98.75% (four-test Bonferroni), with the remaining intervals 95% descriptive.

The study's acceptance rule accepts a specified surrogate when its calibration regret has a one-sided 95% bootstrap upper bound at most $\tau = 0.005$ nats. On the separate evaluation set, a two-sided 95% interval is classified as supporting the tolerance if its upper endpoint is at most τ, violating it if its lower endpoint exceeds τ, and indeterminate otherwise. Calibration uses the 95th percentile of the bootstrap mean, whereas evaluation support uses the 97.5th percentile. Because these rules apply to separate samples, neither decision implies the other. Calibration acceptance, evaluation support, and improvement over a comparator are separate statements. Calibration selects and tests ablations; it

does not estimate the tolerance. A regret of 0.005 nats corresponds to about a 0.5% relative change in the answer token's probability. It is a project criterion, not an established community standard or a finite-sample certification guarantee, and the input bootstrap is not an anytime-valid certification method (Asiaee, 2026). With 32 calibration inputs, a heavy-tailed regret distribution can make the rule sensitive to one observation.

B MODELS, INPUTS, AND SEPARATION OF FITTING FROM EVALUATION

This appendix lists the models, inputs, and input splits used in Sections 2–4 (the acronym and greater-than inputs are in Appendix M). All reported main experiments use fixed pretrained weights (GPT-2 small, Radford et al., 2019; Qwen2.5-0.5B and -1.5B base, Qwen et al., 2025), evaluation mode, eager attention, and float32 arithmetic on CPU. The following are 12-character prefixes of the checkpoint revisions and weight-file hashes. Layer and head indices are zero-based throughout.

Fitting inputs determine ablation parameters; calibration inputs select or assess ablations; evaluation (test) inputs provide the reported effects. Qualification inputs are used for preliminary checks of full-model competence.

Table 5: Model checkpoints and weight-file identifiers.

Model	Checkpoint revision prefix	Weight-file SHA-256 prefix
GPT-2 small, 124M	`607a30d783df`	`248dfc391186`
Qwen2.5-0.5B	`060db6499f32`	`88c142557820`
Qwen2.5-1.5B	`8faed761d45a`	`a961db72e75d`

GPT-2 synthetic inputs prefix token 50256 (the end-of-text token) at position 0. Synthetic induction samples 32 distinct token IDs, appends a four-token subsequence, and asks for the token following that subsequence in the prefix. IOI uses a fixed grammatical family with two initially introduced names, a repeated subject, and a final indirect-object completion. Evaluation varies lexical choices and random seeds within the same grammatical family; it does not test held-out language constructions. Most synthetic qualification panels contain 32 separate inputs, with 80% full-model accuracy as the competence check. Test inputs are not filtered for correct full-model answers.

Natural-text evaluation samples WikiText-2 articles (Merity et al., 2017), prepares separate fit, calibration, and test documents from distinct articles, and excludes documents used for implementation checks. GPT-2 and Qwen use the same source-document identities but tokenizer-specific random 128-token windows; their windows and tokens need not be identical. Donors are a fixed cyclic derangement within each split, chosen so recipient and donor first tokens differ for both tokenizers. Consequently the document is the resampling unit used here, but donor reuse induces a fixed paired design; no independent-donor-population guarantee is asserted. The primary evaluation uses unmodified starting tokens, local ablations targeting attention to position 0, and norm-preserving donor resampling. Position-0 values change under knockout, so the natural-text test does not rely on a fixed start-token value. Its scope is limited to these window and donor settings.

Selector panel. Per task, the selector panel uses 32 qualification inputs, 64 fit inputs, and 256 calibration inputs organized as eight disjoint panels of 32. Fitting inputs estimate mean attention mass, norm parameters, and the Kullback–Leibler (KL) divergence between clean output distributions. Selection by MAE uses singletons and pairs under zero and half-scale ablation. Evaluation uses 256 inputs per task and all 15 nonempty head subsets under zero, half-scale, and resample knockouts. GPT-2 sites 6 and 9 and Qwen sites 12 and 18 share the relevant input instances. The MAE-based selector chooses value removal in all six conditions.

Replay and adjusted attention. Implementation-check inputs are separate from evaluation inputs. The fixed source is L5 heads [5,1,0,8], with local-ablation site 6; fixed mean-attention parameters are estimated on fitting data and held fixed for evaluation. Evaluation uses 128 induction inputs. The adjusted-attention experiment reuses the replay experiment's inputs as a paired companion. All 16 subsets, including the empty reference, are executed; the 15 nonempty subsets define the knockout-effect summaries.

Second task and model comparison. This experiment evaluates GPT-2 small on IOI and Qwen2.5-0.5B on induction, with 32 qualification and 128 held-out evaluation inputs per model-task setting. Each setting reuses mean-attention-mass parameters estimated from the selector panel's 64 fit inputs, with no new fitting or calibration.

Natural text. The natural-text comparison uses 32 fit documents to estimate mean attention mass and 128 evaluation documents for both raw and magnitude-matched comparisons in each model. The knocked-out source heads are L5 heads [5,1,0,8] in GPT-2 (local-ablation site 6) and L10 heads [0,4,8,12] in Qwen2.5-0.5B (site 12). The 64 prepared calibration documents are retained in the input files but are not used for fitting, selection, or reported evaluation. Mean-attention parameters remain fixed for the raw and matched evaluations. Matching is a paired auxiliary intervention, not an independent replication.

Eight-head decisions. The selector-panel parameters and the fixed candidate sets remain fixed. An earlier decision study and its independent repetition each generate 32 calibration and 128 test inputs per task; the repetition's inputs exclude those of the earlier study. Qwen uses the same underlying induction sequences with its model-specific beginning-of-sequence (BOS) convention. The repetition selects between value removal and fixed mean attention using calibration D, separately for raw and matched versions, then evaluates the selected ablation's R on test inputs.

Adjusted attention and strong controls. The selector-panel parameters remain fixed. The new calibration and test sets contain 32 and 128 inputs, respectively. The location and freeze controls of Appendix F are analyzed from the same multi-branch run. The natural-text adjusted-attention comparison reuses the 128 natural-text test documents; it is not a new-document replication.

Complement ablation. The complement-ablation evaluation uses 32 qualification, 32 fit, 32 calibration, and 128 test inputs. Non-circuit-head means are computed from the fitting inputs before any knockout outcome is evaluated. Appendix D.4 reuses these input sets and the fitted ablation for paired comparisons.

Third model. For Qwen2.5-1.5B, the fixed source is L12 heads [0,4,5,11], with local-ablation site 14 and model-specific fitted mean attention mass. This experiment reuses the natural-text source documents and the Qwen2.5-0.5B token windows, after verifying tokenizer compatibility. It adds a model, not new text.

Self-repair and backup ranking. Name-mover sets and the arXiv v1 backup labels are fixed. The self-repair experiment and the backup-ranking experiment each have a separate qualification set of 32 and their own 128 new IOI test inputs. The ranking experiment tests the first-token attention decrease score on inputs independent of the self-repair experiment, where the pattern was first inspected; the labels remain fixed.

Diagnostic boundaries. The channel-size and sensitivity-prediction analyses of Appendix F reuse the replay and second-task test inputs; they are follow-up analyses, not untouched-test prediction validations. The prospectively selected Qwen condition has separate selection, qualification, and fit sets of 32 each and fixes the selected condition before testing on 128 new inputs. The other-knockout and destination-sign experiments use 128 new inputs per condition.

The eight-head GPT-2 candidate set is L5 [0,1,2,4,5,7,8,11]; Qwen uses L10 [0,1,4,5,8,9,12,13]. These are different from the candidate circuit heads: L9H9, L9H6, L10H0, L8H6, L7H9, L5H5, L3H0, and L10H7 (three name movers, two S-inhibition heads, one induction head, one duplicate-token head, and one negative name mover, drawn from five of the circuit's seven head classes). The latter all belong to the fixed 26-head IOI circuit of arXiv v1 (Wang et al., 2023); the ICLR 2023 version describes 28 heads and lists L11H3 instead of L9H0 among the backup name movers, and Appendix E uses the arXiv v1 labels. Every attention head outside that circuit—118 heads—is subject to the ablation operation at every layer and position. This experiment uses our IOI generator, a 32-input fit mean, and NLL endpoints; it is not a reproduction or correction of the original published evaluation.

The 1,536-input test, prediction, and selection. The 1,536-input test retains the original fitted mean and circuit but uses a separate 64-input development set and 1,536 fresh inputs (Appendix H). The prediction experiment separates both input pools and source-head identities across training, validation, and evaluation (Appendix I). The selection experiment uses 32 qualification, 32 score-fitting, and 128 evaluation inputs and chooses the pairs on the score-fitting inputs; its test source ablations are

qualified on the qualification inputs and held out from score fitting, with overlapping source heads (Appendix J).

C Ablations and knockouts

This appendix specifies the ablations and knockouts of Sections 2–4 and of Appendices D–M. For one query and head, write the attention output as $z = av_0 + (1-a)u$, where a is position-0 attention mass, v_0 is the value at that position, and u is the value mixture renormalized over all other causal positions. The local ablations are:

Table 6: Local attention ablations and their information requirements.

Ablation	Head-output expression	Information requirement
Value removal	$z - av_0$	Current attention and values; preserves current mass and non-sink routing
Fixed mean attention	$z + (\bar{a} - a)(v_0 - u)$	Mean attention $\bar{a}$ fitted over clean fit inputs, separately by query/head/site
Fixed mean attention with value removal	$(1-\bar{a})u$	Same fitted mean; changes both mass and the position-0 value term
Sink removal	u	Removes the sink and renormalizes other positions
Norm-fitted sink removal	$\bar{k}u$	Fit mean of $\|z\|/\|u\|$, used as a strong selector baseline
Same-input mass freeze	$z + (a_{clean}(x) - a)(v_0 - u)$	One clean pass for this input
Adjusted attention	Substitute the adjusted attention below	Fitted mean, one clean pass, current site state

With $t(a) = \mathrm{logit}(a)$ and σ denoting the logistic sigmoid, the adjusted attention is

$$a_{offset} = \sigma\big[t(\bar{a}) + t(a_{current}) - t(a_{clean})\big].$$

Here $\mathrm{logit}(\bar{a})$ is the logit of the **mean fitted mass**, not the mean of log-odds. The adjustment reads the current pre-ablation state at the intervention site, where upstream computation is shared. It does not require a separate completed full-model knocked-out trajectory. A no-knockout numerical check verifies that it equals fixed mean attention. By contrast, downstream mass replay (Appendices D.4 and E) reads full-model knocked-out downstream masses, so it tests a mechanism and is not a usable ablation.

The ideal logit and conditional-mixture expressions assume $0 < a < 1$. In the implementation, each logit argument is clipped to $[\epsilon_p, 1-\epsilon_p]$, where ϵ_p is the floating-point machine epsilon (approximately 1.19×10^{-7} for float32). Denote this clipped logit by t_ϵ, and let $\Delta t_\epsilon = t_\epsilon(a_{\mathrm{current}}) - t_\epsilon(a_{\mathrm{clean}})$. The code evaluates the adjusted attention in difference form:

$$a_{\mathrm{implemented}} = \bar{a} + \left[\sigma(t_\epsilon(\bar{a}) + \Delta t_\epsilon) - \sigma(t_\epsilon(\bar{a}))\right].$$

This equals the ideal expression in exact arithmetic without clipping and preserves the fitted mean exactly when the logit change is zero. Protected initial queries are left unchanged. The numerical convention is distinct from fitting or tuning the ablation.

Local ablations leave query 0 unchanged. In natural-text runs, queries 0–3 are protected and local ablation begins at query 4. Position-1 drop removes position-1 attention and renormalizes the remaining distribution; it is a location control that replaces no sink computation, so its low regret does not establish an optimal sink ablation.

Upstream output-zero knockout zeros the selected attention-head outputs. Half-scale knockout multiplies them by 0.5. Norm-preserving resampling replaces a selected head's output vector by its

donor direction rescaled to the recipient vector norm, with numerical checks of norm preservation. When the other head of a knocked-out pair lies upstream, the recipient norm can differ between the frozen and full models only because non-circuit heads reacted to the first knockout, so this difference is part of the blocked indirect effect, and the frozen model still measures a natural direct effect under resample knockouts. This head-output constraint does not preserve the norm of the projected sum of heads. Drawing a donor input follows the resampling ablations of causal scrubbing (Chan et al., 2022); unlike causal scrubbing, we keep only the donor direction and rescale it to the recipient norm. Donor head outputs come from the full model's clean pass on the donor input and are shared by all ablations; the donor is fixed within input comparisons.

Mean ablation replaces each non-circuit head output by its clean fit-set mean at the corresponding layer and position. Zero ablation of the complement zeros those outputs; resample ablation of the complement uses norm-preserving donor head outputs. Freezing instead substitutes each non-circuit head's output from the same input's full-model clean pass. All circuit heads recompute except for the explicitly knocked-out members of the candidate pairs. This frozen model preserves the clean computation by construction but suppresses the non-circuit heads' reactions to the knockout.

In Appendix D.4, first-token attention replay starts from that frozen output. For each non-circuit head and query after position 0, it uses

$$z_{\text{freeze+replay}} = z_{\text{clean}} + (a_{\text{knockout}} - a_{\text{clean}})(v_{0,\text{clean}} - u_{\text{clean}}).$$

All clean quantities come from the same input's full-model clean pass; a_{knockout} comes from its full-model run under the knockout. Thus the replay retains clean values and clean conditional non-sink routing, while changing only the first-token attention. Query 0 remains at its clean output. This differs from the local replay of Appendix E, which retains the current branch's values and conditional routing.

For magnitude matching, let z concatenate the head outputs, let $\Delta z_b = z_b - z$, and let $\delta_b = W_O \Delta z_b$ be the bias-free residual update at each token. For ablations b, c, set $m = \min(\|\delta_b\|_2, \|\delta_c\|_2)$ and use

$$z_b^{\text{matched}} = z + \frac{m}{\max(\|\delta_b\|_2, \epsilon)} \Delta z_b.$$

The corresponding operation is applied to c, with $\epsilon = 10^{-12}$. Neither branch is enlarged. If either projected norm is zero, both matched projected updates are zero. Matching is performed separately on clean and knocked-out branches before their own clean references are subtracted. Projection bias is omitted because it cancels from an output difference. This comparison tests whether the raw faithfulness gap persists when the projected update magnitudes are matched. Because it interpolates toward the unmodified output, it can weaken the intended removal; it does not establish equivalent semantic content removal or a universally preferable intervention. A matched ablation is therefore defined relative to its comparator and branch and is not a standalone ablation; in GPT-2 natural text, value removal is the smaller update, so matching leaves it unchanged (Figure 2).

D Quantitative effect and decision comparisons

This appendix gives the comparisons behind Sections 3.1–3.2 (Appendix D.4), an endpoint check (Appendix D.5), and two first-token attention ablations compared at matched size. **Ablation magnitude.** The two local ablations of Appendix C, value removal and fixed mean attention, act on attention to position 0 rather than on whole heads. Value removal looks the better of the two at raw magnitude (Appendix D.2), but their updates differ in size. We compare them at matched norms on 128 natural-text documents (all 15 nonempty subsets of four ablated heads under donor resampling; the local-ablation site is layer 6 in GPT-2, 12 in Qwen2.5-0.5B, and 14 in Qwen2.5-1.5B; Appendix B), with the matching rule of Appendix C applied in both clean and knocked-out comparisons before the downstream outputs are recomputed. In GPT-2, value removal has lower raw MAE than fixed mean attention (0.00352 versus 0.01319 nats). After matching, fixed mean attention falls to 0.00180 nats and is the lower. Both contrasts (raw and matched) have 97.5% intervals excluding zero, as do the smaller post-matching advantages of value removal in both Qwen models (Figure 2; Appendix D.1). The ordering also depends on task: value removal keeps lower MAE after matching in the 28-pair GPT-2 IOI comparison (Appendix D.2).

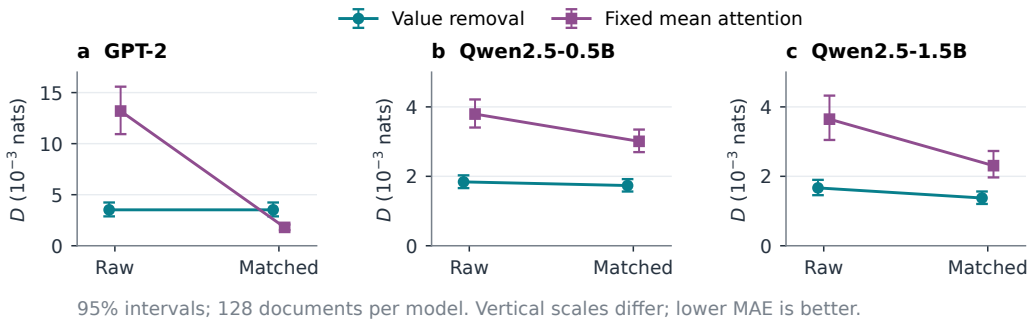


Figure 2: Matching ablation magnitude reverses which first-token attention ablation has the lower MAE in GPT-2, while value removal keeps the lower error in both Qwen models (MAE of knockout effects on 128 WikiText-2 documents per model). Intervals are 95% document bootstraps; paired 97.5% contrasts are in Appendix D.1.

Do these faithfulness comparisons change component decisions? This study covers six task–site conditions on GPT-2 and Qwen2.5-0.5B (Appendix D.2); GPT-2 induction at site 6 is primary. Each condition uses 32 calibration inputs, 128 new evaluation inputs, and all 28 pairs of eight candidate heads. On calibration inputs we select the lower-D ablation, for the raw and for the matched versions; each then chooses a head pair per evaluation input and ablation type, scored in the full model. Raw value removal has lower regret in all six conditions. Matching removes most of the gap in the four GPT-2 conditions and, in the primary one, reverses its sign: value removal minus fixed mean attention changes from -0.016 to $+0.002$ nats (unresolved). Choosing the ablation by matched rather than raw MAE does not detectably change regret in any condition. Tolerance and superiority differ: the calibration-selected matched ablation meets the tolerance in all six conditions and raw value removal misses it in the primary one. Much of value removal's raw advantage thus reflects its smaller update. The adjusted attention of Appendix C, which shifts the fitted mean attention by the current-minus-clean change in log-odds, is compared with both in Appendix D.3.

D.1 Natural-text magnitude comparison

The table reports $D_{\mathrm{VR}} - D_{\mathrm{FMA}}$, the MAE of value removal (VR) minus that of fixed mean attention (FMA); negative values favor value removal. Every estimate uses the same 128 documents within its paired comparison. Intervals are 97.5%.

Table 7: Natural-text MAE before and after magnitude matching.

Model/site	Comparison	VR D	FMA D	Difference, 97.5% interval
GPT-2/6	Raw	0.003518	0.013193	−0.009675 [−0.011700, −0.007780]
GPT-2/6	Matched	0.003518	0.001799	+0.001719 [0.001229, 0.002294]
Qwen-0.5B/12	Raw	0.001839	0.003793	−0.001954 [−0.002471, −0.001469]
Qwen-0.5B/12	Matched	0.001735	0.003008	−0.001273 [−0.001685, −0.000889]
Qwen-1.5B/14	Raw	0.001666	0.003647	−0.001980 [−0.002737, −0.001325]

Model/site	Comparison	VR D	FMA D	Difference, 97.5% interval
Qwen-1.5B/14	Matched	0.001376	0.002307	−0.000931 [−0.001311, −0.000614]

The GPT-2 ordering reverses. Both Qwen models retain lower error for value removal at matched magnitude. The three observations do not identify whether attention architecture, normalization, training data, or another model property causes the difference.

D.2 TWENTY-EIGHT-PAIR COMPARISONS

The IOI task comparison at the start of this appendix concerns MAE D: matched value removal minus matched fixed mean attention is −0.005923 [−0.006968, −0.004890] nats at GPT-2 IOI/site 6 and −0.002812 [−0.003321, −0.002321] at IOI/site 9 (paired 97.5% intervals; 128 inputs). Both favor value removal. These are separate IOI inputs from the natural-text comparison, and the MAE result does not imply a resolved regret advantage at both sites.

The following table gives paired differences in full-model regret, in nats. The final column compares the evaluation regret of the calibration-selected matched ablation with that of the calibration-selected raw ablation. Each selected ablation chooses its own pair for each evaluation input and ablation type.

Table 8: Full-model decision-regret contrasts in the 28-pair evaluations.

Model/task/site	Raw VR − FMA R	Matched VR − FMA R	Regret of matched-selected minus raw-selected ablation
GPT-2 induction/6	−0.01646 [−0.03211, −0.00474]	+0.00162 [−0.00120, 0.00498]	−0.00162 [−0.00498, 0.00120]
GPT-2 induction/9	−0.00834 [−0.01507, −0.00393]	−0.00032 [−0.00093, 0.00000]	0
GPT-2 IOI/6	−0.04468 [−0.07570, −0.02325]	−0.00266 [−0.00542, −0.00050]	−0.00009 [−0.00047, 0.00011]
GPT-2 IOI/9	−0.01879 [−0.03456, −0.00703]	−0.00018 [−0.00056, 0.00019]	+0.00009 [−0.00001, 0.00036]
Qwen induction/12	−0.00228 [−0.00381, −0.00104]	−0.00116 [−0.00215, −0.00031]	−0.00015 [−0.00051, 0.00016]
Qwen induction/18	−0.00009 [−0.00022, −0.00001]	−0.00003 [−0.00017, 0.00009]	+0.00005 [−0.00001, 0.00013]

These are 97.5% intervals. GPT-2 induction/site 6 is primary, GPT-2 IOI/site 6 and Qwen/site 12 are secondary, and the other conditions are auxiliary. Matching removes 94–110% of the signed raw regret gap in the four GPT-2 conditions and 49% at Qwen/site 12; a value above 100% denotes a sign reversal, not a causal variance fraction. Calibration selects a different ablation in the primary condition, but this change does not establish lower evaluation regret. Fixed value removal already improves raw regret over fixed mean attention in all six conditions; that improvement is not an incremental gain attributable to the selector.

An earlier decision study accepted value removal in all six conditions, with five evaluation sets supporting the tolerance and IOI/site 6 indeterminate. In the independent repetition reported here, value removal is accepted on calibration in five conditions. In the primary condition it is not accepted on calibration and is indeterminate on evaluation: one of the 32 calibration inputs has regret 0.121, raising the calibration upper bound to 0.01355. Matched fixed mean attention is accepted on calibration and supported on evaluation in this primary condition. These acceptance outcomes coexist with an unresolved regret advantage over the fixed raw value-removal baseline.

Table 9: Calibration and evaluation of the matched ablation selected by lower calibration MAE in each condition. All six are accepted and supported at $\tau = 0.005$; bounds and intervals are in nats.

Model/task/site	Selected matched ablation	Calibration upper 95% bound	Evaluation R, 95% interval
GPT-2 induction/6	Fixed mean attention	0.000554	0.002201 [0.000399, 0.004837]
GPT-2 induction/9	Value removal	0.001821	0.000064 [0.000005, 0.000154]
GPT-2 IOI/6	Value removal	0.002307	0.001281 [0.000611, 0.002074]
GPT-2 IOI/9	Value removal	0.000307	0.000142 [0.000020, 0.000358]
Qwen induction/12	Value removal	0.002274	0.000618 [0.000328, 0.000957]
Qwen induction/18	Value removal	0	0.000071 [0.000018, 0.000141]

The exact zero in the induction/site-9 row of the regret-contrast table arises because raw and matched value removal select the same pair on every evaluation input; the zero calibration bound for Qwen/site 18 arises because all 32 calibration regrets are zero. Selected-matched acceptance does not mean every matched mean-attention branch passes: GPT-2 IOI/site 6 and Qwen/site 12 have calibration upper bounds 0.006007 and 0.005537. In the primary condition, raw and matched value removal select the same pair on every evaluation input: value removal is the smaller projected update on 99% of evaluation tokens, so matching leaves it essentially unchanged, and the final column of the regret-contrast table is the exact negative of the matched contrast.

Calibration sensitivity. Repeating the rule after each one-input deletion (31 inputs, 8,000 resamples) shows that omitting the largest primary raw regret, 0.121240 nats, changes its upper bound from 0.013546 to 0.004831 and reverses acceptance; the other 31 deletions do not. Twenty-four of the 32 raw regrets are zero. After lower-D reselection, the chosen matched ablation remains accepted under every deletion in all six conditions. Only auxiliary induction/site 9 switches matched choice once; both alternatives are accepted and supported on the fixed evaluation set. This conditional influence check does not justify dropping the observation or changing the reported outcome.

With the stored full-sample bounds, the primary labels persist for $0.0048371 \leq \tau < 0.0070012$ nats; τ does not affect lower-D selection or paired superiority. In a 200,000-resample check with three seeds, the no-maximum raw upper bound was 0.004831–0.004848 and the matched evaluation upper endpoint 0.004833–0.004851. The reported 8,000-resample estimates are unchanged; this check does not establish bootstrap coverage.

D.3 Conditional adjustment of attention mass

The adjusted-attention ablation of Appendix C is compared with fixed mean attention and with value removal, raw and matched, in six 28-pair synthetic conditions and two natural-text conditions.

Table 10: MAE contrasts for adjusted attention.

Condition	Adjusted − FMA D	Raw adjusted − VR D	Matched adjusted − matched VR D	Decision qualification
GPT-2 natural text/6	−0.00337 [−0.00457, −0.00231]	+0.00630 [0.00512, 0.00760]	−0.00186 [−0.00244, −0.00137]	Matched R: 0.00004 versus 0.00023; descriptive secondary comparison
Qwen natural text/12	−0.00020 [−0.00045, 0.00004]	+0.00175 [0.00130, 0.00220]	+0.00111 [0.00076, 0.00147]	Secondary matched R difference includes zero
GPT-2 induction/6	−0.0139 [−0.0178, −0.0101]	+0.0237 [0.0193, 0.0285]	−0.0089 [−0.0111, −0.0070]	Matched R difference includes zero
GPT-2 induction/9	−0.0098 [−0.0144, −0.0052]	+0.0299 [0.0258, 0.0343]	−0.0013 [−0.0021, −0.0005]	Matched R difference includes zero
GPT-2 IOI/6	−0.0313 [−0.0351, −0.0277]	+0.0034 [0.0018, 0.0050]	−0.0011 [−0.0019, −0.0003]	Matched R difference includes zero
GPT-2 IOI/9	−0.0109 [−0.0133, −0.0087]	+0.0073 [0.0058, 0.0090]	+0.0007 [0.0003, 0.0010]	Matched R difference includes zero
Qwen induction/12	−0.0040 [−0.0054, −0.0028]	+0.0041 [0.0024, 0.0061]	+0.0033 [0.0020, 0.0050]	Matched VR has lower regret
Qwen induction/18	−0.0017 [−0.0021, −0.0013]	+0.0004 [0.0000, 0.0007]	−0.0005 [−0.0007, −0.0003]	Matched R difference includes zero

Displayed MAE intervals are 97.5%; the natural-text regret comparisons are secondary and use 95% intervals. Adjusted attention mass reduces D relative to fixed mean attention in all six synthetic conditions and GPT-2 natural text. It reduces synthetic R in four of six conditions, with 97.5% intervals excluding zero. At raw magnitude, value removal has lower MAE than adjusted attention in all eight conditions and lower regret in three synthetic conditions (induction/6, IOI/9, and Qwen/12); the matched comparison in the fourth column is the headline comparison here. Raw GPT-2 adjusted attention is not accepted on calibration; matched adjusted attention is accepted and supported in its two induction conditions. Natural-text GPT-2 D falls from 0.01319 to 0.00982, about 26%, and reaches 0.00166 after matching. These are conditional improvements, not evidence for a universally optimal ablation. The location and freeze controls of Appendix F use the same synthetic runs.

D.4 Mean ablation, freezing, and first-token attention replay

All complement ablations below act on the 118 non-circuit attention heads. **Frozen** denotes freezing at the same input's clean outputs; **replay** substitutes full-model knocked-out first-token attention while retaining the clean values and conditional routing, as defined in Appendix C; **matched** denotes magnitude matching of the non-circuit-head updates in which, at each layer and token, the mean and zero updates are shrunk, never enlarged, to the smaller of their two projected norms computed from the same input's full-model clean outputs, so the matched ablations require one full-model clean pass per input, unlike the per-branch rule of Appendix C. Every row other than the full model is **not accepted on calibration**. The second table gives the separate evaluation classification.

Table 11: Clean performance of complement ablations on the historical 128-input evaluation.

Condition	Clean accuracy	Clean NLL
Full model	94.5%	0.893
Mean	93.0%	1.117
Resample	43.0%	2.005
Zero	0.0%	6.918
Frozen	94.5%	0.893
Frozen + replay	94.5%	0.893
Matched mean	93.8%	1.080
Matched zero	89.1%	1.205

Table 12: Knockout-effect and decision faithfulness on the same historical 128-input evaluation.

Condition	D	R, 95% interval	Agreement	Evaluation
Full model	0	0	100%	Reference
Mean	0.244	0.039 [0.027, 0.052]	77.0%	Violated
Resample	0.375	0.090 [0.069, 0.112]	67.6%	Violated
Zero	0.636	1.066 [0.963, 1.168]	16.0%	Violated
Frozen	0.072	0.0077 [0.0038, 0.0125]	90.6%	Indeterminate
Frozen + replay	0.060	0.0058 [0.0032, 0.0089]	91.0%	Indeterminate
Matched mean	0.222	0.031 [0.021, 0.042]	80.1%	Violated
Matched zero	0.191	0.028 [0.016, 0.041]	82.4%	Violated

The mean ablation's similar accuracy does not imply preservation of every clean metric: its clean NLL increases; the original paper's logit-difference endpoint is examined in Appendix D.5. The 0.93 single-head rank correlation is the **mean per-input Spearman correlation** across the effects of eight single-head zero ablations; it is not a correlation computed after pooling all inputs. Pair agreement averages separate zero/resample decisions. Of the 59 input–ablation-type selections in which the mean ablation and the full model choose different pairs, 58 have full-model regret above a descriptive 0.005-nat cutoff and 46 above 0.05 nats (median over the 59: 0.104 nats); across all 256 selections the full-model best-minus-second-best gap has median 0.240 nats, and only 7 selections have a gap of at most 0.005 nats. The cutoff equals the tolerance numerically but is not validated for single selections, and this re-aggregation of stored rows does not replace the primary intervals.

Replacing a fitted mean by the same input's clean outputs reduces D by approximately 71% and R by approximately 80%. The ratio $D_{freeze}/D_{mean} = 0.294$ has 97.5% interval [0.273, 0.316]; the corresponding regret ratio $R_{freeze}/R_{mean} = 0.198$ has 95% interval [0.105, 0.312]. Replaying the non-circuit heads' first-token attention reduces the remaining D by 0.01145 [0.00874, 0.01435], or 16.0% [12.5%, 19.4%]. These sequential MAE reductions are not independent additive causal shares; 16% applies to the residual after clean freeze, not the original mean-ablation error. None of these substitute ablations is accepted on calibration under the stated rule.

Let $C = D_{\text{freeze+replay}} - D_{\text{freeze}}$, $f = -C/D_{\text{freeze}}$, and $f_1 = -(D_{\text{freeze+pos1replay}} - D_{\text{freeze}})/D_{\text{freeze}}$, where the position-1 replay acts at queries $q \geq 2$. The ratio intervals below use the same input resamples in numerator and denominator.

Table 13: Sequential-ablation quantities on the 128-input evaluation, with 97.5% intervals.

Quantity	Estimate	97.5% interval
$D_{\text{freeze}}/D_{\text{mean}}$	0.294	[0.273, 0.316]
Replay change C	-0.0115	$[-0.0144, -0.0087]$
Replay fraction f	0.160	[0.125, 0.194]
Position contrast $f - f_1$	0.163	[0.130, 0.197]

Clean freezing thus removed more than half of the mean-ablation error, and replay less than a quarter of the remainder. Three practical candidates were also checked against the tolerance: mean ablation with recomputed first-token attention (the fitted mean ablation with the non-circuit heads' position-0 masses recomputed, $\bar{z}+(a-\bar{a})(\bar{v}_0-\bar{u})$ at $q \geq 1$ with fitted $\bar{a}$, $\bar{v}_0$, and $\bar{u}$), matched mean ablation, and matched zero ablation of the complement; replay branches, which read the knocked-out full model, do not count. All three are unaccepted on calibration and violated on evaluation, although replay improves knockout-effect faithfulness and exceeds the unmatched position-1 replay comparison. That comparison does not use the matched support and magnitude of the 1,536-input test and is distinct from its unresolved position-specificity result (Appendix H).

Fixed versus input-specific choices. On both the 128-input evaluation and the 1,536-input test, full-model and mean-ablation cohort-average knockout effects select the same pair: L9H9+L8H6 under zero ablation and L9H9+L10H0 under resampling. Fixing these full-model choices, one per ablation type, on the 32 calibration inputs nevertheless leaves room for input-specific selection. The table measures full-model NLL-effect gains over those fixed choices, making selections separately for each ablation type before averaging.

Table 14: Full-model-effect gains over pairs fixed on the complement-ablation calibration inputs. Descriptive 95% input-bootstrap intervals, in nats.

Evaluation cohort	Full-model choice gain	Mean-ablation choice gain
Historical evaluation, 128 inputs	0.120518 [0.098121, 0.144609]	0.081599 [0.057979, 0.106106]
1,536-input test	0.128833 [0.121373, 0.136386]	0.088550 [0.080740, 0.096598]

On the 128 inputs the full model's most damaging pair raises the answer loss by 1.62 nats on average over input–ablation-type selections, against 1.50 for the calibration-fixed pairs. The difference between these gains is full-model decision regret. Intervals use 8,000 paired input resamples, holding the calibration-fixed pairs unchanged. These gains describe larger selected intervention effects, not task-performance or computational improvements, and do not replace the primary intervals.

D.5 Endpoint sensitivity: indirect-object–subject logit difference

The primary complement-ablation analyses score the designated answer token's NLL. As a sensitivity check, we recomputed the 32 calibration and 128 evaluation inputs of the complement-ablation evaluation with the same fitted mean ablation, 26-head circuit, eight candidate heads, 28 pairs, and both ablation types, additionally recording the logit difference between the indirect object and the repeated subject, the endpoint of the original circuit study (Wang et al., 2023). Its intervention effect is the clean-minus-knocked-out decrease of that margin, so positive values denote harm; pair selection and full-model-scored regret follow Appendix A with this endpoint in place of NLL. The two endpoints, five conditions (the full model, mean ablation, freezing, the first-token attention freeze K_0, and mean ablation plus δ_0), and two primary contrasts (B_M and T_R; notation of Appendix H) make up this check; primary intervals are 98.75% (four-test Bonferroni) and the others 95% descriptive. Values are in logits, not nats, and the NLL tolerance does not transfer. The full-model and mean NLL branches of the same run agree with the Appendix D.4 estimates to displayed precision, and an independent re-analysis from the saved rows reproduces every point estimate below.

Table 15: Indirect-object–subject endpoint on the 128 complement-ablation evaluation inputs (paired input bootstrap).

Condition	D (logit)	R (logit), 95% interval	Agreement
Mean	0.402	0.0762 [0.0556, 0.0990]	64.5%
Frozen	0.199	0.0316 [0.0220, 0.0424]	79.3%
Full-model first-token attention freeze K_0	0.089	0.0145 [0.0088, 0.0213]	85.9%
Mean + δ_0	0.409	0.0632 [0.0467, 0.0822]	66.4%

Under this endpoint the mean ablation changes the selected pair in 35.5% of input–ablation-type selections (NLL: 23.0%), and freezing reduces D by 50.5% and R by 58.5% (NLL: 71% and 80%). The first-token attention freeze contrast $B_M = 0.0398$ [0.0208, 0.0610] excludes zero at the family level and is carried by zero ablation (0.0766 [0.0470, 0.1085]; resample 0.0029 [−0.0044, 0.0102]; 95%). The patch contrast $T_R = 0.0130$ [−0.0033, 0.0345] is unresolved at the family level and positive under zero ablation alone (0.0279 [0.0050, 0.0557]; 95%). The two endpoints' full-model argmax pairs coincide on 36.3% [30.1%, 42.6%] of selections: under zero ablation the calibration-fixed full-model pair is L8H6+L7H9 for the logit difference and L9H9+L8H6 for NLL, and a pair selected under one endpoint scores a regret of 0.287 nats or 0.313 logits under the other. The ablation-induced mismatch and the value of retained clean outputs therefore persist under this endpoint, with different magnitudes, whereas the identity of the most damaging pair does not. This check reuses earlier inputs rather than fresh ones.

E Causal channel and backup ranking

Summary. The induction-task replay supports the knockout dependence noted in Section 6; the backup ranking tests whether documented backups show a larger first-token attention decrease. In GPT-2 induction, we ablate subsets of four layer-5 heads and run two branches. Both freeze first-token attention at layer 6; one also replays the full-model knocked-out first-token attention at later layers, restoring the downstream heads' first-token attention to its full-model knocked-out values. Relative to the freeze-only branch, the replay branch changes loss by −0.0938 nats under zero ablation [−0.1176, −0.0714; 97.5% interval]. Contrasts against position-1 replay and three random controls matched in norm and angle also exclude zero. Under resampling, the replay effect is only +0.0014 nats [−0.0053, +0.0084]. This downstream replay effect depends on the ablation type. Better knockout-effect faithfulness does not guarantee better decisions: under zero ablation, replay slightly lowers MAE but raises the layer-6 freeze's regret from 0.0269 to 0.0335.

For the backup ranking, which asks whether documented backups show a larger first-token attention decrease: ranked by first-token attention decrease among all 141 candidate heads on 128 further inputs, in the benchmark design of Gong et al. (2026), the eight documented backups (two in layer 9, which holds two of the knocked-out name movers) reach an area under the ROC curve (ROC-AUC) of 0.859. That study's conditional co-ablation (ablating each candidate together with the name movers) reaches 0.91 on its inputs in its first arXiv version and 0.94 in its third. A layer-10/11 indicator alone gives 0.811, but within the 23 downstream heads the decrease score still reaches 0.922, against 0.647 for single ablation. The details follow.

The local replay test summarized above (GPT-2 induction) first freezes the first-token attention at a designated site to its same-input clean value, then replaces later layers' position-0 masses by those from full-model knockout while retaining the current branch's values and relative non-sink routing. The designated-site freeze remains active in every replay and control branch. Its contrast is $C_{replay} = L_{replay0} - L_{freeze}$. In GPT-2 induction/site 6, zero-ablation $C_{replay} = -0.09381$ [−0.11756, −0.07136], whereas norm-preserving resample yields 0.00142 [−0.00527, 0.00843] (97.5%). Position-1 and three matched projected-geometry random controls do not reproduce the large

zero-ablation effect. This is a pathway result, not automatic decision repair: under zero knockout, MAE changes from 0.2152 to 0.2053 and regret from 0.02686 to 0.03352 under replay.

The random controls here preserve the projected update's coordinate mean, norm, and component along the centered residual, while randomizing the orthogonal component. Their reference is the centered residual, rather than the mean-ablation write used by the random controls of the 1,536-input test (Appendix H).

Table 16: Specificity contrasts for the local replay test on 128 induction inputs. Each entry is position-0 replay loss minus control loss, in nats, with a paired 97.5% interval. Negative values favor position-0 replay. All four zero-ablation intervals exclude zero; all four resample intervals include zero. The designated-site freeze remains active throughout.

Control	Zero ablation	Norm-preserving resample
Position-1 replay	−0.09242 [−0.11595, −0.07006]	0.00199 [−0.00487, 0.00931]
Random 1	−0.10837 [−0.14370, −0.07591]	0.00379 [−0.00621, 0.01458]
Random 2	−0.11069 [−0.14717, −0.07680]	0.00216 [−0.00625, 0.01077]
Random 3	−0.10839 [−0.14495, −0.07361]	0.00352 [−0.00452, 0.01219]

Adjusted attention mass is a separate local intervention. Under zero ablation, relative to fixed mean attention, it reduces D by 0.09984 [0.07127, 0.13072] and six-pair regret by 0.02786 [0.00835, 0.05557]. In this induction condition, fixed value removal remains stronger. Appendix C states the adjustment's local information requirements.

For the name-mover experiment, the primary knockout set is {L9H9,L9H6,L10H0}. The downstream set comprises the other 11 L10 attention heads, all 12 L11 attention heads, and multilayer perceptrons (MLPs) at layers 9–11. A direct-effect branch freezes these downstream outputs and final layer normalization (LayerNorm) statistics to the same input's clean values. Define $TE = L_{knockout} - L_{clean}$, $DE = L_{direct} - L_{clean}$, and $SR = DE - TE$, following the total- and direct-effect decomposition of causal mediation analysis (Vig et al., 2020). For sink, position-1, attention, and MLP channels, $C_j = L_{freeze(j)} - L_{knockout}$. The LayerNorm contrast is specifically $L_{direct} - L_{direct,\text{LN recomputed}}$, conditional on the other downstream outputs being frozen. The latter branch retains the same frozen downstream attention and MLP outputs but recomputes final LayerNorm statistics.

Channel shares are ratios of input means, with the ratio recomputed in each bootstrap; they are not averages of per-input ratios. The secondary margin is the correct-name logit minus the logit of a sampled third name absent from the prompt; it is not the indirect-object–subject logit difference and is not used in the reported results. We reverse its effect sign so positive denotes harm.

Table 17: Name-mover self-repair and separately frozen downstream channels.

Name-mover zero-ablation quantity	NLL effect, 97.5% interval	Share of SR
Total effect TE	0.504 [0.414, 0.597]	—
Direct branch DE	5.196 [4.994, 5.387]	—
Self-repair SR	4.692 [4.502, 4.870]	Reference
Freeze the first-token attention of 23 downstream heads	1.748 [1.661, 1.833]	37.2% [35.8%, 38.7%]
Position-1 control	−0.00536 [−0.00956, −0.00165]	−0.11%
Freeze downstream attention outputs	4.475 [4.282, 4.659]	95.38%

Name-mover zero-ablation quantity	NLL effect, 97.5% interval	Share of SR
Freeze downstream MLP outputs	0.127 [0.076, 0.179]	2.71%
Final LayerNorm contrast	0.00266 [0.00062, 0.00477]	0.057%

Position-1 and LayerNorm effects are small but not literally zero. These separately frozen channels need not add: the attention/MLP/LayerNorm sum leaves a residual 0.0869 nats. The sink is a subchannel of attention and cannot be added again to the attention total. The 37% statement pertains to all 23 downstream heads, not only the six documented downstream backups. In this experiment the six documented downstream backups have a mean position-0 mass decrease of 0.055 against 0.012 for the other 17 downstream heads; L11H3, a backup only in the ICLR 2023 circuit, has the second-largest decrease.

The first-token attention decrease score is the decrease in a head's position-0 mass from the clean forward to the primary-knocked-out forward, averaged over queries $q \geq 1$; its clean-normalized variant divides that decrease by the head's clean position-0 mass. Scores are then averaged over the 128 inputs before ranking. The primary candidate set is all 144 GPT-2 heads except the three knocked-out name movers (141 candidates); the eight positive labels are the backup name movers of the arXiv v1 circuit (Appendix B). This benchmark design (three ablated name movers, a candidate set of the remaining heads, fixed backup labels, and ROC-AUC) follows the conditional co-ablation study of Gong et al. (2026), whose own conditional score reaches 0.91 on their inputs in the first arXiv version and 0.94 in the third. The area under the receiver operating characteristic curve (ROC-AUC) gives half credit to tied positive/negative scores; top-20 counts break ties by head index, and only 19 candidates have a positive mean first-token attention decrease. Confidence intervals resample inputs, recompute mean scores, and rerank against the same labels.

Table 18: Backup-label ranking and candidate-specific forward costs.

Score	ROC-AUC, 97.5% interval	Backups in top 20	Candidate-specific knocked-out forwards
First-token attention decrease	0.859 [0.857, 0.859]	6/8	0
Clean-normalized first-token attention decrease	0.851 [0.849, 0.851]	6/8	0
Conditional co-ablation, our implementation	0.665 [0.633, 0.687]	3/8	141
Single ablation	0.597 [0.573, 0.616]	2/8	141

The first-token attention decrease requires the clean reference and one primary-knockout forward; "zero candidate-specific forwards" does not mean no ablation. The two backups at L9 cannot respond to other heads knocked out in that same attention layer. Excluding only these two positives gives the recorded auxiliary AUC 0.990, but retains structurally upstream negative candidates. Narrow bootstrap intervals reflect a stable ranking on one fixed circuit; they do not establish accuracy across circuits or provide an uncertainty estimate for new head labels. The conditional baseline shares our inputs and implementation, and does not reproduce the separately reported CoAx (conditional co-ablation) benchmark. This result verifies a mechanism signature; end-to-end discovery performance and compute savings are not established.

Structural control. All 118 candidates at layers 0–9 that are not knocked out have exactly zero first-token attention change in these recorded runs. A score that merely indicates layer 10 or 11 already gives AUC 0.811 on the original 141 candidates. We therefore restrict the first-token attention decrease, single-ablation, and conditional co-ablation scores to the same 23 downstream candidates, with six documented backups and 17 negatives. On this restricted candidate set, AUC is 0.922 for first-token attention decrease, 0.647 for single ablation, and 0.657 for our conditional co-ablation implementation; all six backups occur in the decrease score's top 10. This comparison shows that the ranking contains information beyond the upstream/downstream split. It reanalyses the same 128 inputs, with no new model evaluations or confidence intervals. The primary 141-head analysis remains unchanged.

F Scope and boundary experiments

Summary (Section 6). The compensatory role of the change in first-token attention is not general: in two other GPT-2 IOI ablations (of S-inhibition and induction heads, two other classes of the circuit), downstream reactions amplify rather than offset the knockout, and sink interventions carry about 12% and 30% of that indirect, downstream effect. In two induction-task ablations the first-token attention channel is unresolved or compensatory. In Qwen2.5-0.5B, replay at a site selected for a large mass change changes loss by only -0.008 nats $[-0.018, 0.004]$.

For the prospectively selected Qwen2.5-0.5B condition, we count a downstream pathway as established only if the replay effect's 97.5% interval excludes zero and its magnitude is at least 0.01 nats; the observed -0.0079 $[-0.0183, 0.0037]$ meets neither condition, and it also fails the stronger check that the entire interval lie outside $[-0.01, 0.01]$. These are experiment-specific thresholds, not universal criteria for the existence of a mechanism.

Table 19: Scope and boundary experiments.

Question	Result retained in the manuscript's scope
Second task and corresponding Qwen2.5-0.5B site	GPT-2 IOI zero-ablation replay: -0.0971; adjusted-attention D/R differences: $-0.1128/-0.0631$ relative to fixed mean attention. Qwen/site 12 replay: +0.0005 $[-0.0001, 0.0011]$; no resolved decision repair.
Predicting where the channel is large	Zero moves GPT-2 site mass about 3.2 times more than resample in induction and 15 times more in IOI. Input-level tercile/correlation predictors fail. Ablation-type-level association does not establish an individual-input predictor.
Strong location and freeze controls	Position-1 drop, a location control that replaces no sink computation, has the lowest regret in three of the four GPT-2 conditions. Local same-input mass freeze is worse than value removal in all four GPT-2 conditions. This local freeze differs from the complement-output substitution that reduces error in Appendix D.4.
Model-size extension	Qwen-1.5B retains lower MAE for value removal after matching on the shared natural-text documents. It is not evidence for a cause specific to one architectural difference.
Sensitivity-based prediction	Clean sensitivity $\times$ knockout-induced mass change does not predict the channel on both tasks; a freeze-state predictor succeeds in IOI but not induction. No general input-level predictor is claimed.
Prospectively selected Qwen2.5-0.5B condition	The prospectively selected layer-2 knockout/site-3 condition has a mass change of 0.057, but its downstream replay effect is only -0.0079 $[-0.0183, 0.0037]$. The local adjusted-attention intervention reduces D by 0.0822 [0.0612, 0.1043]; its R change includes zero. Mass movement alone is insufficient for the downstream pathway.
Other knocked-out head sets	None of six conditions shows positive self-repair. In GPT-2 IOI S-inhibition and induction-head knockout, downstream reactions amplify harm; the first-token attention channel carries about 12% [8%, 15%] and 30% [25%, 35%] of that indirect effect (97.5% intervals). In GPT-2 induction with four layer-5 heads knocked out, the downstream reactions also amplify harm (self-repair -0.44 $[-0.81, -0.09]$ nats) but the first-token attention channel is unresolved (-0.004 $[-0.258, 0.257]$); with L5H5 and L6H9 knocked out, self-repair is unresolved (+0.10 $[-0.09, 0.30]$) and freezing first-token attention raises loss (+0.12 [0.02, 0.21]), the compensating sign. Low-effect Qwen2.5-0.5B conditions cannot establish channel absence.
Destination and sign	Answer-token attention changes correlate with channel effect at only 0.26–0.48. One of four sign tests is unresolved. A site's sign can differ from that of all downstream layers.

These boundaries are why we speak of downstream reactions rather than treating every recomputation as self-repair. Self-repair is established for the specified name-mover ablations. The tests do not support a universal beneficial change in first-token attention, a general sign rule, or claims about refusal and alignment outcomes.

G Numerical validation and cost

This appendix describes the numerical checks behind the results of Sections 3–4 and the other appendices, and the compute cost. The runs record numerical identity checks that verify agreement with full-model execution and suffix recomputation, unchanged protected positions and non-target heads, attention normalization and causal masking, donor norms, and matching scales. No-knockout sink-freeze and adjusted-attention identities check whether numerical artifacts could be mistaken for intervention effects. Failed implementation checks and initial evaluation attempts were kept in the experiment history and excluded from the reported estimates.

Representative runs used PyTorch 2.12.1 with Transformers 5.12.1 (the complement ablation and the comparisons of Appendix D.4, adjusted attention, self-repair, backup ranking, and the third model) or PyTorch 2.13.0 with Transformers 5.13.1 (the natural-text magnitude comparison, eight-head decisions, the 1,536-input test, prediction, and selection). Most later runs use two CPU threads; earlier Qwen/GPT-2 runs sometimes use four. Recorded wall time varies with concurrent jobs and pauses and is not a method-speed benchmark.

Table 20: Recorded execution cost of representative completed runs.

Experiment	Full forward calls	Suffix calls, where recorded	Elapsed seconds
Natural text, GPT-2, raw	554	1,056	1,198
Natural text, Qwen2.5-0.5B, raw	1,106	2,112	1,258
Eight-head decisions, GPT-2	2,394	4,640	1,642
Adjusted attention, GPT-2 synthetic	2,394	4,640	8,572
Complement ablation	5,230	—	4,010
Freezing and replay (Appendix D.4)	11,750	—	2,181
Self-repair channel	550	—	126
Backup ranking	6,966	—	909
Natural text, Qwen2.5-1.5B	554	2,080	11,318

Counts describe complete experimental runs, including comparator and integrity work, rather than the marginal cost of one deployed ablation. Suffix calls may batch branches and cannot be added as if they were independent full forwards.

H Linking the mean-ablation error to the change in first-token attention

This experiment tests whether changes in position-0 mass contribute to the complement-ablation decision mismatch in Section 3. It retains the same GPT-2 checkpoint, 26-head arXiv v1 circuit, eight

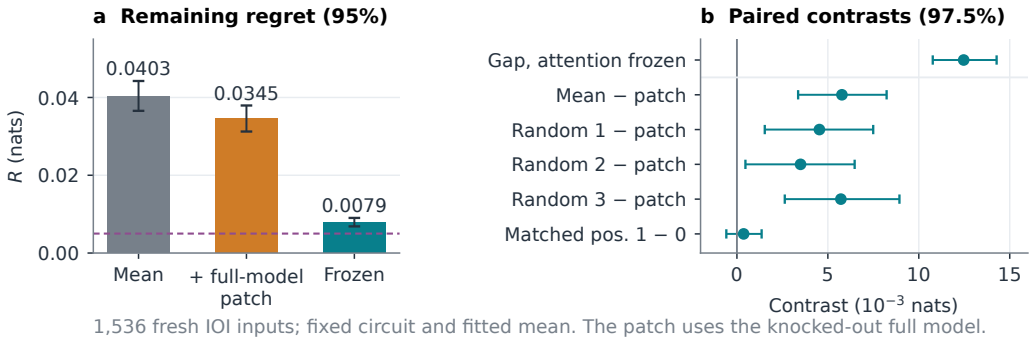


Figure 3: The change in first-token attention is linked to part of the mean ablation's decision error (1,536 fresh inputs; Section 4.2). (a) Regret with 95% intervals; the dashed line marks the tolerance. Patching in the full model's change in first-token attention lowers the mean ablation's regret, and freezing at clean activations gives lower regret still; both stay above the tolerance. (b) Paired contrasts with 97.5% intervals; positive values favor the link. Freezing first-token attention narrows the full-model effect gap between the two selected pairs, the patch beats three random patches, and matched position-0 and position-1 patches are indistinguishable. The contrasts in (b) are separate tests, not shares of one effect.

target heads, 28 candidate pairs, zero and norm-preserving resample knockouts, and 118 non-circuit heads. The mean ablation M uses the original 32-input fitting means, unchanged. The test uses **1,536 fresh IOI inputs**, disjoint from the historical 128-input evaluation and from all earlier inputs; evaluation inputs are not selected by correctness or intervention effects.

Design. A separate 64-input development set sets the number of inputs, 1,536, from the largest input-level standard deviation among the six primary contrasts below. The test includes all 12 branches and 57 conditions per input: an empty intervention and 28 pairs under each ablation type. The input is the bootstrap unit; both ablation types, all pairs, and all branches remain together in 8,000 resamples. Ablation types receive equal weight after making their selections separately. Primary intervals are 97.5%; descriptive intervals are 95%.

Full-model first-token attention freeze, and a patch into mean ablation. Write a head output as $z = av_0 + (1-a)u$, using Appendix C's notation. Superscripts c, d denote full-model clean and full-model knocked-out caches. The full-model first-token attention freeze K_0 replaces a non-circuit head's current output by

$$z_{K_0} = z + (a^c - a)(v_0 - u), \qquad q \geq 1.$$

Here a, v_0, u are recomputed on the current branch, including upstream effects of the freeze. Query 0 is unchanged. There is no subsequent full-output ablation; values, conditional non-sink routing, and downstream computation recompute. A separate patch branch first installs the complete original mean ablation and then adds

$$\delta_0 = (a^d - a^c)(v_0^d - u^d), \qquad q \geq 1,$$

to non-circuit heads, with zero increment at query 0. This head-space increment is distinct from the projected update δ_b of Appendix C. Thus its clean execution equals M exactly. This branch reads full-model knocked-out activations for every candidate pair, making it a diagnostic rather than a practical shortcut.

Three random controls preserve each projected increment's norm and angle to the unmodified mean-ablation write, randomizing its remaining direction within the non-circuit-head write subspace. They share the full-model donor information and use fixed seeds; degenerate cases are retained. The position control compares δ_0 with the analogous knocked-out-content δ_1. Both are restricted to $q \geq 2$, projected into the residual stream, and scaled to the smaller of their two norms before addition to M.

Endpoints. Each branch retains its own clean reference, $e_b(x, S, o) = L_b(x, S, o) - L_b(x, \varnothing)$, and selects $S_b^*(x, o) = \arg\max_S e_b(x, S, o)$ using fixed pair ordering for ties. All regrets evaluate that selection in the original full model: $R_b = \mathbb{E}[e_N(S_N^*) - e_N(S_b^*)]$. The first-token attention freeze endpoint holds the original full-model and mean selections fixed:

$$B_M = \mathbb{E}[e_N(S_N^*) - e_N(S_M^*) - \{e_{K_0}(S_N^*) - e_{K_0}(S_M^*)\}] .$$

It measures how the first-token attention freeze changes their effect gap, not K0's own regret or an additive mediated fraction. All inputs enter; matching full-model/mean choices contribute zero. Patch improvement is $T_R = R_M - R_{M+\delta_0}$.

Table 21: The 1,536-input test: the six primary contrasts.

Primary contrast, pooled across ablation types	Estimate	97.5% interval
Full-model first-token attention freeze B_M	0.012468	[0.010768, 0.014279]
Patch into mean ablation, T_R	0.005773	[0.003365, 0.008231]
Random 1 minus patch regret	0.004535	[0.001532, 0.007494]
Random 2 minus patch regret	0.003498	[0.000468, 0.006473]
Random 3 minus patch regret	0.005715	[0.002631, 0.008936]
Matched position 1 minus position 0 regret	0.000369	[−0.000583, 0.001358]

Five of the six lower endpoints exceed zero; the matched position contrast does not, so position specificity remains unresolved. The matched regret difference is exactly zero for 1,513 of 1,536 inputs. A reconstruction confirms that the two matched additions select identical pairs under both ablation types on these inputs, although their knockout-effect vectors differ. This equality of choices does not make the manipulations identities. The constrained position-0 branch also has little observed improvement over M, unlike the unrestricted patch. Because matching changes both query support and increment magnitude, this result neither establishes specificity nor demonstrates absence of the unrestricted channel.

Ablation faithfulness. Four factorial conditions cross clean or knocked-out content with clean or knocked-out first-token attention for non-circuit heads at $q \geq 1$: frozen, z^c; clean content with knocked-out mass, $z^c + (a^d - a^c)(v_0^c - u^c)$; knocked-out content with clean mass, $z^d - \delta_0$; and knocked-out content with knocked-out mass, z^d. Query 0 follows the content factor: it keeps its clean output in the two clean-content conditions and takes its output under the knockout in the two knocked-out-content ones. The last equals the full model by construction.

Table 22: Ablation faithfulness on the 1,536 fresh inputs.

Branch	Regret R, 95% interval	MAE D	Agreement
Original mean M	0.040283 [0.036564, 0.044239]	0.255304	75.9%
Frozen	0.007910 [0.006850, 0.009027]	0.073903	88.9%
Clean content, knocked-out mass	0.006684 [0.005677, 0.007713]	0.060356	90.4%
Knocked-out content, clean mass	0.002612 [0.002124, 0.003142]	0.038055	93.8%

Branch	Regret R, 95% interval	MAE D	Agreement
Full-model	0.002984 [0.002424, 0.003599]	0.040783	93.7%
first-token attention freeze K0			
Mean + δ_0	0.034510 [0.031243, 0.037940]	0.229200	77.6%
Matched position 0	0.040071 [0.036353, 0.043955]	0.252996	76.1%
Matched position 1	0.040440 [0.036722, 0.044362]	0.256433	76.0%

Clean accuracy on this cohort is 94.1% in the full model and in the four branches that keep each input's clean computation (frozen, both factorial branches, and K0), and 90.8% under M and the three branches built on it.

The patch reduces mean regret by 14.3%, with substantial residual error. On this fresh cohort, the pooled regret of the frozen and clean-content conditions both exceed the evaluation tolerance $\tau = 0.005$ with intervals above it; the historical smaller cohort was indeterminate. No calibration stage was run on this cohort, so this is not a new calibration acceptance decision. The knocked-out-content condition and K_0 retain knocked-out content or unfrozen non-circuit computation and do not demonstrate circuit sufficiency. The signed factorial interaction, $\mathbb{E}[e_N - e_{\text{knocked-out content}} - e_{\text{knocked-out mass}} + e_{\text{frozen}}]$, is -0.000811 [-0.001019, -0.000608] (95%). Absolute errors and their sequential reductions are not additive causal shares.

Ablation-type scope. The first-token attention freeze contrast B_M is larger under zero, 0.018652 [0.015458, 0.022095], than resample, 0.006284 [0.005077, 0.007591]. Patch remains positive under both: 0.006878 [0.002889, 0.010989] and 0.004668 [0.002178, 0.007611], respectively (97.5%). Random 2 versus patch is unresolved within zero, 0.003590 [-0.001394, 0.008600]; the pooled result does not establish superiority to every control separately under both ablation types. Matched-position specificity remains unresolved for each ablation type. The frozen model's regret is 0.009996 [0.008184, 0.011885] under zero knockouts, which insert the same value in the frozen and full models, with the whole interval above τ; under resample knockouts it is 0.005823 [0.004673, 0.007045], indeterminate relative to τ despite the pooled violation (95% intervals).

Scale and checks. The test scored 1,050,624 rows over 1,052,196 forward passes on CPU in float32 (batch 8, two threads). Validation reproduced the original fitted means and sampled historical full-model and mean NLLs exactly, and an independent re-analysis from the saved rows reproduces all primary estimates and intervals.

I Predicting conditional first-token attention effects on unseen knockouts

Summary (Section 6). A predictor built from mass change, update magnitude, and alignment with the terminal NLL readout aims to forecast the effect of freezing downstream first-token attention at clean values after a GPT-2 ablation. On disjoint groups of ablated heads it fails to beat a constant predictor (MAE 0.0524 versus 0.0479). Full-model gradients are accurate locally (MAE 0.0100). On the larger joint target, freezing first-token attention at every later layer at once, their MAE is 0.2131 against 0.2099 for a constant predictor. A score, based on attention change or on AtP, can therefore propose which heads to leave unfrozen (Section 3.3), but only the audit can say whether the circuit with its unfrozen heads works.

I.1 Design and estimands

We tested whether local mass, value, and readout geometry predicts sink-intervention effects beyond simpler controls. GPT-2 small used the pinned weights in Appendix B. Four-head output-zero ablations were allocated without effect-based selection across early (layers 0–2), middle (3–5), and late (7–9) source bands. Training, validation, and evaluation used 6, 3, and 12 ablations,

respectively: 84 mutually disjoint source-head identities. Per task, these splits contained 32, 16, and 64 independently generated prompts, with disjoint IOI name pools and induction token pools. Qualification and implementation-check prompts were separate; evaluation retained incorrect, weak-effect, and harmful cases. Evaluation prompts were paired across ablations. IOI scored one final-query target. Unlike the single-target induction protocol elsewhere, this experiment averaged eight next-token NLLs after at least five repeated-context tokens.

At layer l, full-model knocked-out quantities define the clean-mass correction

$$d_{lqh} = (a_{lqh}^{\text{clean}} - a_{lqh}^{\text{knockout}})(v_{0h}^{\text{knockout}} - u_{lqh}^{\text{knockout}}), \qquad \delta_{lq} = \sum_h W_{O,h} d_{lqh}.$$

The local target is $C_l = L(\text{knockout + local mass freeze}) - L(\text{knockout})$, with all other computation recomputed. Positive values mean full-model mass adaptation was beneficial relative to this freeze; negative values mean harmful. It is not a self-repair fraction. Freezes apply at queries $q \geq 1$, retaining recomputed values and relative non-sink routing. Local sites are 6/9/11 for early/middle ablations and 11 for late ablations.

The separate joint target freezes every eligible attention layer after the earliest source layer, excluding ablated heads. Predictions sum individual-layer predictions computed from full-model knocked-out features. This includes untrained depths and, sometimes, layers preceding later members of a multi-layer ablation. Joint effects are not assumed additive.

I.2 Predictors and evaluation

The candidate uses 13 features: task, source-depth mean/range, site, knocked-out-minus-clean NLL, signed/absolute mass movement, projected correction RMS, radial RMS, unscored-query norm fraction, readout alignment A_{geom}, cosine V, and readout norm. For scored queries Q_i,

$$A_{\text{geom}} = \sum_{q \in Q_i} r_{iq}^{\mathsf{T}} \delta_{lq}, \qquad V = \frac{A_{\text{geom}}}{\| \operatorname{vec}(r_{iQ_i}) \| \, \| \operatorname{vec}(\delta_{lQ_i}) \| + 10^{-12}},$$

where r_{iq} is the terminal NLL derivative, including the actual final LayerNorm and unembedding and the per-input reduction weight. This readout omits propagation through later layers, including contributions from earlier unscored queries. The candidate requires clean and full-model knocked-out passes and target labels; it forecasts an additional freeze, not the original knockout effect.

Feature scaling and ridge coefficients (penalty 0.001; unpenalized intercept) were fitted exclusively on training data and remained unchanged after validation. Controls use depth/knockout features, then mass, norm, or both; additional controls are a task/site constant, an answer-label permutation in the terminal readout, and paired-input mass substitution. The answer-label null retains the regression target while substituting another same-split input's answer IDs for the readout calculation. The Taylor references use full-model gradients: first-order attribution patching (Syed et al., 2024; Kramár et al., 2024) and our head-space and projected radial restrictions of it. Predictions were recorded before the target branches were run. Their additional backward cost is included in execution accounting.

All table intervals are paired, crossed 95% percentile bootstrap intervals from 8,000 replicates, resampling prompts within task and ablations within source band, with ablation draws shared across tasks. Sites are averaged within each prompt/ablation before equally weighting tasks and bands. Thus uncertainty uses 12 ablation groups, not thousands of independent intervention rows.

Table 23: Local and joint prediction errors with crossed 95% intervals.

Predictor	Local MAE [95% CI], nats	Joint MAE [95% CI], nats
Geometric candidate	0.0524 [0.0310, 0.0739]	0.2144 [0.1243, 0.3311]
Constant	0.0479 [0.0299, 0.0653]	0.2099 [0.1265, 0.3068]
Depth/knockout	0.0460 [0.0283, 0.0631]	0.2125 [0.1243, 0.3244]
Mass	0.0473 [0.0294, 0.0649]	0.2029 [0.1313, 0.2821]
Norm	0.0499 [0.0306, 0.0690]	0.2212 [0.1260, 0.3379]
Mass + norm	0.0501 [0.0307, 0.0695]	0.2092 [0.1241, 0.3153]
Answer-label readout null	0.0508 [0.0310, 0.0703]	0.1901 [0.1198, 0.2644]

Predictor	Local MAE [95% CI], nats	Joint MAE [95% CI], nats
Input-pairing null	0.0519 [0.0321, 0.0718]	0.2086 [0.1394, 0.2850]
Full Taylor	0.0100 [0.0052, 0.0143]	0.2131 [0.1143, 0.3204]
Head-space radial Taylor	0.0199 [0.0113, 0.0290]	0.1737 [0.1025, 0.2436]
Projected radial Taylor	0.0428 [0.0230, 0.0634]	0.2004 [0.1184, 0.3137]

The candidate does not improve on the depth, mass, norm, or mass-plus-norm controls: candidate-minus-mass-plus-norm MAE was +0.00228 [0.00025, 0.00512] locally and +0.00518 [−0.00126, 0.01583] jointly. Full Taylor's local advantage over the candidate was 0.04239 [0.02513, 0.05982] nats; the joint candidate-minus-Taylor difference was 0.00128 [−0.04704, 0.04313]. The local advantage therefore did not establish an advantage over the candidate for the larger joint target; joint MAE was also close to the constant baseline's 0.2099.

With a 0.02-nat weak threshold, nonweak target coverage was 0.404 [0.337, 0.468] locally and 0.806 [0.740, 0.862] jointly. Candidate balanced sign scores were 0.140 [0.104, 0.173] locally and 0.402 [0.358, 0.439] jointly, with beneficial/harmful recalls of 0.225/0.054 and 0.784/0.021: the candidate does not predict the sign of the effect. Weak predictions count as failures on nonweak targets, so these scores should not be compared with a 50% random-sign benchmark. Each nonweak sign had at least eight observations in 8 of 12 ablation groups locally and 11 of 12 jointly.

I.3 Geometry interventions and scope

A fixed subset of 16 prompts per task crossed all 12 evaluation ablations (384 instances). At scored queries, we reflected the projected correction across the terminal-readout normal, or applied a random correction matched in norm and angle from the full-model correction. Unscored queries retained the full-model correction. These synthetic output interventions were crossed with recomputation or freezing the current MLP and all later attention/MLP outputs at full-model knocked-out values; residual skips and final LayerNorm recomputed. Frozen infinitesimal reflection effects are partly guaranteed by construction.

Let E_x denote the mean scored-loss effect of intervention x, and $\Delta_{\mathrm{geom}} = E_{\mathrm{random}} - E_{\mathrm{reflected}}$. The contrast with recomputation was -0.0993 [-0.2299, -0.0078]; frozen Δ_{geom} was -0.1758 [-0.4003, -0.0114]. Their interaction was +0.0765 [0.0026, 0.1700]. Centered two-sided bootstrap tests gave raw $p_{\mathrm{recomputed}} = 0.0676$ and $p_{\mathrm{interaction}} = 0.0666$, and Holm-adjusted $p = 0.1332$ for both, so neither is established at the 0.05 level. These approximate tests and percentile intervals are not exact inverses. Geometry effects were not limited to recomputation. Unadjusted task-stratified interaction estimates were: induction -0.0006 [-0.0242, 0.0242], IOI +0.1536 [0.0167, 0.3258].

On the same subset, joint-minus-summed-individual effects averaged -0.0247 [-0.0935, 0.0383], while mean absolute nonadditivity was 0.1376 [0.0808, 0.2040]. Signed cancellation therefore conceals substantial discrepancies. On all 64 prompts per task, matched-support position-0 minus position-1 effects, both using $q \geq 2$, were 0.0140 [-0.0008, 0.0287]. Half-dose curvature, $E_{1/2} - 0.5E_1$ with $E_{1/2}$ the effect of the correction applied at half scale, was -0.00177 [-0.00367, -0.00062].

These results limit this fixed predictor and ablation panel; they establish neither a general sign rule nor impossibility of conditional prediction. Total execution used 22,680 forwards and 1,644 backwards, and an independent re-analysis from the saved rows reproduces the primary estimates.

J Fixed-pair selection from a first-token attention decrease score

Summary (Section 6). Can first-token attention decrease, without backup labels, select an important pair? Scoring every head of layers 10–11 (all 24, not the 24 heads of Section 3.3 ranked by attention change) by that decrease selects L10H0 and L11H3. The pair is chosen on separate inputs; evaluation uses 128 fresh inputs and four held-out test ablations; these overlap the scoring ablations without repeating them (each contains L8H6 or L9H9) and were screened for a large NLL increase on other inputs. Under zero ablation, removing this pair is 0.1435 nats more harmful after the test ablations

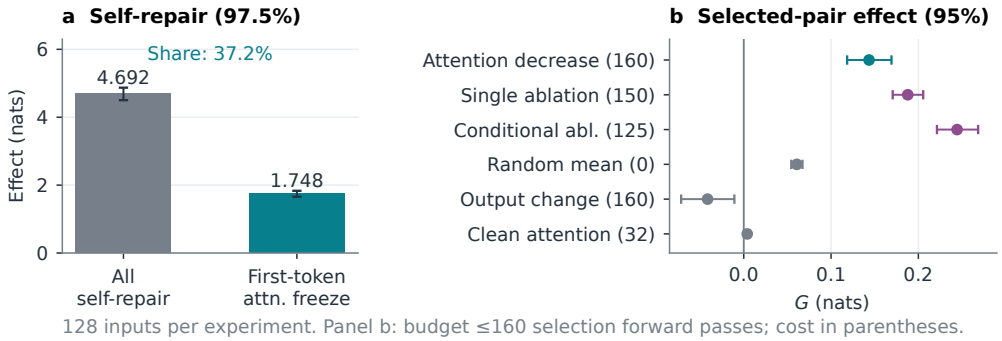


Figure 4: First-token attention carries part of the self-repair, but selecting a pair by its decrease is weaker than ablation-based selection. (a) Freezing first-token attention in 23 downstream heads removes 1.748 of 4.692 nats of measured self-repair (97.5% intervals; 128 inputs; the result of Figure 1c). (b) Conditional pair effect G for six selectors within 160 forward passes (95% intervals; 128 fresh inputs); "clean attention" uses clean first-token attention alone. Single and conditional ablation beat both first-token attention selectors at this budget.

than on clean inputs; call this conditional-importance increase G. G and its differences from the mean of 20 random pairs and from a head-output-change selector all have intervals above zero. Single and conditional ablation select pairs with larger G, 0.1879 and 0.2444, using 150 and 125 selection forward passes versus 160 for the decrease score (Figure 4b). L10H0 is a known name mover, L11H3 is a backup in the later ICLR 2023 circuit version, and the pair already matters on clean inputs.

This experiment tests first-token attention selection without backup labels. It selects L10H0 and L11H3, with larger conditional effects than random and output-change controls but smaller effects than ablation-based selectors. L10H0 is a known name mover and L11H3 is a backup name mover in the ICLR 2023 circuit (Appendix B); this supports diagnostic relevance, not a newly discovered backup.

J.1 Qualification, selection, and evaluation

We use GPT-2 IOI with three disjoint input sets of 32 qualification, 32 score-fitting, and 128 evaluation inputs, whose token sequences are distinct from each other and from all earlier experiments. A source coalition P contains three heads in L8 or three in L9; ablation zeros their outputs before output projection at every query position. On the qualification inputs, we measure 64 randomly drawn candidate coalitions, 32 per layer, and retain the four per layer with the largest mean NLL increase. The experiment proceeds only if qualification clean accuracy is at least 0.8 and every selected coalition's mean NLL increase is at least 0.05 nats; accuracy is 0.9375 and the selected NLL increases range from 0.3035 to 1.0532 nats.

A random partition assigns two selected coalitions per layer to score fitting and two to evaluation:

Table 24: Qualified three-head ablations assigned to score fitting and evaluation.

Layer	Score-fitting coalitions, head indices	Evaluation coalitions, head indices
L8	{0,6,11}; {6,9,11}	{2,3,6}; {3,5,6}
L9	{2,4,9}; {7,8,9}	{1,3,9}; {2,7,9}

The evaluation coalitions are **held out from score fitting but were qualified on the same qualification inputs**, so their NLL increase had already been observed before evaluation. Every selected L8 coalition contains L8H6 and every selected L9 coalition contains L9H9. Thus the test is neither head-disjoint nor representative of an unqualified ablation distribution.

All scored methods share 24 downstream candidates in L10/L11 and select the highest-scoring head in each layer. The first-token attention decrease score averages clean-minus-knocked-out position-0 attention mass over the score-fitting inputs, four training coalitions, and queries $q \geq 1$. Comparators use clean first-token attention, head-output change magnitude, clean single-head ablation loss, or the increase in single-head ablation effect after primary knockout (the G definition below with singleton H). The conditional baseline is our implementation of this score, not a reproduction of another discovery system. Twenty random pairs are drawn once; one of them coincides with the decrease-selected pair, so the contrasts against their mean are slightly conservative. No backup labels enter selection or evaluation. Every selected pair, including the budget-specific choices, is chosen on separate inputs and held fixed across evaluation inputs and coalitions.

An independent re-analysis from the saved rows reproduces the summaries below.

J.2 Conditional ablation importance

Let L denote correct-answer NLL at the final position. For pair H and coalition P, we evaluate

$$G(H, P, x) = [L(P \cup H, x) - L(P, x)] - [L(H, x) - L(\text{clean}, x)].$$

Positive G means removing H is more harmful after P than in the clean computation. It does not itself prove adaptation, beneficial self-repair, or improved accuracy. For the selected pair, clean-only ablation costs 0.2886 nats and ablation after P costs 0.4321, giving G = 0.1435. The pair is already important before the knockout.

The three primary tests are one positivity test for G and two baseline comparisons. We use 8,000 paired-input bootstrap replicates, retaining all branches and four equally weighted test coalitions within each resampled input. Each primary interval is $98\frac{1}{3}\%$, targeting nominal 95% familywise coverage by Bonferroni adjustment; all three intervals lie above zero.

Table 25: Primary conditional-effect tests for the fixed decrease-selected pair.

Primary quantity	Mean, nats	$98\frac{1}{3}\%$ CI
First-token attention decrease G	0.14353	[0.11321, 0.17431]
G minus mean G of 20 random pairs	0.08277	[0.04909, 0.11649]
G minus output-change G	0.18497	[0.12678, 0.24380]

The full comparison on the same 32 score-fitting inputs includes stronger ablation baselines:

Table 26: Complete selector comparison using the same 32 fitting inputs.

Selection method	Selected pair	Selection forward passes	Evaluation mean G	95% CI
First-token attention decrease	L10H0 + L11H3	160	0.14353	[0.11842, 0.16927]
Mean of 20 random pairs	Twenty pairs	0	0.06076	[0.05443, 0.06726]
Output change	L10H7 + L11H8	160	−0.04145	[−0.07170, −0.01075]
Clean first-token attention	L10H8 + L11H9	32	0.00397	[0.00099, 0.00692]
Single ablation	L10H6 + L11H0	800	0.18787	[0.17062, 0.20556]
Conditional ablation	L10H2 + L11H2	4,000	0.31533	[0.29154, 0.33839]

Paired decrease-minus-single is -0.04434, 95% CI $[-0.07298, -0.01546]$; decrease-minus-conditional is -0.17180 $[-0.21004, -0.13249]$. Random selection costs zero forwards: the random mean is not a same-cost comparator. Two random pairs have larger point G than the decrease-selected pair, which exceeds the ensemble mean, not every draw.

J.3 Selection budgets

Standalone forward-pass costs for n fitting inputs are $5n$ for the first-token attention decrease and output change, n for clean first-token attention, $25n$ for single ablation, and $125n$ for conditional ablation. Each budget uses the longest affordable prefix of the score-fitting inputs, capped at 32. At a budget of 160, single and conditional ablation use six and one inputs (150 and 125 forwards). Single ablation selects its 32-input pair; conditional ablation selects L10H10 + L11H6 with $G = 0.24438$ [0.22131, 0.26849; 95%]. Decrease-minus-conditional is -0.10085 $[-0.13367, -0.06776]$. Both ablation baselines outperform the decrease score with fewer forward passes at this budget.

At budgets 640 and 2,560, conditional ablation uses five and twenty inputs and selects its 32-input pair; single ablation also retains its pair. The decrease score is capped at 32 fitting inputs with $G = 0.14353$. Costs exclude the shared measurement harness and the final evaluation. Wall-clock latency was not measured, and no 128-input fitting extension was run.

J.4 Selective mass clamps and interpretation

For the selected pair and the 20 random pairs, we additionally measure a clamp loss difference (a scalar contrast, not the first-token attention freeze K_0 of Appendix H)

$$K_p(H, P) = L(P, \mathrm{clamp}_p(H)) - L(P).$$

The recursive clamp fixes attention mass at its same-input clean value while values and relative attention over the other positions recompute. Positions 0 and 1 use the same heads and $q \geq 2$ queries, matching identity and support but **not intervention magnitude**; later components recompute.

Table 27: Selective mass-clamp effects for the chosen pair and random controls.

Auxiliary quantity	Mean, nats	CI	Level
Selected pair, clamp K_0	0.05733	[0.04755, 0.06795]	95%
Selected pair, clamp K_1	0.00063	[0.00050, 0.00076]	95%
Selected pair, $K_0 - K_1$	0.05670	[0.04544, 0.06857]	97.5%
Selected pair K_0 minus mean K_0 of 20 random pairs	0.02551	[0.01487, 0.03671]	97.5%

The larger position-0 effect establishes causal relevance under this clamp, not an exhaustive decomposition: G uses zero ablation, K clamps mass, and K/G is neither a mediation nor a self-repair fraction. Intervals resample evaluation inputs, conditional on the fixed qualified coalition panel, fitted choices, and the 20 random pairs; they do not include variation over new qualification or score-fitting samples, ablation populations, or models.

K The self-repair measurement under complement ablation

The protocol of Sections 3.5 and 4.3 (zeroing the three name movers; total, direct, and self-repair effects; the first-token attention freeze; and the conditional importance of the eight documented backups, $G_h = [L(\mathrm{NM3} \cup h) - L(\mathrm{NM3})] - [L(h) - L(\mathrm{clean})]$) was repeated with the 118 heads outside the Appendix B circuit replaced at every layer and position, under three conditions: the full model (all heads recompute), mean ablation with the fitted means of Section 3 (per position,

as in the IOI completeness test), and freezing at the same input's clean outputs. Downstream freezing operations act on the downstream components that recompute under each condition: all 23 downstream heads in the full model, the eight downstream circuit members under either ablation, and the layer 9–11 MLPs and final LayerNorm throughout. Each condition has its own clean reference. Two cohorts of 128 new inputs were used; the second repeats the first. Distortion is the mean over the eight documented backups of $|G_h^b - G_h^N|$ per input; contrasts are paired 97.5% intervals.

Table 28: Name-mover measurements under three conditions on two cohorts of 128 new inputs (name movers zeroed; the main text's 'about twice as harmful' and 'by half or more' are the ratios and distortions here). Entries are full model / mean ablation / frozen unless stated.

Quantity	Cohort 1	Cohort 2
Total effect TE (nats)	0.588 / 1.118 / 0.752	0.532 / 1.095 / 0.711
TE ratio to the full model	1 / 1.90 / 1.28	1 / 2.06 / 1.34
Self-repair SR (nats)	4.660 / 4.870 / 4.496	4.597 / 4.673 / 4.418
First-token attention share of SR	0.371 / 0.423 / 0.359	0.378 / 0.427 / 0.368
Backups with resolved G (of 8)	5 / 6 / 6	5 / 6 / 6
Distortion of G (nats per head)	— / 0.151 / 0.067	— / 0.151 / 0.068
Distortion, mean ablation minus frozen	0.085 [0.076, 0.094]	0.082 [0.074, 0.091]
TE, mean ablation minus frozen	0.365 [0.314, 0.419]	0.384 [0.324, 0.442]
SR, full model minus mean ablation	-0.210 [-0.314, -0.115]	-0.076 [-0.169, +0.015]

Under mean ablation the knockout reads about twice as harmful, the first-token attention share rises to 0.42 although only eight downstream circuit heads recompute, and L10H1 enters the set of backups with resolved conditional importance (a backup is resolved when its 97.5% interval lies above zero). Under freezing, the distortion of that importance falls by 56% and 55% and the inflation falls to 1.28× and 1.34×, but L10H1 still enters. The measured self-repair is not smaller under fitted means: full model minus mean ablation is -0.210 [-0.314, -0.115] in the first cohort and -0.126 [-0.242, -0.016] in the third (introduced below); it is unresolved in the second and borderline in the fourth, whose upper bound is -0.001 or +0.001 depending on the bootstrap seed. Backup ranks are preserved under fitted means (Spearman 0.98 and 0.95 against the full model on the two cohorts; L11H2 first throughout). These measurements use the circuit and fitted means of Sections 3.1–3.3 and do not test the original paper's ablation protocol, which Appendices N and O address.

Circuit with unfrozen heads. A third cohort of 128 new inputs repeats the protocol under two further ablations in which 24 non-circuit heads recompute and the other 94 are frozen at their same-input clean outputs. The 24 are chosen on the 32 fitting inputs of Section 3 from one clean pass and eight single-head zero ablations of the candidate heads, by either the attention-change score of Section 3.3 (the mean absolute change in position-0 mass at queries $q \geq 1$) or an AtP estimate of the effect of freezing each head's reaction on the answer NLL (the gradient of the NLL with respect to the head output, contracted with the head's output change under the knockout; a central-difference check on the largest term agrees within 1%). The two sets share 13 heads; the attention-change set has 10 heads in the downstream layers and the AtP set 7, so 18 and 15 of the 23 downstream heads are unfrozen and subject to the freezing operations. Without knockout both ablations reproduce the full-model clean head outputs exactly.

Table 29: Name-mover measurements on the third cohort of 128 new inputs (name movers zeroed). Contrasts are paired 97.5% intervals.

Quantity	Full model	Mean ablation	Frozen	Frozen + 24 unfrozen, attention change	Frozen + 24 unfrozen, AtP
Total effect TE (ratio to the full model)	0.514 (1)	1.023 (1.99)	0.692 (1.34)	0.507 (0.99)	0.500 (0.97)
TE ratio under resampling	1	1.36	0.95	0.98	0.99
Self-repair SR	4.603	4.729	4.426	4.610	4.617
First-token attention share of SR	0.367	0.412	0.357	0.359	0.366
Backups with resolved G (of 8)	5	6 (+L10H1)	6 (+L10H1)	5 (the full model's five)	5 (the full model's five)
Distortion of G (nats per head)	—	0.148	0.067	0.007	0.008
Spearman of G against the full model	—	0.95	0.90	1.00	1.00

Relative to the frozen model, unfreezing the 24 attention-change heads reduces the distortion by 0.060 [0.053, 0.066] nats per head (89%) and the total effect by 0.185 [0.159, 0.210] nats, and its self-repair exceeds the full model's by 0.007 [0.002, 0.013] nats; the AtP score gives 0.059 [0.053, 0.066], 0.192 [0.165, 0.217], and 0.014 [0.009, 0.020]. The mean-ablation and frozen columns reproduce the two earlier cohorts (distortion 0.148 and 0.067, ratios 1.99 and 1.34, L10H1 entering under both), so the residual of the frozen model is carried by the reactions of the non-circuit heads that the score selects. Under both sets of unfrozen heads L10H1's interval lies slightly below zero (-0.029 [-0.049, -0.007] with the attention-change score), so it is not resolved.

Fourth cohort, second knockout set, and a structural control. A fourth cohort of 128 new inputs repeats the unfrozen-head measurement with the same sets and adds a structural control: the 15 non-circuit heads of layers 10–11 left unfrozen, all other non-circuit heads frozen. A fifth cohort of 128 inputs applies the same ablations to a second knockout set, zeroing L9H9 alone (its downstream set is then all 24 heads of layers 10–11 and the layer 9–11 MLPs).

Table 30: Measurements with unfrozen heads on the fourth cohort (three name movers zeroed) and the fifth cohort (L9H9 zeroed). Entries are mean ablation / frozen / frozen + 24 unfrozen (attention change) / frozen + the 15 layer 10–11 non-circuit heads unfrozen.

Quantity	Three name movers, cohort 4	L9H9 alone, cohort 5
TE ratio to the full model	2.01 / 1.32 / 0.99 / 1.00	1.83 / 1.21 / 0.99 / 1.00
Distortion of G (nats per head)	0.151 / 0.067 / 0.007 / 0.000	0.065 / 0.025 / 0.004 / 0.000
Backups with resolved G (full model: 5 and 6)	6 (+L10H1) / 6 (+L10H1) / 5 / 5	6 / 6 / 6 / 6
Spearman of G against the full model	0.95 / 0.90 / 1.00 / 1.00	0.98 / 0.98 / 0.98 / 1.00

The fourth cohort reproduces the third (distortion contrast, frozen minus unfrozen heads, 0.060 [0.053, 0.068] nats per head; total-effect contrast 0.181 [0.153, 0.208]), and the second knockout set gives the same ordering (0.021 [0.018, 0.025] and 0.083 [0.071, 0.096]); under L9H9 alone L10H1 is a backup in the full model, so no ablation changes the backup set. The layer control reproduces the full-model reading exactly in both cases: knockout at layers 9–10 can be answered only by non-circuit heads at layers 10–11, so keeping all 15 unfrozen is the full model for this measurement, and the residual of the circuit with its unfrozen heads comes from the five layer-10–11 non-circuit heads it does not include (L10H8, L11H0, L11H5, L11H7, L11H8). The 24-head AtP set, run on the fourth and fifth

cohorts as well, gives total-effect ratios 0.98 and 0.97 and distortions 0.008 and 0.004 nats per head, and recovers the full-model backup sets. Under resample knockouts the mean-ablation ratio is 1.39, 1.35, 1.36 and 1.34 on the four three-name-mover cohorts. The pair-choice audit of Sections 3.1–3.3, whose knockout spans layers 3–10, does not have the layer control's property; there unfreezing the same heads is not accepted (Appendix L). Section 3.3 reports the same score-selected sets of unfrozen heads under the pair-choice audit.

L UNFREEZING SELECTED HEADS, AND CHANNEL-SPLIT ABLATIONS

Design. The complement-ablation audit of Section 3 (28 candidate pairs of eight candidate heads, zero and norm-preserving resample knockouts, 57 conditions, tolerance 0.005 nats) was repeated on 32 fresh calibration and 256 fresh evaluation inputs with ablations that keep part of the non-circuit heads unfrozen over the frozen model. The k non-circuit heads ($k = 8, 16, 24, 48$) with the largest score on the 32 fitting inputs recompute, the score being the attention-change score (mean absolute change in position-0 mass at queries $q \geq 1$ under eight single-candidate-head zero ablations; Section 3.3), an attribution-patching (AtP) estimate of the effect of freezing the head's reaction on the answer NLL (the NLL gradient with respect to the head output contracted with its output change under the knockout; a central-difference check on the largest term agrees within 1%, and the AtP estimate of freezing all non-circuit heads correlates with the measured effect at $r = 0.92$ over 256 input–ablation pairs), or the norm of the head-output change; 24 random non-circuit heads (three draws) and the 15 non-circuit heads of layers 10–11 are controls. Only the 74 non-circuit heads in layers 4–11 lie downstream of the eight candidate heads, whose earliest head is L3H0; all three scores are exactly zero for the 44 non-circuit heads in layers 0–3, so every score-selected set of unfrozen heads lies in layers 4–11, whereas the random sets contain 9 to 12 heads from layers 0–3. The remaining non-circuit heads are frozen at the same input's clean outputs. Keeping none reproduces the frozen model and keeping all 118 reproduces the full model (both identities hold to 10^{-6}). Two earlier variants motivated this design: mean-replacing the remaining heads instead reduced regret by only 40% at $k = 24$ and was not monotone in k (at $k = 48$ regret exceeded mean ablation), and adding each non-circuit head's per-condition mean full-model output change under the knockout to its frozen output, a constant correction fitted on the same 32 inputs, reduced MAE by 16% but left regret unchanged.

Table 31: Complement-ablation audit on 256 fresh evaluation inputs with partly unfrozen non-circuit heads over the frozen model (and the three channel-split ablations). D is MAE, R decision regret (nats), and the calibration column the one-sided 95% upper bound on 32 fresh calibration inputs, starred when at most the 0.005-nat tolerance (accepted); clean accuracy is 95.7% for the full model and every ablation except mean ablation (93.0%).

Condition	R, 95% interval	D	Agreement	Calibration	Evaluation
Full model	0.0000 [0.0000, 0.0000]	0.000	100.0%	0.0000*	Supported
Mean ablation	0.0410 [0.0326, 0.0502]	0.247	75.2%	0.0774	Violated
Frozen	0.0063 [0.0040, 0.0089]	0.072	90.4%	0.0212	Indeterminate
Frozen + 8, attention change	0.0036 [0.0018, 0.0057]	0.044	94.3%	0.0216	Indeterminate
Frozen + 8, AtP	0.0024 [0.0009, 0.0044]	0.050	95.3%	0.0030*	Supported
Frozen + 16, attention change	0.0028 [0.0014, 0.0044]	0.040	94.3%	0.0030*	Supported
Frozen + 16, AtP	0.0162 [0.0107, 0.0224]	0.051	91.2%	0.0315	Violated
Frozen + 24, attention change	0.0022 [0.0011, 0.0036]	0.040	94.5%	0.0030*	Supported

Condition	R, 95% interval	D	Agreement	Calibration	Evaluation
Frozen + 24, AtP	0.0065 [0.0036, 0.0101]	0.045	94.1%	0.0246	Indeterminate
Frozen + 48, attention change	0.0237 [0.0179, 0.0298]	0.044	87.3%	0.0614	Violated
Frozen + 48, AtP	0.0001 [0.0000, 0.0002]	0.010	98.8%	0.0009*	Supported
Frozen + 24, output norm	0.0154 [0.0108, 0.0206]	0.052	89.8%	0.0414	Violated
Frozen + 24, random 1	0.0084 [0.0051, 0.0125]	0.074	90.4%	0.0212	Violated
Frozen + 24, random 2	0.0045 [0.0025, 0.0068]	0.061	92.4%	0.0058	Indeterminate
Frozen + 24, random 3	0.0055 [0.0034, 0.0080]	0.075	90.8%	0.0116	Indeterminate
Frozen + non-circuit L10–L11	0.0037 [0.0019, 0.0059]	0.039	94.7%	0.0054	Indeterminate
Recomputed routing	0.0112 [0.0071, 0.0158]	0.068	90.4%	0.0255	Violated
Recomputed values	0.0073 [0.0047, 0.0101]	0.079	89.6%	0.0123	Indeterminate
Recomputed first-token attention	0.0069 [0.0041, 0.0102]	0.062	90.2%	0.0164	Indeterminate

Table 32: Regret contrasts (nats; positive favors the second ablation; "attention 24" abbreviates the frozen model with 24 unfrozen heads ranked by attention change, and likewise for the other scores).

Contrast (regret, nats)	Estimate	97.5% interval
Frozen − attention 24	0.00403	[0.00170, 0.00680]
Random 24 (mean of three) − attention 24	0.00392	[0.00164, 0.00666]
Attention 24 − attention 48	−0.02145	[−0.02885, −0.01471]
Attention 24 − AtP 24	−0.00428	[−0.00849, −0.00075]
Output change 24 − attention 24	0.01322	[0.00781, 0.01926]
Non-circuit L10–L11 − attention 24	0.00150	[−0.00067, 0.00396]
Frozen − AtP 24	−0.00025	[−0.00496, 0.00428]
Random 24 (mean of three) − AtP 24	−0.00036	[−0.00442, 0.00374]
Mean ablation − frozen	0.03474	[0.02546, 0.04525]

Table 33: Regret and MAE against the number of unfrozen non-circuit heads k.

k	R, attention-change score	D, attention-change score	R, AtP score	D, AtP score
8	0.0036	0.044	0.00245	0.050
16	0.0028	0.040	0.01623	0.051
24	0.0022	0.040	0.00650	0.045
48	0.0237	0.044	0.00009	0.010

Leaving unfrozen the 24 non-circuit heads with the largest attention-change score reduces regret from 0.0063 to 0.0022 nats (contrast 0.0040 [0.0017, 0.0068]), raises pair agreement from 90.4% to 94.5%, and on this input set is accepted on calibration (one-sided upper 0.0030) and supported on evaluation;

the 16-head set is accepted as well. Three random sets of 24 average 0.0061 (contrast 0.0039 [0.0016, 0.0067] against the attention-change set), the output-change score is worse than the frozen model, and the 15 non-circuit heads of layers 10–11 come close (0.0037; calibration upper 0.0054, not accepted). Regret is not monotone in k: 48 heads by the attention-change score raise regret to 0.0237 while leaving MAE at 0.044, and the AtP score is accepted at 8 and 48 heads (0.0024 and 0.00009 nats, the latter with 98.8% agreement) but not at 16 or 24. The increases occur under zero ablation only (48 attention-change heads: 0.0473 under zero ablation, 0.00002 under resampling; 16 AtP heads, recounted from the saved rows: 0.0318 and 0.0007). One untested explanation is that unfrozen heads respond to the knockout while heads they interact with stay frozen. The circuit with its unfrozen heads must itself pass the audit; k is not a free parameter. Its residual regret partly measures the remaining incompleteness of the circuit with its unfrozen heads for these knockouts.

A second input set. A second fresh set of 32 calibration and 256 evaluation inputs (newly drawn random controls; the scored sets, recomputed on the same fitting inputs, are identical) repeats the primary comparison. It re-tested the 24 attention-ranked heads, not the top 8 AtP heads, which also passed on the first set. The attention-change 24-head set again reduces regret (0.0064 to 0.0038; contrast 0.0026 [0.0007, 0.0048]) and beats the random sets (mean 0.0105; contrast 0.0067 [0.0038, 0.0098]), but is not accepted on this set (calibration upper 0.0078; evaluation interval [0.0020, 0.0059] crosses the tolerance), and neither is the 16-head set. The 48-head AtP set is accepted and supported on both sets (0.0001 and 0.0003 nats; 98.8% and 98.2% agreement), so of the groups re-tested here, only the 48-head AtP set meets the tolerance on both sets; the attention-change 24-head set passes on the first set but not the second, although it brings the point estimate of the name-mover knockout effect of Appendix K to 0.99 of the full model's. Unfreezing the layer 10–11 non-circuit heads gives 0.0052 (not accepted). Acceptance on 32 calibration inputs is a heavy-tailed statistic: the same 24 heads give calibration bounds of 0.0030 and 0.0078 on the two sets.

Table 34: The second input set (256 fresh evaluation inputs; same columns as the previous audit table; clean accuracy 91.8% for the full model and every ablation except mean ablation, 87.1%).

Condition	R, 95% interval	D	Agreement	Calibration	Evaluation
Full model	0.0000 [0.0000, 0.0000]	0.000	100.0%	0.0000*	Supported
Mean ablation	0.0448 [0.0367, 0.0534]	0.254	73.0%	0.0505	Violated
Frozen	0.0064 [0.0041, 0.0091]	0.075	90.0%	0.0066	Indeterminate
Frozen + 16, attention change	0.0035 [0.0019, 0.0054]	0.042	93.2%	0.0125	Indeterminate
Frozen + 24, attention change	0.0038 [0.0020, 0.0059]	0.042	93.4%	0.0078	Indeterminate
Frozen + 48, AtP	0.0003 [0.0000, 0.0007]	0.010	98.2%	0.0000*	Supported
Frozen + non-circuit L10–L11	0.0052 [0.0029, 0.0081]	0.041	92.6%	0.0094	Indeterminate
Frozen + 24, random 1	0.0132 [0.0085, 0.0185]	0.079	88.5%	0.0281	Violated
Frozen + 24, random 2	0.0113 [0.0078, 0.0152]	0.085	87.5%	0.0125	Violated
Frozen + 24, random 3	0.0069 [0.0045, 0.0098]	0.069	89.6%	0.0067	Indeterminate

Channel split. Three further ablations keep all 118 non-circuit heads' routing recomputed over clean values, their values recomputed over clean routing, or only their position-0 mass recomputed with clean conditional routing and values (the practical analogue of the replay of Appendix D.4). Without knockout each equals the full model.

Table 35: Channel-split contrasts against the frozen model (nats).

Contrast	Estimate	97.5% interval
R: frozen minus recomputed routing	-0.00495	$[-0.01056, 0.00018]$
R: frozen minus recomputed values	-0.00101	$[-0.00319, 0.00097]$
R: frozen minus recomputed first-token attention	-0.00061	$[-0.00395, 0.00223]$
D: frozen minus recomputed routing	0.00463	[0.00056, 0.00886]
D: frozen minus recomputed values	-0.00621	$[-0.00854, -0.00392]$
D: frozen minus recomputed first-token attention	0.01049	[0.00854, 0.01251]
R: recomputed values minus recomputed routing	-0.00394	$[-0.00975, 0.00135]$

None of the three changes regret relative to the frozen model (contrasts -0.0050 $[-0.0106, 0.0002]$, -0.0010 $[-0.0032, 0.0010]$, and -0.0006 $[-0.0040, 0.0022]$ for recomputed routing, recomputed values, and recomputed first-token attention); recomputed routing over clean values raises zero-ablation regret from 0.0083 to 0.0212. Recomputed first-token attention reduces MAE by 14.5% (0.0105 [0.0085, 0.0125] nats), close to the 16.0% of the replay of full-model first-token attention in Appendix D.4 (on other inputs), without reading any full-model knocked-out activation, whereas recomputed values raise it by 8.6%. MAE moves with a single channel or with the constant correction above; in these tests, regret improves only when selected heads are unfrozen in both channels.

M TWO MORE CIRCUITS

This appendix gives the design and full results behind Sections 3.4 and 3.5 and the amplification noted in Section 6.

M.1 CIRCUITS, INPUTS, AND DESIGN

Circuits. Both are GPT-2 small circuits whose heads are listed in peer-reviewed papers. The acronym circuit (García-Carrasco et al., 2024) has eight heads, L1H0, L2H2, L4H11, L5H8, L8H11, L9H9, L10H10, and L11H4, of which L8H11, L9H9, L10H10, and L11H4 move the letters; its authors evaluated it by mean-ablating the heads outside it. All eight are candidate heads (28 pairs). The greater-than circuit (Hanna et al., 2023) has the ten heads that the text of that paper assigns to its full circuit, L0H1, L0H3, L0H5, L5H1, L5H5, L6H9, L7H10, L8H8, L8H11, and L9H1 (its Section 3.2 and Appendix B; its full-circuit diagram, Figure 14, instead draws L6H1 and omits L5H1 and L8H8); the seven that feed MLPs 8 and 9 (L5H1, L5H5, L6H9, L7H10, L8H8, L8H11, and L9H1) are the candidate heads (21 pairs). As in IOI, MLPs are never ablated, and the other 136 and 134 heads are the non-circuit heads. Of these, 113 and 67 lie in layers after the first candidate head (74 in IOI) and can react to the knockouts; unfreezing all of them reproduces the full model, and no set larger than 48 was tested.

Inputs and endpoints. Acronym inputs follow the construction of García-Carrasco et al. (2024): the letters X, Q, and U are excluded; each word is a space-and-capital token followed by one lowercase token; and the three capitals tokenize separately (1,154 acronyms, as in the original). The words are WordNet noun lemmas (Miller, 1995) that satisfy these constraints, appear in the web2 word list (Webster's Second International Dictionary), and are single GPT-2 tokens in lowercase (422 words, after removing a short list of offensive words). The prompt "The C1T1 C2T2 C3T3 (A1A2" follows the end-of-text token, and the target is the third capital. Greater-than inputs use the 120 nouns and the template of Hanna et al. (2023), with start years from 1000–1899 that split into century and two-digit tokens and end in 02–98 (771 years; the original reports 768). The endpoint is the NLL of the answer token for acronyms and the NLL of the set of valid years, $-\log \sum_{y>YY} p(y)$, for greater-than; the original papers use the logit or probability difference. Clean accuracy is whether the answer has the largest logit among the 26 capitals, or whether the most probable two-digit year exceeds the start year. On the 32 qualification inputs of each circuit the full model answers all correctly; mean ablation answers 30 correctly for acronyms and all 32 for greater-than.

Design. Everything else follows Appendices A, C, and L: zero and norm-preserving resample knockouts of every candidate pair; means fitted per layer, position, and head on 32 clean inputs; the

frozen model and its variants with the top 8, 16, 24, or 48 non-circuit heads by AtP or the top 24 or 48 by attention change unfrozen, scored on the fitting inputs; one set of 24 and two sets of 48 random heads drawn only from the non-circuit heads downstream of the knockouts; and the non-circuit heads of layers 10–11. The calibration rule and tolerance are unchanged. Each circuit has two cohorts of 32 calibration and 256 evaluation inputs, with all sequences distinct across splits. Every knocked-out branch recomputes the blocks from the earliest knocked-out layer on the ablated model's own clean residual stream, which is exact because nothing upstream differs.

M.2 The audit

Table 36: The audit in the acronym and greater-than circuits (regret in nats, 95% intervals where shown; agreement with the full model's choice, %; set 1 / set 2).

Condition	Acronyms, regret	Acronyms, agreement	Greater-than, regret	Greater-than, agreement
Mean	0.0700 [0.0514, 0.0896] / 0.0920 [0.0700, 0.1159]	82.6 / 80.5	0.0176 [0.0141, 0.0215] / 0.0224 [0.0176, 0.0278]	44.5 / 42.8
Frozen	0.0408 [0.0281, 0.0548] / 0.0477 [0.0329, 0.0644]	87.7 / 86.1	0.0091 [0.0068, 0.0118] / 0.0093 [0.0069, 0.0121]	64.5 / 64.5
+8 by AtP	0.0160 / 0.0178	91.8 / 91.8	0.0030 / 0.0025	77.3 / 78.3
+16 by AtP	0.0083 / 0.0081	93.8 / 94.3	0.0008 / 0.0010	84.0 / 83.4
+24 by AtP	0.0048 / 0.0071	94.9 / 94.5	0.0003 / 0.0003	89.5 / 89.8
+48 by AtP	0.0035 [0.0013, 0.0061] / 0.0036 [0.0015, 0.0061]	96.5 / 96.7	0.00001 [0.00000, 0.00002] / 0.00001 [0.00000, 0.00003]	96.9 / 97.9
+24 by attention change	0.0375 / 0.0487	89.1 / 87.7	0.0040 / 0.0052	70.9 / 72.1
+48 by attention change	0.0355 / 0.0421	88.9 / 89.1	0.0004 / 0.0003	90.6 / 90.0
+24 random downstream	0.0443 / 0.0514	86.7 / 86.3	0.0011 / 0.0013	80.1 / 79.3
+48 random downstream (two draws)	0.0327, 0.0370 / 0.0374, 0.0362	88.3, 88.1 / 89.6, 88.3	0.0019, 0.0007 / 0.0033, 0.0006	79.3, 85.2 / 79.1, 85.4
Layers 10–11 unfrozen	0.0376 / 0.0466	89.8 / 88.1	0.0099 / 0.0097	63.1 / 65.2

Verdicts and the tolerance. In greater-than, every AtP-ranked set, the 48 heads ranked by attention change, and all three random downstream sets lie within tolerance on both sets; the frozen model, the layer 10–11 heads, and the 24 heads ranked by attention change do not. For acronyms no set of unfrozen heads is within tolerance; the 48 AtP heads come closest, with interval upper ends of 0.0061 nats on both sets, so they would pass at a tolerance of 0.0062 nats. Mean ablation fails greater-than for any tolerance up to 0.014 nats. On the calibration inputs, the rule of Appendix A accepts the 48 AtP heads on the first acronym set but not the second, and accepts the frozen model on the second greater-than set although its evaluation interval lies above the tolerance: with 32 inputs the rule is sensitive to heavy-tailed regret.

Clean behavior. Mean ablation lowers the acronym clean loss from 0.986 to 0.683 nats (set 1) and from 0.932 to 0.681 (set 2), and changes clean accuracy from 91.4% to 94.5% and from 94.9% to 96.1%. For greater-than the clean loss rises from 0.092 to 0.114 and from 0.089 to 0.107, and clean accuracy changes from 99.2% to 99.6% and from 99.2% to 100.0%.

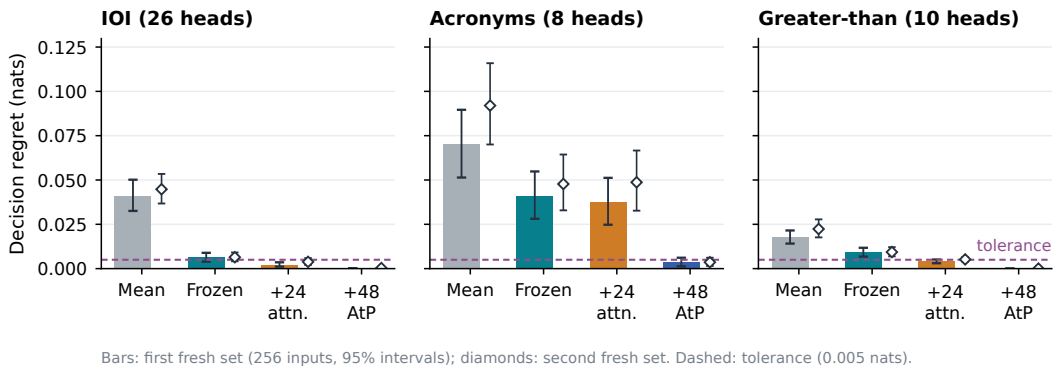


Figure 5: Mean ablation fails the audit in all three published GPT-2 circuits; freezing leaves regret above the tolerance in the two new circuits; unfreezing the 48 heads with the highest AtP scores passes in IOI and greater-than but not for acronyms, and the 24 heads ranked by attention change pass only in IOI's first set. Bars: first fresh set of 256 inputs (95% intervals); diamonds: second set; dashed: the 0.005-nat tolerance.

Disagreement structure. For acronyms, 88 of 89 and 100 of 100 changed selections of the mean ablation miss more than 0.005 nats, with median cost 0.33 and 0.36 nats; for the frozen model, 62 of 63 and 70 of 71, median 0.23 and 0.25. For greater-than, near-ties are common: the full model's two best pairs lie within 0.005 nats in 200 and 201 of 512 selections. Still, 211 of 284 and 214 of 293 changed selections of the mean ablation cost more than the tolerance (median 0.010 and 0.012), as do 107 of 182 and 105 of 182 of the frozen model's. In greater-than, the pair that is most damaging on average under zero knockouts differs between the full model and mean ablation on both sets (L6H9+L7H10 and L7H10+L9H1); under resampling it is the same. Choosing a pair per input gains 0.106 and 0.105 nats over always using the pair that is best on average on the calibration inputs for acronyms and 0.019 and 0.020 for greater-than; mean ablation forfeits 66% and 88%, and 91% and 111%, of these gains.

M.3 Completeness readings

Table 37: Completeness-style readings: the knockout effect of a group of the circuit's heads under each condition, as a ratio to the full model (95% intervals; set 1 / set 2; zero knockout).

Knockout	Full-model effect (nats)	Mean	Frozen	+48 by AtP
Acronyms: four letter movers	2.78 / 2.82	1.50 [1.44, 1.55] / 1.47 [1.42, 1.52]	1.28 [1.24, 1.31] / 1.30 [1.27, 1.35]	1.01 [1.01, 1.02] / 1.01 [1.01, 1.02]
Greater-than: seven candidate heads	0.38 / 0.34	0.75 [0.69, 0.81] / 0.76 [0.68, 0.84]	0.84 [0.77, 0.91] / 0.88 [0.81, 0.97]	0.98 [0.97, 1.00] / 0.98 [0.96, 0.99]

Under resample knockouts the ratios are 1.29 / 1.26 (mean ablation) and 1.07 / 1.10 (frozen) for acronyms, and 1.04 / 1.00 and 1.09 / 1.10 for greater-than.

M.4 Downstream self-repair and checks

Downstream self-repair and the first-token channel. In the full model we zeroed the four acronym letter movers, or the seven greater-than candidate heads, and repeated the channel measurement of Section 4.3 on 128 new inputs per circuit (97.5% intervals). Downstream reactions amplify both

knockouts: freezing every downstream component lowers the loss increase by 1.28 [1.06, 1.49] nats for acronyms and 0.40 [0.28, 0.52] for greater-than. For acronyms, freezing the downstream MLPs alone lowers it by 1.71 nats, whereas freezing the downstream attention outputs alone raises it by 0.85 [0.69, 1.01], and freezing only their first-token attention raises it by 0.40 [0.32, 0.48]. First-token attention thus carries 47% [41%, 53%] of the attention heads' offset (95%); the position-1 control changes the loss by -0.003. Several of the heads whose first-token attention drops most, L10H1, L10H6, L11H2, and L11H9, are backup name movers of the IOI circuit. For greater-than, the attention reaction is indistinguishable from zero, -0.001 [-0.068, 0.070]; this interval includes the differences between the frozen and full models in the completeness readings above (-0.063 and -0.040 nats on 256 inputs each), whose own intervals exclude zero. Its first-token channel is small, and its sign depends on the knockout type: -0.055 [-0.099, -0.010] under zero ablation and 0.064 [0.039, 0.090] under resampling.

Checks and cost. All numerical identity checks of Appendix G hold at float precision (maximum 3.6×10^{-7}), the AtP central-difference check agrees within 0.63%, and the recomputation from the earliest knocked-out layer matches full forward passes exactly. The runs took 8,113 s and 6,149 s on CPU (two threads each, concurrently).

N Robustness of the audit and transfer to other knockouts

This appendix tests three choices behind Sections 3.1–3.4: how many inputs the means are fitted on, which replacement the original studies used, and which random heads serve as a comparison. It then asks whether a pass carries over to knockouts the audit did not use (Section 6).

N.1 Design

Inputs. All comparisons use the evaluation sets of Tables 2 and 4: the two fresh sets of 256 IOI inputs, and the two sets of 256 acronym and greater-than inputs of Appendix M. The full model, the 32-input mean ablation, freezing, and the reported groups of unfrozen heads are computed again on these inputs (N.2), so every new ablation is compared with them on the same inputs. New fitting inputs share no token sequence with any evaluation or earlier fitting input.

Number of fitting inputs. Means are fitted per layer, position, and head, as in Section 2.3. All three input generators produce prompts of a fixed length whose positions hold the same roles (for IOI, one grammatical family with single-token names, places, and objects), so per-position means are aligned with the template. We fit IOI means on 128, 512, and 2,048 clean inputs and on nine further independent sets of 32; acronym means on 128 and 800 inputs, the size of the dataset over which García-Carrasco et al. (2024) average, and on four further sets of 32; greater-than means on 128 and 512 inputs and on four further sets of 32.

The original studies' ablations. Wang et al. (2023) mean-ablate over the ABC distribution, in which the three names are unrelated; our IOI inputs become ABC inputs by replacing the repeated subject with a third name, which changes one position and keeps the length. We fit ABC means on 32 and 512 inputs. We also replace every non-circuit head's output with its output on the same input's ABC counterpart (counterfactual patching). Hanna et al. (2023) patch from the 01 dataset, which replaces the last two digits of the start year with 01; we replace every non-circuit head's output with its output on the same input's 01 counterpart. Their procedure patches edges and includes MLPs in the circuit, so ours is a node-level approximation.

Random downstream sets. We draw 50 (IOI) or 20 (acronyms and greater-than) random sets each of 24 and 48 non-circuit heads from the heads that lie downstream of the first candidate head (74, 113, and 67 heads) and unfreeze them in turn, with all other non-circuit heads frozen. For IOI we also unfreeze the top 32, 40, 56, and 64 heads by the AtP scores of Section 3.3.

Knockouts the audit did not use. In IOI and greater-than, where the 48 AtP heads pass the pair audit on both sets, we evaluate the same ablations on four further families of knockouts, each under zero and resample ablation: single candidate heads (8 and 7), triples of candidate heads (56 and 35), single circuit heads outside the candidate set (18 and 3), and pairs of them (28 of the 153 IOI pairs, drawn at random, and all 3 greater-than pairs). Within each family, the decision is the most damaging member per input and ablation type, scored by regret as in Section 2.2. The AtP scores that choose

the unfrozen heads come from zeroing single candidate heads, so the first family overlaps with that choice; the other three do not. The circuit heads outside the candidate set include heads in layers 0–2 (IOI: L0H1, L0H10, L2H2) and layer 0 (greater-than: all three), upstream of every unfrozen head, whose knockouts can change non-circuit heads that stay frozen in every ablation of this appendix. We also zero and resample whole head classes (the seven IOI classes of Wang et al. (2023); for greater-than, the candidate heads, the other three heads, and all ten) and read their knockout effects as in Section 3.5.

Estimation. Everything else follows Appendices A and L: regret, agreement, and knockout-effect MAE per input, averaged over the two ablation types; 95% input-bootstrap intervals (8,000 resamples); a verdict of within tolerance, above it, or undecided at 0.005 nats; paired contrasts at 97.5%. As a secondary endpoint we repeat the decisions with each study's own metric (IOI logit difference between the indirect object and the subject, the letter logit difference for acronyms, the probability difference for greater-than).

Computation. Every knocked-out branch recomputes the blocks from the earliest knocked-out layer on the ablated model's own clean residual stream, as in Appendix M. Most runs use the Apple M2 GPU through PyTorch's MPS backend; six IOI random-group runs (draws 0–29 on each set) use the CPU. Each ablation is scored against the full model computed in the same run; MPS and CPU log-probabilities for the same input differ by at most about 10^{-4}.

N.2 Reproduction

On the evaluation sets of Tables 2 and 4, the full model, mean ablation, freezing, and the reported unfrozen groups computed again here differ from the per-input losses behind Tables 2 and 4 by at most 6.2×10^{-5} nats (57,600–75,520 knocked-out losses per set), reproduce every reported regret to within 4×10^{-7} nats, and give the same verdicts.

N.3 Fitted means and the original studies' replacements

Table 38: Regret of the complement ablations by fitted means and by the original studies' replacements (nats, 95% intervals; set 1 / set 2). "Within": yes if the interval lies below 0.005 nats, no if above, "?" if it crosses. Clean accuracy in %; the last column is the zeroed group's knockout effect as a ratio to the full model (Section 3.5).

Circuit	Complement ablation	Regret	Within	Clean accuracy	Group knockout
IOI	Mean, 32 inputs (Sections 3.1–3.4)	0.0410 [0.0326, 0.0502] / 0.0448 [0.0364, 0.0533]	no / no	93.0 / 87.1	2.07 / 2.02
	9 further 32-input fits	0.0403–0.0462 / 0.0439–0.0525	no / no	—	—
	Mean, 128 inputs	0.0440 [0.0353, 0.0536] / 0.0455 [0.0368, 0.0543]	no / no	94.9 / 93.0	2.03 / 1.96
	Mean, 512 inputs	0.0454 [0.0366, 0.0547] / 0.0490 [0.0395, 0.0589]	no / no	94.9 / 92.6	2.03 / 1.97
	Mean, 2,048 inputs	0.0444 [0.0356, 0.0538] / 0.0488 [0.0393, 0.0586]	no / no	93.8 / 91.0	2.03 / 1.97
	ABC mean, 32 inputs (IOI study)	0.0344 [0.0267, 0.0431] / 0.0360 [0.0285, 0.0442]	no / no	84.4 / 82.4	1.94 / 1.87
	ABC mean, 512 inputs (IOI study)	0.0437 [0.0349, 0.0532] / 0.0417 [0.0335, 0.0503]	no / no	88.7 / 84.4	1.98 / 1.91
	Each input's ABC counterpart	0.0183 [0.0135, 0.0237] / 0.0182 [0.0131, 0.0237]	no / no	81.6 / 85.2	1.44 / 1.40
	Freezing (reference)	0.0063 [0.0040, 0.0089] / 0.0064 [0.0041, 0.0092]	? / ?	95.7 / 91.8	1.35 / 1.33
Acronyms	Mean, 32 inputs (Sections 3.1–3.4)	0.0700 [0.0514, 0.0900] / 0.0920 [0.0700, 0.1162]	no / no	94.5 / 96.1	1.50 / 1.47
	4 further 32-input fits	0.0700–0.0783 / 0.0898–0.0968	no / no	—	—

Circuit	Complement ablation	Regret	Within	Clean accuracy	Group knockout
Greater-than	Mean, 128 inputs	0.0751 [0.0556, 0.0961] / 0.0928 [0.0707, 0.1168]	no / no	95.3 / 97.7	1.50 / 1.48
	Mean, 800 inputs (acronym authors' sample size)	0.0674 [0.0502, 0.0858] / 0.0913 [0.0692, 0.1153]	no / no	95.3 / 98.4	1.50 / 1.48
	Freezing (reference)	0.0408 [0.0281, 0.0552] / 0.0477 [0.0329, 0.0645]	no / no	91.4 / 94.9	1.28 / 1.30
	Mean, 32 inputs (Sections 3.1–3.4)	0.0176 [0.0142, 0.0215] / 0.0224 [0.0176, 0.0278]	no / no	99.6 / 100.0	0.75 / 0.76
	4 further 32-input fits	0.0141–0.0176 / 0.0191–0.0224	no / no	—	—
	Mean, 128 inputs	0.0173 [0.0138, 0.0212] / 0.0207 [0.0163, 0.0258]	no / no	99.2 / 99.6	0.73 / 0.73
	Mean, 512 inputs	0.0157 [0.0126, 0.0191] / 0.0205 [0.0161, 0.0256]	no / no	99.2 / 99.6	0.75 / 0.76
	Each input's 01 counterpart (greater-than study, node-level)	0.0174 [0.0129, 0.0228] / 0.0229 [0.0169, 0.0300]	no / no	98.4 / 99.2	0.65 / 0.65
	Freezing (reference)	0.0091 [0.0068, 0.0118] / 0.0093 [0.0069, 0.0122]	no / no	99.2 / 99.2	0.84 / 0.88

Full-model clean accuracy on set 1: IOI, acronyms, greater-than = 95.7 / 91.4 / 99.2%.

Every fit and every original study's replacement fails the audit on both sets of each circuit, and the zeroed group's knockout ratio barely moves with the number of fitting inputs. By ablation type, the IOI regret of the 32-input mean ablation is 0.047 (zero) and 0.035 (resample) nats on set 1 and 0.049 and 0.041 on set 2. The IOI study's ABC means also cost 7–11 points of clean accuracy on these sets, against 2.7 and 4.7 for our 32-input means. Replacing each non-circuit head's output with its output on the same input's ABC counterpart cuts the IOI regret by more than half (0.018 against 0.041 and 0.045) but stays above the tolerance; in greater-than, the 01 counterparts distort the group knockout further than the means (0.65 against 0.75 and 0.76).

Each study's own metric. Scored with each study's own metric (N.1), the 32-input mean ablation changes 35–38% of the IOI choices (freezing 21–23%, the 48 AtP heads 4%), 21–24% of the acronym choices, and 56–57% of the greater-than choices; regret in these units is not compared with the NLL tolerance.

N.4 More unfrozen heads and random downstream groups

Table 39: IOI regret by the number k of unfrozen heads ranked by AtP (nats; set 1 / set 2; Table 3 gives $k = 8, 16, 24$, and 48 on set 1).

k	32	40	48	56	64
Regret	0.0003 / 0.0008	0.0002 / 0.0005	0.0001 / 0.0003	<0.0001 / <0.0001	<0.0001 / <0.0001

Table 40: Random groups of non-circuit heads drawn from the heads downstream of the knockouts, unfrozen in turn (regret in nats; contrasts are paired over inputs with 97.5% intervals).

Circuit	Heads	Set 1	Set 2
IOI	48	50 groups: median 0.0047 (range 0.0003–0.0699); within 14, ? 22, above 14; 48 by AtP 0.0001; random mean minus it 0.0094 [0.0077, 0.0112]	50 groups: median 0.0045 (range 0.0007–0.0538); within 14, ? 23, above 13; 48 by AtP 0.0003; random mean minus it 0.0078 [0.0063, 0.0094]
IOI	24	50 groups: median 0.0055 (range 0.0032–0.0288); within 1, ? 38, above 11; 24 by attention change 0.0022; random mean minus it 0.0053 [0.0035, 0.0074]	50 groups: median 0.0054 (range 0.0035–0.0231); within 0, ? 39, above 11; 24 by attention change 0.0038; random mean minus it 0.0032 [0.0015, 0.0048]
Acronyms	48	20 groups: median 0.0337 (range 0.0151–0.0386); within 0, ? 0, above 20; 48 by AtP 0.0035; random mean minus it 0.0285 [0.0190, 0.0393]	20 groups: median 0.0394 (range 0.0205–0.0486); within 0, ? 0, above 20; 48 by AtP 0.0036; random mean minus it 0.0336 [0.0218, 0.0468]
Acronyms	24	20 groups: median 0.0381 (range 0.0263–0.0446); within 0, ? 0, above 20; 24 by attention change 0.0375; random mean minus it 0.0003 [−0.0067, 0.0068]	20 groups: median 0.0442 (range 0.0332–0.0549); within 0, ? 0, above 20; 24 by attention change 0.0487; random mean minus it −0.0045 [−0.0126, 0.0030]
Greater-than	48	20 groups: median 0.0014 (range 0.0005–0.0048); within 18, ? 2, above 0; 48 by AtP <0.0001; random mean minus it 0.0015 [0.0012, 0.0018]	20 groups: median 0.0012 (range 0.0007–0.0047); within 18, ? 2, above 0; 48 by AtP <0.0001; random mean minus it 0.0016 [0.0013, 0.0019]
Greater-than	24	20 groups: median 0.0043 (range 0.0017–0.0129); within 5, ? 10, above 5; 24 by attention change 0.0040; random mean minus it 0.0015 [0.0002, 0.0030]	20 groups: median 0.0051 (range 0.0014–0.0159); within 9, ? 6, above 5; 24 by attention change 0.0052; random mean minus it 0.0007 [−0.0009, 0.0023]

From 32 heads on, IOI regret stays within tolerance on both sets and does not rise as heads are added; the non-monotone values of Table 3 occur only below 32 heads. The 48 AtP heads beat every random group of 48 in all three circuits, but how many random groups pass depends on the circuit: 28 of 100 in IOI, 36 of 40 in greater-than, and none of 40 for acronyms. The 24 heads ranked by attention change beat the mean of the random groups of 24 on both IOI sets, are not distinguishable from it for acronyms, and beat it in greater-than on the first set only.

N.5 Knockouts the audit did not use

Table 41: IOI: regret on knockouts the audit did not use (nats; set 1 / set 2; zero and resample ablation averaged within input; "yes", "no", "?" as in Table 2). The first column repeats the pair audit.

Complement ablation	Pair audit	Candidate heads, single (8)	Candidate heads, triples (56)	Other circuit heads, single (18)	Other circuit heads, pairs (28)	Head classes (7)
Mean	no / no	0.0305 no / 0.0303 no	0.0474 no / 0.0471 no	0.0648 no / 0.0448 no	0.1632 no / 0.1639 no	0.1982 no / 0.1902 no
Freezing	? / ?	0.0022 yes / 0.0034 ?	0.0264 no / 0.0270 no	0.0435 no / 0.0179 no	0.1499 no / 0.1283 no	0.1145 no / 0.1424 no

Complement ablation	Pair audit	Candidate heads, single (8)	Candidate heads, triples (56)	Other circuit heads, single (18)	Other circuit heads, pairs (28)	Head classes (7)
+8 by AtP	yes / —	0.0010 yes / 0.0018 yes	0.0234 no / 0.0254 no	0.0225 no / 0.0197 no	0.3125 no / 0.2565 no	0.1184 no / 0.1361 no
+48 by AtP	yes / yes	<0.0001 yes / <0.0001 yes	0.0010 yes / 0.0015 yes	0.0436 no / 0.0299 no	0.1873 no / 0.1855 no	0.1151 no / 0.1373 no
+24 by attention change	yes / ?	0.0002 yes / 0.0003 yes	0.0196 no / 0.0192 no	0.0309 no / 0.0263 no	0.2464 no / 0.2122 no	0.1056 no / 0.1273 no
+48 random downstream (1)	? / ?	0.0004 yes / 0.0006 yes	0.0187 no / 0.0212 no	0.0500 no / 0.0518 no	0.6933 no / 0.6939 no	0.0959 no / 0.0854 no
+48 random downstream (2)	no / no	0.0012 yes / 0.0023 yes	0.0276 no / 0.0236 no	0.0325 no / 0.0192 no	0.3773 no / 0.3569 no	0.1000 no / 0.1276 no

Stakes on set 1 (largest full-model knockout effect, and the gain of choosing per input over the best member on average; nats): candidate heads, single: 0.96 / 0.045 · candidate heads, triples: 2.28 / 0.069 · other circuit heads, single: 0.73 / 0.101 · other circuit heads, pairs: 1.67 / 0.071 · head classes: 2.15 / 0.215.

Table 42: Greater-than: regret on knockouts the audit did not use (nats; set 1 / set 2; zero and resample ablation averaged within input; “yes”, “no”, “?” as in Table 2). The first column repeats the pair audit.

Complement ablation	Pair audit	Candidate heads, single (7)	Candidate heads, triples (35)	Other circuit heads, single (3)	Other circuit heads, pairs (3)	Head classes (3)
Mean	no / no	0.0064 ? / 0.0080 no	0.0182 no / 0.0186 no	0.0053 ? / 0.0050 ?	0.0050 ? / 0.0040 yes	0.0266 no / 0.0246 no
Freezing	no / no	0.0034 ? / 0.0026 yes	0.0184 no / 0.0210 no	0.0036 ? / 0.0028 yes	0.0019 yes / 0.0017 yes	0.0121 no / 0.0161 no
+8 by AtP	yes / yes	0.0008 yes / 0.0008 yes	0.0068 no / 0.0066 ?	0.0034 yes / 0.0027 yes	0.0033 yes / 0.0027 yes	0.0135 no / 0.0196 no
+24 by AtP	yes / yes	0.0002 yes / 0.0002 yes	0.0018 yes / 0.0017 yes	0.0046 ? / 0.0037 ?	0.0044 ? / 0.0035 yes	0.0185 no / 0.0193 no
+48 by AtP	yes / yes	<0.0001 yes / <0.0001 yes	<0.0001 yes / <0.0001 yes	0.0047 ? / 0.0039 ?	0.0044 ? / 0.0037 ?	0.0157 no / 0.0157 no
+48 random downstream	yes / yes	0.0006 yes / 0.0011 yes	0.0032 yes / 0.0057 ?	0.0047 ? / 0.0033 yes	0.0023 yes / 0.0023 yes	0.0205 no / 0.0218 no

Stakes on set 1 (largest full-model knockout effect, and the gain of choosing per input over the best member on average; nats): candidate heads, single: 0.08 / 0.009 · candidate heads, triples: 0.26 / 0.037 · other circuit heads, single: 0.02 / 0.006 · other circuit heads, pairs: 0.01 / 0.005 · head classes: 0.54 / 0.020.

With the 48 AtP heads unfrozen, the pass carries over to single candidate heads and their triples in both circuits, whereas the other groups that pass the IOI pair audit on the first set (8 AtP heads, 24 heads ranked by attention change) fail on triples. In IOI the pass does not carry over to the other circuit heads. In greater-than, the largest knockout effect among the three layer-0 heads averages at most 0.016 nats per input, too little to separate the ablations.

Splitting the IOI circuit heads outside the candidate set by whether they lie upstream of every unfrozen head (for the 48 AtP heads the first unfrozen layer is 4, which leaves L0H1, L0H10, L2H2 upstream): with the 48 AtP heads unfrozen, regret for the 15 downstream heads is 0.0006 yes / 0.0010 yes for single knockouts and 0.0012 yes / 0.0011 yes for pairs, and for the upstream heads 0.0126 no / 0.0044 ? for single knockouts and 0.1669 no / 0.1777 no for pairs containing one of them (set 1 / set 2). The knockout effects with the largest error are those of L0H10 (0.30 nats), L0H1 (0.06 nats), L2H2 (0.05 nats).

Table 43: IOI: zeroing each head class of the circuit, as a ratio to the full model's knockout effect (set 1 / set 2; the full-model effect in nats on set 1).

Class (full-model effect)	Mean	Freezing	+48 by AtP
Backup name movers (+0.60)	2.07 / 2.18	1.40 / 1.42	1.00 / 1.00
Duplicate-token heads (+1.18)	1.98 / 1.97	1.42 / 1.43	1.72 / 1.74
Induction heads (+0.64)	1.26 / 1.25	1.00 / 0.99	1.10 / 1.11
Name movers (+0.52)	2.07 / 2.02	1.35 / 1.33	1.03 / 1.02
Negative name movers (−0.78)	1.32 / 1.37	1.00 / 1.00	1.00 / 1.00
Previous-token heads (−0.10)	−5.87 / −7.00	−0.19 / −0.35	1.15 / 1.10
S-inhibition heads (+1.18)	1.17 / 1.15	1.10 / 1.10	1.00 / 1.00

Mean ablation distorts the knockout effect of every IOI head class (every 95% interval of the ratio excludes 1) and flips its sign for the previous-token heads. The 48 AtP heads bring four classes within 3% of the full model, but not the duplicate-token heads L0H1, L0H10, and L3H0, all upstream of every unfrozen head.

O Re-implementing a recent completeness test

This appendix supports Section 3.5. Gong et al. (2026), following the completeness test of Wang et al. (2023), test whether a circuit is complete by comparing its response to a knockout of the name movers with the full model's response, with every head outside the circuit mean-ablated (Section 4 and Table 2 of their third arXiv version); they released the code and its reference values. Their test differs from that of Wang et al. (2023) in two ways: it fits the means on the evaluation prompts themselves, pooling all positions, rather than on the ABC distribution per template position, and it defines circuits by heads rather than by heads at positions. We re-implemented this test and then changed how the non-circuit heads are ablated.

O.1 Procedure and reproduction

The original test. The inputs are GPT-2 small prompts from one template, "When [IO] and [S] went to the [place], [S] gave a [object] to", with the indirect object first and no end-of-text token; each of four seeds (1, 8, 15, 22) draws 96 prompts. The circuits are the documented 26-head IOI circuit, a primary circuit without its eight backup name movers (18 heads), and four completions that add eight further heads to the primary circuit: the top eight by three scores of conditional co-ablation (CoAx, the amplification ratio, and conditional energy) or eight random heads, drawn once per seed. Each head outside a circuit is replaced by its output averaged over all positions of the evaluation prompts. The knockout replaces the name movers L9H9, L9H6, and L10H0 by the same means, in the full model and in every circuit, and a model's response is the drop in the logit difference between the indirect object and the subject at the last position. A circuit's incompleteness is the absolute difference between its mean response and the full model's for a seed, averaged over the four seeds.

Reproduction. We re-implemented this procedure from the released code (commit 332a699c). Our implementation reproduces the six incompleteness values of the released reference file, which match

Table 2 of Gong et al. (2026), and the CoAx backup ROC-AUC of 0.941, to within 4×10^{-6} (first table below). The full model's margin falls from 2.57 to 2.29 under the knockout, as reported there.

What we changed. In the main comparison only the treatment of the non-circuit heads changes: freezing replaces each by its clean output on the same prompt at every position, which keeps every circuit's clean behavior exact. Under freezing the incompleteness no longer contains a term from substituted values; it measures only the reactions of non-circuit heads that the circuit lacks (Section 2.3). Section O.3 also refits the means per position, as ours are fitted, on 32 held-out prompts from the same template, and varies the knockout and the metric. Intervals are bootstraps over prompts within seeds (10,000 draws) with the means held fixed; estimates use 95% and the four comparisons of the paper's statements 97.5%, each without a correction for the others. Ratios to the full model are computed per seed and averaged.

Table 44: The completeness test re-implemented (GPT-2 small; four seeds of 96 prompts). Incompleteness and knockout effects are in units of the logit difference between the indirect object and the subject; 95% intervals over prompts. "Clean" is the circuit's clean logit difference as a fraction of the full model's under the original means (exactly 1 under freezing). The name movers' knockout effect in the full model is 0.288.

Circuit (heads)	Incompleteness, reported / re-implemented	Clean, original means	Knockout effect, original means / freezing	Incompleteness, freezing
Primary (18)	0.474 / 0.474	0.023	−0.187 / 1.049	0.762 [0.682, 0.844]
+ CoAx	0.236 / 0.236	−0.255	0.052 / 0.232	0.056 [0.031, 0.085]
+ amplification ratio	0.258 / 0.258	−0.192	0.030 / 0.322	0.048 [0.027, 0.086]
+ conditional energy	0.181 / 0.181	−0.073	0.107 / 0.608	0.320 [0.244, 0.398]
+ random	0.499 / 0.499	−0.003	−0.211 / 1.019	0.732 [0.652, 0.816]
Documented (26)	0.300 / 0.300	−0.149	−0.012 / 0.368	0.081 [0.051, 0.116]

O.2 What the incompleteness values measure

Under the original means no circuit performs the task: the circuits' clean logit differences range from −26% to 2.3% of the full model's, and they favor the indirect object on 25–52% of prompts, against 99.7% for the full model. A clean check would reveal this failure. The knockout then barely moves the circuits' margins, or moves them the wrong way (effects from −0.211 to 0.107, against 0.288 in the full model), so the incompleteness values are dominated by, and take their sign from, the substituted means. Split as in Section 2.3, the documented circuit's error has a substitution term of −0.38 and a blocked term of +0.08, and the primary circuit's −1.24 and +0.76 (point estimates); in every circuit the substitution term is the larger. Under freezing only the blocked term remains. The documented circuit's incompleteness is then 0.08 and the CoAx completion's 0.06, 28% and 19% of the full model's knockout effect, and the documented circuit's knockout reads 1.30 times the full model's. This residual, not the failure of the clean check, is what the audit adds. The primary circuit's knockout reads 3.68 times the full model's under freezing, the direction in which missing backup heads should push it, and −0.69 times under the original means.

Four comparisons (97.5% intervals) test the paper's statements. That the CoAx completion lowers the primary circuit's incompleteness holds, more strongly under freezing (−0.70 [−0.81, −0.61], against −0.24 [−0.27, −0.21] under the original means), and so does that it matches the documented circuit (−0.02 [−0.09, 0.04]). A random top-up does not lower the incompleteness under the original means (0.499 against 0.474); under freezing eight random heads lower it by 4% (−0.030 [−0.048, −0.012]), as heads that can react are expected to, and far less than CoAx (0.73 against 0.06). The primary circuit's error under freezing has the opposite sign to its error under the original means (their ratio is −1.61 [−2.05, −1.28]), so its reported incompleteness does not measure missing backup reactions. The test's main conclusion, that the recovered backup heads complete the circuit, holds for the CoAx and amplification-ratio completions (93% of the primary circuit's incompleteness removed under freezing by CoAx, against 50% under the original means). It does not hold for the conditional-energy

completion, which has the lowest value under the original means (0.18) but under freezing the highest of the three score-based completions (0.32, against 0.05 and 0.06); the test's authors already flag its knockout as a poor match to the documented backup intervention.

O.3 Per-position means and other readings

Table 45: The name movers' knockout effect as a ratio to the full model's, by how the non-circuit heads are treated (same prompts as the previous table; ratios averaged over seeds). Clean accuracy is the share of prompts whose clean logit difference favors the indirect object (full model 99.7%). Mean knockout and logit difference are the original test's; zero knockout and answer loss are those of Section 3.5.

Circuit	Non-circuit heads	Clean accuracy (%)	Mean knockout, logit difference	Zero knockout, answer loss
Documented	Original means	34.4	−0.05	0.36
	Per-position means	97.4	2.07	1.91
	Freezing	99.7	1.30	1.07
	Freezing, 48 AtP heads unfrozen	99.7	1.02	1.02
Primary	Original means	51.6	−0.69	1.01
	Per-position means	94.3	4.77	5.07
	Freezing	99.7	3.68	3.19
+ CoAx	Original means	25.0	0.18	0.24
	Per-position means	94.0	1.64	1.57
	Freezing	99.7	0.81	0.81

Per-position means keep the task (92–98% of prompts), and the documented circuit's name-mover knockout then reads 2.07 times the full model's in the logit difference and 1.91 times in answer loss, close to the two times of Section 3.5 on our inputs; under freezing it reads 1.30 and 1.07 times on these prompts. In the terms of Section 2.3, the blocked term follows the non-circuit heads' reactions, which here offset the knockout, so freezing makes it look more harmful. The substitution term follows the means: per-position means add to the blocked term, whereas means pooled over positions reverse it and make the knockout look almost harmless (−0.05). How the means are fitted thus sets the direction of the distortion; its size also depends on the knockout and the metric. One reading can look complete by chance: under the original means the primary circuit's answer-loss ratio for a zero knockout is 1.01, while its logit-difference ratio for the same knockout is −0.73. On the evaluation sets of Table 2, the name-mover zero-knockout reading in the logit difference is 2.26 and 1.75 under our means and 1.83 and 1.40 under means fitted on 512 ABC inputs as Wang et al. (2023) did (1.69 and 1.31 with 32), so the inflation of Section 3.5 does not come from the answer-loss metric.

The other head classes behave alike. Under the original means the documented circuit misreads the knockouts of the negative name movers, S-inhibition heads, induction heads, and previous-token heads (ratios −0.21 to 0.17, for full-model effects of 0.5 to 1.7 in absolute value); under freezing the ratios are 0.61 to 0.93. As in Section 3.1, the dataset-level number hides per-input error: under freezing the documented circuit's mean absolute per-prompt error is 0.28, 3.4 times its incompleteness. Unfreezing the 48 AtP heads of Section 3.3, chosen on our inputs, brings the documented circuit's name-mover reading to 1.02 on this template and those of the four other classes to 0.99–1.09; the exception is the duplicate-token class, which includes L0H1 and L0H10 (Appendix N).

O.4 Implementation notes

We did not run the released code; reproducing all six reference values and the ROC-AUC checks our implementation, which applies the unembedding only at the last position and runs on a CPU, and freezing every head reproduces each clean output exactly. The first arXiv version of the paper reports other values (0.72, 0.15, 0.16, and 0.61), which the released reference file does not contain, so we test the third version. Besides the per-position fitting of Section 2.3, the only choice carried over from our experiments is the set of 48 AtP heads.

P THE ORIGINAL COMPLETENESS TEST WITH THE COMPLEMENT FROZEN

This appendix supports Section 3.5. It re-runs the completeness test of Wang et al. (2023) with its original procedure and changes one thing, what the nodes outside the circuit are replaced with. Section 3.5's name-mover measurement uses our circuit, means, and answer loss; here every other choice is the original study's.

P.1 PROCEDURE

The original test. We follow the released code of Wang et al. (2023) (`completeness.py`, `ioi_circuit_extraction.py`, and `ioi_dataset.py`). Inputs come from its IOI dataset class: 14 templates, no end-of-text token; the search uses 100 inputs, as in the original study. A node is a head at one position: name movers, negative name movers, and S-inhibition heads at the final token; induction and duplicate-token heads at the repeated subject; previous-token heads at the token after the first subject. The published circuit has 26 heads, among them the eight backup name movers; the naive circuit of the same study has 13 and lacks the backups, the negative name movers, and three other heads. Each replaced node takes the mean of the head's output over the ABC version of the same inputs, per template and position. F is the logit difference between the indirect object and the subject at the final token. For a node set K, $F(C \setminus K)$ keeps the circuit's other nodes and replaces every other node, and $F(M \setminus K)$ replaces only K's nodes. The incompleteness score is $|F(C \setminus K) - F(M \setminus K)|$, for K empty, each head class, 100 random subsets (each node with probability one half), and the final set of each of 10 greedy searches (10 steps, 10 sampled candidates per step, 5 for the naive circuit), which add the node that most increases the score.

What we changed. Freezing replaces every node outside the circuit by its clean activation on the same input and still knocks K out with the ABC means, so $F(C) = F(M)$ exactly and the two differ only through the reactions of the replaced nodes. The signed gap $F(C \setminus K) - F(M \setminus K)$ then splits, as in Section 2.3, into a replacement term (original minus frozen) and a blocked term (frozen). The greedy search was run under each complement. The search uses one set of 100 inputs; two fresh sets of 250 inputs re-evaluate the sets it found, each with its own ABC means. Intervals are 95% bootstraps over inputs (2,000 resamples) with the means held fixed.

Reproduction. The circuit keeps 83% of the full model's logit difference (3.39) on the search set and 91% and 89% on the fresh sets, against 87% reported by Wang et al. (2023). The original inputs were drawn once without a recorded seed, so the searches cannot reproduce its reported values input for input. The model code matches the reference implementation to within 1.1e-4 in log-probability, and a literal re-implementation of the original hook, which overwrites each head's output position by position, matches our node-level replacement to within 3.1e-5 in logit difference. These runs used GPT-2 small weights assembled from two public exports of the checkpoint of Appendix B (the ONNX model zoo and KerasNLP), which agree bit for bit; on the first evaluation set of Table 2 they reproduce the full model's accuracy exactly and its mean answer loss to within 3e-5 nats.

P.2 HEAD CLASSES AND RANDOM SETS

Table 46: The original test and the frozen complement on the search set of 100 inputs. The gap is $F(C \setminus K) - F(M \setminus K)$ in logit difference; the incompleteness score is its absolute value. The replacement term is the original gap minus the frozen one.

Circuit (heads)	K	$F(M \setminus K)$	Gap, original	Gap, frozen	Replacement term
Published (26)	$\varnothing$	3.39	−0.58	0.00	−0.58
	Name movers (with the 8 backups)	1.40	−0.75	−0.46	−0.29
	Negative name movers	5.88	−1.43	−0.06	−1.37
	S-inhibition	0.41	−0.18	0.15	−0.34
	Induction	1.66	−0.68	0.11	−0.79
	Duplicate token	1.64	−0.94	0.22	−1.16

Circuit (heads)	K	$F(M \setminus K)$	Gap, original	Gap, frozen	Replacement term
Naive (13)	Previous token	2.99	−1.58	−0.13	−1.44
	∅	3.39	−0.58	0.00	−0.58
	Name movers (3)	3.28	−2.95	−4.15	1.20
	S-inhibition	0.41	−1.37	−2.10	0.72
	Induction (2)	2.45	−1.98	−1.16	−0.82
	Duplicate token (2)	2.69	−2.48	0.04	−2.53
	Previous token	2.99	−1.99	−0.17	−1.82

On the fresh sets, freezing shrinks the absolute gap of every class of the published circuit except the S-inhibition heads, whose small gap grows and changes sign; the largest remaining gap, the name movers', is mostly blocked reactions (−0.40 of −0.56 on the first fresh set). Averaged over the six classes, the absolute gap falls from 0.73 to 0.18 and from 0.74 to 0.20 on the two fresh sets, and the least complete class changes from the previous-token heads to the name movers. For the naive circuit, freezing enlarges the name movers' gap (−2.95 to −4.15 on the search set): its missing backups cannot react under either complement, and the original means partly offset that blocked term. Over the 100 random sets, the mean incompleteness changes from 0.54 to 0.38 for the published circuit and from 1.52 to 1.74 for the naive one; for the published circuit the two scores are negatively correlated across sets (−0.48), and freezing lowers the score of only 52% of them. Under both complements the published circuit is more complete than the naive one, for random sets and for classes.

P.3 Which sets the greedy search finds

Table 47: The set with the highest search score in each search (10 runs on the search set), re-evaluated under both complements on the two fresh sets of 250 inputs (incompleteness score with 95% intervals). The search score is the set's score on the search set under the complement it was found with.

Circuit (heads)	Search	Search score	Fresh set	Original test	Frozen complement
Published (26)	Original	3.81	1	3.60 [3.35, 3.83]	0.14 [0.10, 0.18]
			2	3.50 [3.25, 3.75]	0.10 [0.05, 0.15]
	Frozen	1.32	1	0.26 [0.10, 0.42]	1.38 [1.31, 1.47]
			2	0.28 [0.11, 0.44]	1.32 [1.23, 1.41]
Naive (13)	Original	3.21	1	3.09 [2.90, 3.28]	0.98 [0.89, 1.08]
			2	3.09 [2.89, 3.28]	0.98 [0.89, 1.06]
	Frozen	4.15	1	2.73 [2.52, 2.93]	4.21 [3.98, 4.44]
			2	2.83 [2.62, 3.05]	4.16 [3.92, 4.39]

For the published circuit the two searches find different sets: the best original search removes L2H2, L3H0, L4H11, L5H8, L9H0, L10H7, L11H2, and L11H10, and the best frozen search L0H1, L0H10, L3H0, L5H5, L6H9, L8H10, L10H0, L10H7, L11H2, and L11H10. The final sets of all ten original searches score 2.29 to 3.61 in the original test and 0.00 to 0.33 with the complement frozen on the fresh sets. They are built around the negative name movers and the previous-token heads, whose removal leaves the full model's logit difference near or above its clean value while the mean-ablated circuit collapses. Nine of the ten frozen searches find sets that score 1.04 to 1.38 with the complement frozen and 0.00 to 0.41 in the original test; there the full model's reactions amplify the knockout. The largest score each search reaches, re-evaluated on the fresh sets, is 1.38 and 1.32 under freezing against 3.60 and 3.50 in the original test (ratios 0.38 and 0.38). For the naive circuit both searches find sets that stay incomplete under either complement: the original search removes L0H10, L2H2, L4H11, and L9H6, and the frozen search the three name movers, L9H6, L9H9, and L10H0.

Search budget. Every frozen search of the published circuit used all ten steps. A deterministic greedy path that evaluates every remaining node at each step until all 26 are removed peaks at 1.35

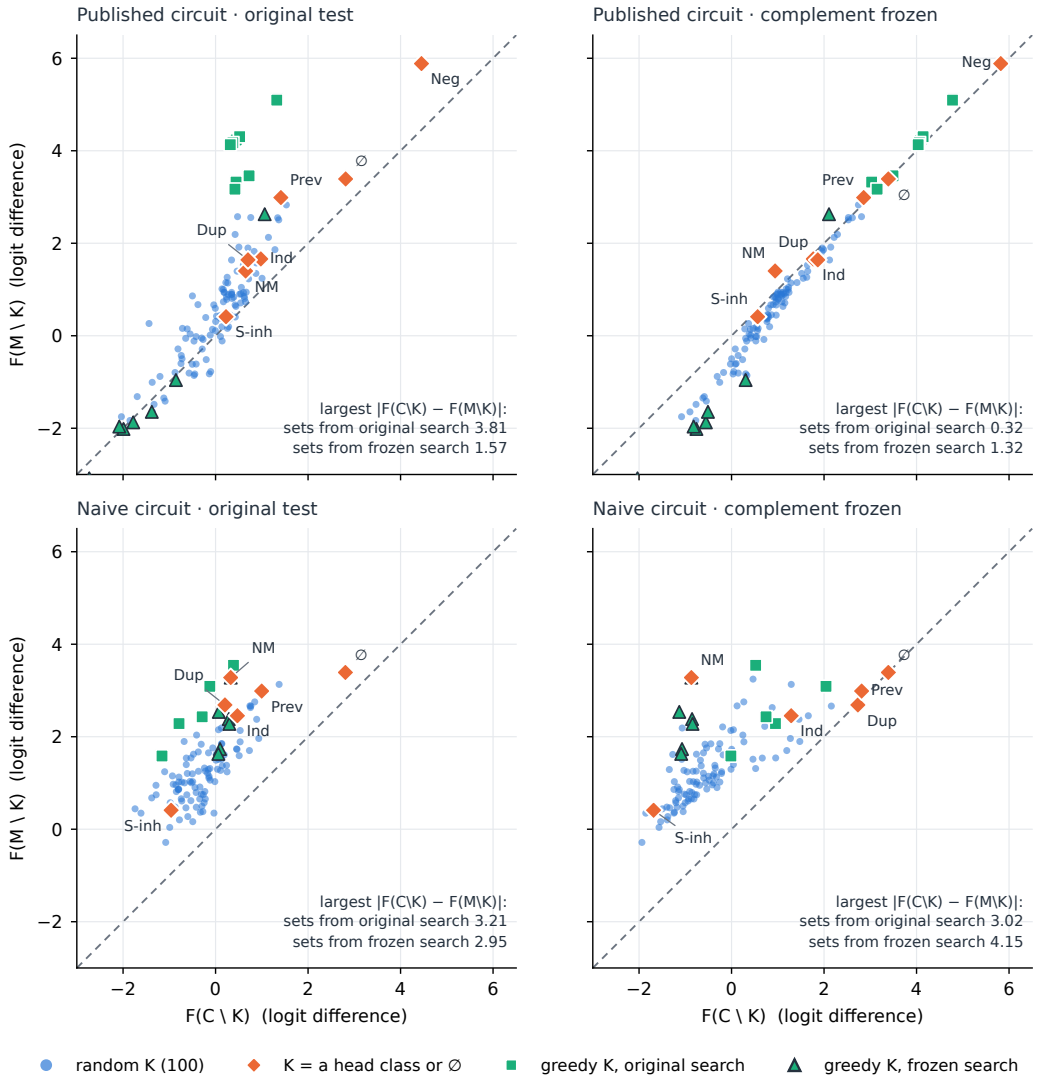


Figure 6: The original completeness test of Wang et al. (2023) with the complement mean-ablated (left, as in the original test) or frozen at clean activations (right), on the search set of 100 inputs; the knockout of K is the same. Points are $(F(C \setminus K), F(M \setminus K))$ in logit difference, and a complete circuit lies on the dashed diagonal. Squares are the final sets of the ten original greedy searches, triangles those of the ten frozen searches, diamonds the head classes, and dots 100 random sets; the same set has the same marker in both panels. Top: the published 26-head circuit; bottom: the naive circuit.

after 13 nodes under freezing and at 3.81 after 8 nodes in the original test; re-evaluated on the fresh sets, 1.41 and 1.36 against 3.60 and 3.50. The budget does not change the comparison.

Group knockout under the original test. The original test's own knockout of the name-mover class, backups included, is 1.13 and 1.14 times the full model's effect in logit difference with the complement mean-ablated and 1.20 and 1.24 with it frozen, and 1.47 and 1.48 against 1.16 and 1.16 in answer loss. Freezing therefore does not move every reading toward the full model's: here it does in answer loss but not in logit difference. The inflation of about two in Section 3.5 is a different measurement (our circuit, the three name movers zeroed, answer loss).