Circuit Condensation: Post-Training that Concentrates a Behavior's Causal Circuit

Anonymous authors
Paper under double-blind review

Abstract

One approach in mechanistic interpretability explains model behavior through circuits: the components and connections that carry it. Frozen discovery often returns hundreds of edges, making them hard to inspect, compare, or verify exhaustively. We introduce **Circuit Condensation**, which post-trains models to concentrate behaviors into smaller causal graphs. Each round prunes low-attribution edges and trains a low-rank adapter to match the original through what remains. It keeps the cut only if task accuracy and held-out perplexity remain within tolerance. Across four behaviors and eight models, condensed circuits are smaller than the strongest published frozen-weight baseline in 30 of 32 settings, by $8.1\times$ on average and up to $316\times$. Repeating the search without weight updates produces larger circuits in 29 of 32 settings, showing that weight updates drive the reduction rather than search alone. For 19 small circuits, exhaustive subset tests show that 11 are exactly minimal and identify removable edges in the rest. Pair ablations reveal dependencies that single-edge tests miss. On indirect object identification, condensation isolates 24 heads, including 17 with documented roles, whereas a matched frozen circuit needs 61 heads, 36 without documented roles. The result is a sufficient subcircuit of the published mechanism rather than a reconstruction of it. The resulting circuit tracks the original model's next-token distribution and, at matched edge count, predicts its errors better than a frozen circuit in 10 of 11 evaluable settings.

1 Introduction

Mechanistic interpretability seeks explanations in the internal computations that produce a language model's behavior and tests them through intervention rather than observation alone. One approach analyzes that computation as a *circuit*: a subgraph of components and connections sufficient to carry a particular behavior (Olah et al., 2020; Elhage et al., 2021; Olsson et al., 2022; Wang et al., 2023). The circuit view makes a precise claim that can be checked: a circuit is *faithful* if the model still performs the behavior when activations outside it are replaced.

Faithfulness is only a starting point. To understand a circuit, we would also like to know whether it contains unnecessary components and how its retained edges work together. These checks quickly become combinatorial, and circuits recovered from real models are often too large to test exhaustively. Proving that no smaller circuit suffices requires testing every subset, while testing whether retained edges depend on one another requires ablating every pair. For example, Wang et al. (2023) found exhaustive completeness testing infeasible for an indirect object identification (IOI) circuit containing 26 attention heads and used three subset-sampling strategies instead. A 13-edge circuit admits $8{,}192$ subsets and 78 pairs; a 400-edge circuit admits 2^{400} subsets. Size decides which questions can be asked at all, which is why it is a practical proxy for interpretability (Gao et al., 2025).

Most existing methods search the fixed model more carefully. Activation and path patching exchange internal states between clean and corrupted inputs (Goldowsky-Dill et al., 2023; Zhang & Nanda, 2024), Edge Attribution Patching and its integrated-gradient and saturation-corrected variants EAP-IG and EAP-GP scale that search with gradients (Syed et al., 2024; Hanna et al., 2024; Zhang et al., 2025), and learned gates and feature-level decompositions offer further routes to sparse structure (Bhaskar et al., 2024; Yu et al., 2025; Marks et al., 2025). These methods may improve the search, but they do not change how the model carries the behavior. Even within a fixed model, the recovered

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

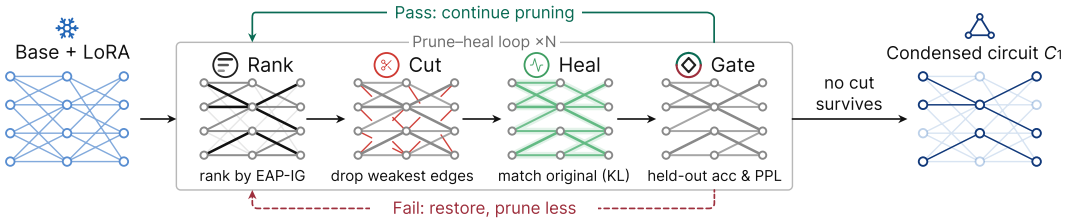


Figure 1: **One round of Circuit Condensation.** Rank surviving edges by causal importance, cut the weakest, then train only the adapter to reproduce the original model through what is left. A round is kept only if the target behavior and perplexity on held-out text remain within tolerance; otherwise the previous state is restored and a smaller cut tried. C_1 is the smallest accepted circuit.

circuit may shift with the estimator, data, and choice among equally faithful graphs (Méloux et al., 2026; Chen et al., 2026b).

We therefore ask a different question: rather than searching harder for where a behavior already lives, can training concentrate it into a smaller circuit? **Circuit Condensation** treats findability as a property a model can be trained to have. Starting from a fresh low-rank adapter (Hu et al., 2022) with every edge open, each round ranks the surviving connections by causal importance, prunes the weakest, and trains only the adapter to reproduce the original model's outputs through what remains. A controller on held-out data accepts a cut only if it preserves both the target behavior and held-out language-modeling performance, restoring the previous state otherwise (Figure 1). We call the smallest accepted circuit C_1.

A smaller circuit is useful only if it makes new analyses feasible and still describes the model we started from rather than only the adapted network. The reduced circuits make exhaustive subset and pairwise tests runnable (§5.3). They also track the original model's next-token distribution, at a median Kullback–Leibler (KL) divergence of 0.215 across the grid, and predict where that model will fail better than a frozen circuit of the same size (§5.4).

Contributions.

- We formulate circuit findability as a trainable model property rather than a fixed target for search. Our adaptive prune, heal, and backtrack procedure concentrates a named behavior under held-out faithfulness and capability constraints and returns a validated circuit in every task–model cell without per-cell tuning.
- Across four behaviors, eight models from four families, and three seeds, C_1 is smaller than frozen EAP-IG at comparable accuracy in 30 of 32 cells, by an average $8.1\times$ and up to $316\times$ (sign test $p = 2.5 \times 10^{-7}$, bootstrap interval $[4.9, 13.7]$). The same search with frozen weights loses in 29 of 32: the gain is reshaping, not search.
- The smaller circuits enable exhaustive subset, all-pairs, and edge-level necessity checks. Matched frozen circuits would require a median $55\times$ as many pair ablations, and subset enumeration is infeasible in all 19 matched cases. We also test whether the condensed circuit still describes the unmodified model and identify where condensation does not help.

2 Related Work

Circuit discovery. Circuits explain behavior through causal subgraphs of model components (Olah et al., 2020; Elhage et al., 2021); causal abstraction asks whether their representations respond to interventions like variables in a higher-level causal model (Geiger et al., 2021; 2022). Detailed studies show what these explanations can reveal, but manual analysis does not scale readily (Wang et al., 2023; Olsson et al., 2022; Heimersheim & Janiak, 2023). Activation and path patching intervene on internal states (Goldowsky-Dill et al., 2023; Zhang & Nanda, 2024); ACDC organizes repeated interventions into greedy pruning (Conmy et al., 2023), while causal scrubbing formalizes hypothesis testing (Chan et al., 2022). EAP, EAP-IG, and EAP-GP replace repeated interventions with gradient-based edge scores (Syed et al., 2024; Hanna et al., 2024; Zhang et al., 2025); EAP-IG generally per-

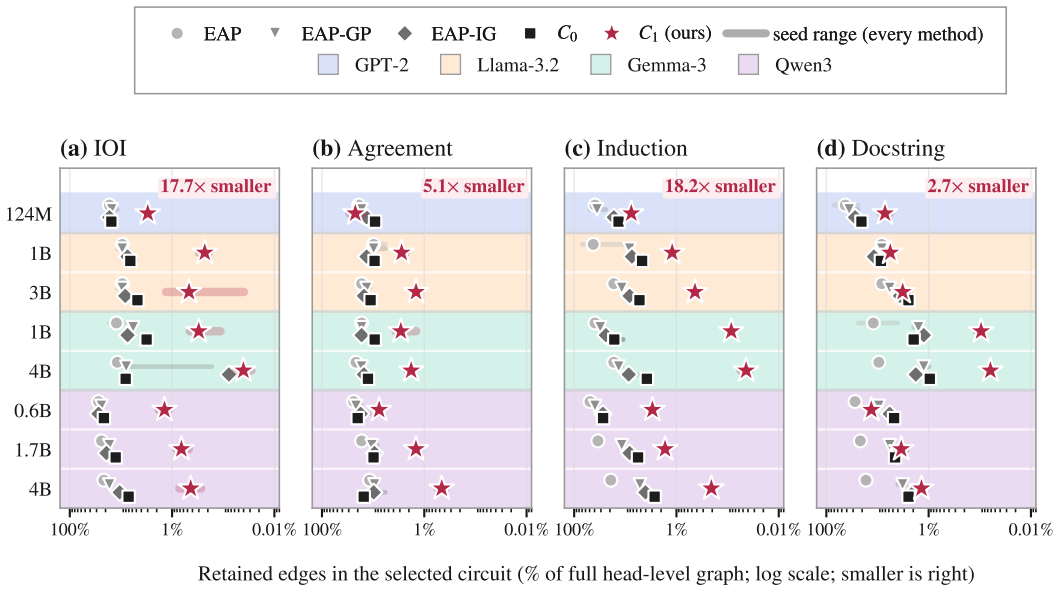


Figure 2: **Condensation usually retains fewer edges than frozen discovery at comparable accuracy.** Panels are behaviors and rows are models, shaded by family. The axis shows the share of the full head-level graph retained on a log scale, so *further right means smaller*. Stars show C_1 and other markers show frozen baselines (C_0 runs the same search with the weights fixed); markers are means over seeds 11/22/33 and bars span their range.

forms best on the MIB circuit-localization track (Mueller et al., 2025). Other methods search over information-flow routes (Ferrando & Voita, 2024), component decompositions (Hsu et al., 2025), position-aware edges (Haklay et al., 2025), or learned gates (Bhaskar et al., 2024; Yu et al., 2025). All keep the model weights fixed and ask how small a faithful circuit can be recovered, making them our main comparison class.

What circuit discovery returns. Even with fixed weights, the recovered graph changes with the ablation rule, metric, data, and estimator (Miller et al., 2024; Zhang & Nanda, 2024; Méloux et al., 2026), and multiple faithful graphs may describe the same behavior (Chen et al., 2026b). Circuits can also vary across datasets or model refits (Rai et al., 2026; Bali et al., 2026), motivating work on agreement and stability (Li & Subramani, 2026; Parekh, 2026; Anani et al., 2026). Feature-level methods instead build circuit nodes from learned features (Cunningham et al., 2024; Dunefsky et al., 2024; Marks et al., 2025; Ameisen et al., 2025; Yin et al., 2026), which other work compares with neurons and alternative representations (Arora et al., 2026; Wu et al., 2025). We ask whether training can make a compact circuit sufficient, changing the target rather than only the estimator.

Training for interpretable structure. Fine-tuning can preserve circuit function even as the heads involved change, but it can also make an earlier localization stale (Prakash et al., 2024; Tigges et al., 2024; Chen et al., 2026a). It has therefore been viewed as a search over subgraphs (Li et al., 2025); we optimize that movement directly. Zhang et al. (2026) find that reinforcement-learning (RL) fine-tuning spreads a behavior's attribution mass when measured with every edge open rather than through a faithful circuit. Other methods encourage sparse weights or modular computation (Gao et al., 2025; Liu et al., 2024). Closest to our setting, Draye et al. (2025) post-train sparse attention connectivity and report faithful task circuits up to 97× smaller by edge count. Their behavior-agnostic constraint limits communication between token positions; Circuit Condensation targets the inter-component graph for one named behavior while preserving held-out task performance and general capability. Other work uses an existing circuit to restrict adaptation (Prakash et al., 2026; Nur'aini et al., 2026) or edits a stored association (Meng et al., 2022; 2023); we optimize the retained circuit itself. The procedure borrows iterative pruning, self-distillation, and low-rank adaptation (Han et al., 2015; Frankle & Carbin, 2019; Hinton et al., 2015; Furlanello et al., 2018; Hu et al., 2022), but removes causal edges rather than parameters.

3 Method: Circuit Condensation

3.1 Problem Setup

We treat a transformer as a directed graph of residual-stream writers and readers (Elhage et al., 2021; Conmy et al., 2023; Bhaskar et al., 2024). Writers are the token embedding, each attention head's residual contribution, and each multilayer perceptron (MLP) output; readers are downstream attention and MLP inputs and the final logits. An edge (w, r) exists whenever w is upstream of r, giving $2{,}041$ candidate edges in GPT-2 up to $43{,}993$ in Qwen3-4B.

A behavior is specified by paired inputs rather than labels: a clean prompt x and a matched corrupted prompt $\tilde{x}$, identical in form but with a different correct answer. The corrupted prompt provides a testable replacement for a removed edge: we replace its clean activation with the corresponding counterfactual activation. Each edge carries a binary gate $z_{wr} \in \{0, 1\}$. Let a_w and $\tilde{a}_w$ be writer w's contributions on x and $\tilde{x}$. The input to reader r is

$$x_r = \sum_{w \prec r} \left[z_{wr} a_w + (1 - z_{wr}) \tilde{a}_w \right],$$

where $w \prec r$ ranges over writers upstream of r. An open edge passes the clean activation. A closed edge passes the corresponding corrupted activation rather than zero, following the *interchange ablation* convention. Zeroing would push activations off the data manifold, confounding "this edge mattered" with "this input is unfamiliar" (Miller et al., 2024).

We want the smallest graph that preserves both the target behavior and general model capability. Let $\mathrm{acc}(z, \theta)$ denote task accuracy under edge mask z and adapter parameters θ, and let $\rho(\theta)$ be the adapted model's off-target perplexity divided by the base model's. Condensation then seeks

$$\min_{z,\,\theta} \|z\|_0 \quad \text{s.t.} \quad \mathrm{acc}(z, \theta) \;\geq\; \mathrm{acc}_{\text{full}} - \varepsilon, \qquad \rho(\theta) \;\leq\; 1 + \kappa, \tag{1}$$

with $\varepsilon = \kappa = 0.05$ throughout. A mask satisfying the accuracy constraint is *faithful*, and $\|z\|_0$ counts its retained edges. This counts graph edges rather than computation inside each retained component (§6). Optimizing θ alongside the mask allows the adapter to reroute computation instead of only searching the existing graph. Each proposed cut must satisfy both constraints before it is accepted. If a behavior resists concentration, the controller therefore returns a larger valid circuit rather than a smaller circuit with damaged performance.

3.2 The Loop

Exact optimization is out of reach because $|E|$ binary gates define $2^{|E|}$ masks, so we approximate Equation 1 greedily (Figure 1). Beginning with every edge open and a fresh low-rank adaptation (LoRA) module (Hu et al., 2022), each round ranks the active edges, prunes the weakest, heals the adapter against the reduced graph, then accepts or rejects the round. The adapter is not pre-trained, so its final weights come only from accepted rounds.

Rank. Ranking needs a per-edge score cheap enough to recompute every round. Edge attribution patching estimates the effect of closing an edge from a gradient instead of running the ablation, and EAP-IG averages that gradient along an interpolation path (Syed et al., 2024; Hanna et al., 2024); we integrate in gate space between the fully corrupted and current circuits rather than between input embeddings (Appendix C). We use EAP-IG because the risk it addresses grows as the graph shrinks: one gradient sample can under-score an important edge in a saturated region, especially at high sparsity (Zhang et al., 2025), who argue the interpolation path can itself pass through such regions, where the controller spends its later rounds. Scores are recomputed after each cut, reflecting what remains.

Prune. The controller first proposes removing the lowest-scoring 30% of active edges. If a proposal fails, it restores the last accepted state and halves the cut, down to 5%. Because the accuracy curve is flat and then falls sharply, one cut size would waste rounds early or overshoot late; halving on rejection shrinks the cuts as the boundary nears.

Heal. Base weights stay frozen. With the mask fixed, only the adapter trains, matching the original model's output distribution rather than the task label. Writing p_{orig} and p_{masked} for the frozen model's distribution and the adapted model through the mask, healing minimizes $\mathrm{KL}(p_{\text{orig}}(\cdot \mid x) \,\|\, p_{\text{masked}}(\cdot \mid x))$. Self-distillation rather than a task loss is deliberate: a label loss would let the adapter relearn the answer through whatever path survives, giving a small circuit that no longer describes the original computation. Matching the distribution constrains outputs, not internals, so §5.3 checks the mechanism for IOI, where a published circuit provides an external test.

Accept or restore. Validation data are disjoint from those used for ranking and healing. A round is accepted only if both the masked circuit and the fully open adapted model meet the accuracy constraint. The fully open model must also pass the held-out off-target capability probe; otherwise both states are restored and a gentler cut is proposed. We use a hard capability gate because performance stays stable for many rounds and can collapse after a single cut. A soft penalty could trade away that collapse, whereas the gate forces the controller to undo it. The threshold is a fixed perplexity ratio, binding unevenly across families (§6). Appendix B gives the schedules and Appendix B.2 the pseudocode.

Validation guarantee and the frozen-weight arm. The greedy procedure does not guarantee optimality, but every returned C_1 satisfies both constraints on held-out validation data because a failing state is never accepted. Section 5.3 tests directly whether a smaller faithful subset exists for circuits that can be enumerated. Fixing θ gives C_0, which keeps the same controller, ranking, and stopping rule but performs no healing.

4 Experimental Setup

4.1 Tasks, Models, and Graphs

Circuit Condensation requires paired inputs that change the correct answer while preserving the prompt format. We therefore study four token-level behaviors with aligned clean and corrupted inputs: indirect object identification (IOI) (Wang et al., 2023), subject-verb agreement across an attractor (Linzen et al., 2016), repeated-token induction (Olsson et al., 2022), and Python docstring completion (Heimersheim & Janiak, 2023). IOI also has a published head-level circuit with named functional roles, letting us test whether condensation retains a known mechanism rather than merely task accuracy.

Across these tasks, we evaluate GPT-2 small (Radford et al., 2019), Llama-3.2-1B/3B-Instruct (Meta AI, 2024), Gemma-3-1B/4B-it (Gemma Team, 2025), and Qwen3-0.6B/1.7B/4B (Yang et al., 2025). In every model, the graph treats each attention head as a separate component and each MLP as one component. Before pruning, we verify that the fully open graph reproduces the native model across all four families. Per-layer head sums match the native block within 6×10^{-7} relative error, and end-to-end logits within 6×10^{-4} (Appendices A and A.1). This check ensures that later differences arise from edge interventions rather than errors in the graph decomposition.

4.2 Metrics and Baselines

We measure task performance using restricted-choice accuracy at one answer position. The correct token must outrank a task-specific set of alternatives, listed in Appendix C.2, so evaluation is a deterministic comparison of logits rather than a model-graded judgment. Accuracy alone does not show whether a circuit reproduces the full model's output behavior. We therefore measure token-level Kullback–Leibler (KL) divergence, which compares their full next-token probability distributions. Across the grid, the reference is the fully open adapted model. Separate checks compare the circuit with the original model and report how often they choose the same candidate (Appendices E and H.4). We separately evaluate capability retention using perplexity on WikiText-2 (Merity et al., 2017) and accuracy on LAMBADA (Paperno et al., 2016). Neither dataset is used by the controller (Appendix E.1).

The frozen comparisons are EAP, EAP-IG, EAP-GP, a random ranking, and the frozen-weight controller C_0 (§3); we additionally run ACDC (Conmy et al., 2023) on the 8 cells where its cost is tractable (Appendix D). We implement all three attribution estimators on the controller's edge gates

using their published scoring equations (Syed et al., 2024; Hanna et al., 2024; Zhang et al., 2025). Because EAP-IG and EAP-GP require adaptation to this gate space, and EAP-GP has no reference implementation, we audit the shared pipeline against exact single-edge activation patching on a tractable graph (Appendix C). Circuit size alone would not validate the implementation. EAP-IG serves as our primary frozen baseline.

4.3 Data, Selection, and Reporting

Each task–model combination uses disjoint splits of 2000 training, 2000 validation, and 1000 sealed test examples. Appendix C describes their construction; the capability probe is separate. Ranking and healing use only the training split. The controller evaluates each proposed circuit on the full validation split, while the sealed test split is read once, after selection.

For every arm, we select the smallest circuit whose validation accuracy is within 0.05 of that arm's fully open graph. Each method is therefore allowed the same accuracy drop relative to its own full graph before circuit sizes are compared (Appendix C.1). Because C_1 is measured against the adapted full graph, adaptation could in principle lower the reference accuracy and make selection easier. In practice, the adapted graph matches the original model's full-graph accuracy to a median 0.000. Using the original model's accuracy minus the same 0.05 still leaves all 96 endpoints qualifying on validation and 91 on the sealed test split. We use *matched accuracy* in §5.3, where each frozen circuit is selected to match the accuracy of the corresponding C_1 endpoint. We report medians over seeds 11, 22, and 33. Statistical tests treat the 32 task–model combinations, rather than the 96 runs, as independent because their three seeds share a model and data.

5 Results

We evaluate four behaviors on eight models with three seeds, giving 32 task–model combinations and 96 runs. All arms share the graph, data, metric, and selection rule from §4. Circuit size is the number of retained head-level edges.

5.1 Condensation Goes Below the Frozen Discovery Floor

At comparable accuracy, C_1 retains fewer edges than frozen EAP-IG in 30 of 32 task–model cells at the seed median (Figure 2). Across all 96 runs, the average gap narrows from $103\times$ under random ordering to $18.4\times$ for EAP, $10.6\times$ for EAP-GP, and $8.1\times$ for EAP-IG, with one EAP-IG comparison reaching $316\times$. The random arm assigns frozen edges a uniform random order and otherwise sweeps the same budgets. This progression shows that ranking improves frozen discovery, but every frozen arm remains larger on average (Table 1).

Table 1: **Frozen arms against C_1.** Ratios are geometric means (frozen/C_1) over all 96 seeded runs. *Cells* reports wins among the 32 task–model combinations at the seed median. CIs resample those combinations; all five comparisons remain significant after Holm correction (Appendix D).

Arm	vs C_1	cells	95% CI
Random	$103\times$	32/32	$[56.8, 189.7]$
EAP	$18.4\times$	31/32	$[10.7, 32.9]$
EAP-GP	$10.6\times$	30/32	$[6.4, 18.1]$
EAP-IG	$8.1\times$	30/32	$[4.9, 13.7]$
C_0	$7.4\times$	29/32	$[4.3, 12.5]$

The exceptions are docstring/Qwen3-0.6B and agreement/GPT-2. At the seed median, C_1 keeps $1{,}877$ rather than 798 edges in the first and 453 rather than 240 in the second. Agreement and docstring also have the weakest average reductions, $5.1\times$ and $2.7\times$, against $17.7\times$ on IOI and $18.2\times$ on induction (§6).

The endpoints hold up on unseen data. Although validation selects among healed candidates, 91 of 96 C_1 endpoints still meet the 0.05 tolerance on the sealed test split against their own full-graph accuracy (Table 8). The other five miss it by a median 0.009.

The reduction depends on which constraint stops the controller. On IOI and induction, where the behavior's own accuracy sets the stopping point, it averages $18.0\times$ over 16 cells; where the capability gate binds first it is smaller (§6). Because that gate rarely binds on Gemma, we also report the comparison without its two models: $5.8\times$ on average, with C_1 smaller in 22 of 24 cells.

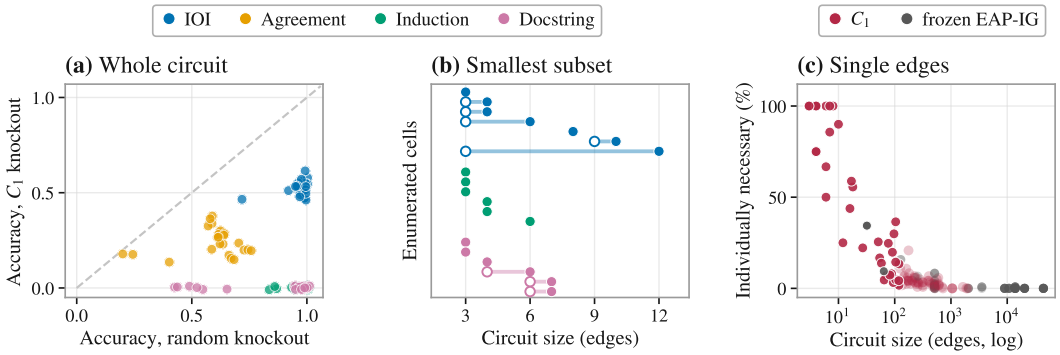


Figure 3: **Three causal checks enabled by smaller circuits.** **(a)** Accuracy after removing C_1 or equally many random edges. **(b)** Smallest faithful subset (open) and C_1 endpoint (filled), 19 enumerable runs. **(c)** Share of individually necessary edges; solid markers test every edge, faded ones estimate from 120 sampled edges.

5.2 The Gain Comes From Reshaping, Not a Better Search

C_1 and C_0 differ only in whether θ may move, isolating the effect of weight updates. With the same ranking, schedule, and stopping rule, C_0 is larger in 29 of 32 cells by an average $7.4\times$ (Table 1). Moving the weights without the loop is also insufficient. DPO, GRPO, and REINFORCE (Rafailov et al., 2023; Shao et al., 2024; Williams, 1992) are inconclusive on these short-answer tasks (Appendix O), so we compare supervised fine-tuning (SFT). A standard LoRA followed by the same EAP-IG sweep yields circuits about half the size of frozen circuits but still $4.7\times$ larger than C_1; C_1 is smaller in 28 of 32 cells (Appendix D). Fine-tuning also concentrates edge attribution under the entropy and kurtosis measures of Zhang et al. (2026), reducing the participation ratio—the effective number of edges carrying attribution mass—to 0.76 of its initial value in 23 of 32 cells. Its probe perplexity is $1.37\times$, above the controller's 1.05 limit. Fine-tuning thus moves circuits toward condensation but only partway, roughly halving the graph where condensation removes most of it. A stronger baseline jointly trains gates and adapter under the same objective without the iterative loop. It finds no non-empty circuit preserving both task accuracy and general capability in 20 of 32 cells, including 11 outright capability-gate failures. Of the 12 valid circuits, 5 are smaller than C_1 (Appendix F.3).

5.3 The Circuit Is Small Enough to Verify

Smaller circuits are useful only if they still carry the behavior and make stronger causal checks tractable. We test both across the grid and compare the IOI circuit with a published mechanism.

The retained set carries the behavior. Across the grid, the circuits returned by the controller range from 3 to 1,976 edges. For each run, we remove the entire retained set while leaving the rest of the graph open; the random control removes exactly the same number of edges. On IOI, median accuracy falls from 0.99 to 0.52, near the 0.50 chance level of its two-choice metric, while random removal leaves it at 0.98. The corresponding intact/circuit-cut/random values are $0.74/0.22/0.62$ for agreement, $1.00/0.00/0.96$ for induction, and $1.00/0.00/0.97$ for docstring (Figure 3a; Appendix Table 21).

On IOI the circuit is a sufficient part of the published mechanism. Wang et al. (2023) assign named roles to 26 GPT-2 heads. Frozen discovery finds 25 of them among 61 heads, so 36 of its heads carry no published role. Condensation returns 24 heads, 17 named, giving a per-seed median precision gain of $2.0\times$ (Table 19; Appendix H). It keeps all three primary Name Movers and sheds three of the eight Backup Name Movers. These heads primarily take over after the primary Name Movers are ablated (Wang et al., 2023), so rankings computed on the intact model may give them little weight. We then interchange the variable assigned by Wang et al.'s algorithm to the subject-inhibition heads, using the same edge indices in every arm. The intervention flips the answer as predicted in 0.930 of pairs for C_1, against 0.735 for a matched frozen circuit and 0.645 for the unmodified model; a size-matched random set scores 0.03 and a sham hypothesis 0.00 (Appendix H.1).

Table 2: **Shorter pruning paths fail at the smallest C_1 endpoints.** Every route targets C_1's edge count. Each cut receives up to 2,000 healing steps, four times one controller round. *Five-stage* re-ranks and heals after five cuts, *single cut* ranks once and prunes directly, and *random* keeps an equal-size mask. We score task accuracy only and do not apply the capability gate. Columns report seed-11 medians across 32 cells by endpoint size. *Matches* counts cells within 0.05 of C_1 (Appendix F).

Route	$k \leq 16$ (n=7)	17–200 (n=10)	>200 (n=15)	matches C_1
C_1, re-ranks every round	0.965	0.962	0.969	—
Five-stage cut	0.450	0.780	0.977	21
Single cut	0.026	0.533	0.823	15
Random mask	0.001	0.250	0.479	0

Single-edge ablations show where slack remains. For each circuit, we switch off every retained edge and count the fraction whose removal lowers accuracy by at least 0.05. Across tasks, the median ranges from 2.5–15%. Among circuits with at most 8 edges it is 100%, whereas no edge in the largest docstring circuit crosses the threshold alone. The matched frozen circuits have a median of 0.0% (Figure 3c; Appendix Table 21). Edges below the threshold may still have smaller or joint effects.

Exhaustive subset tests determine exact minimality. Single-edge tests cannot rule out a smaller faithful combination. For the 19 circuits small enough to enumerate, we test every subset up to $k \leq 13$, totaling $5{,}952$ masks. Eleven have no smaller faithful subset; 7 of the other 8 have only 1 to 3 removable edges. Across all 19 runs, the smallest faithful subset has 3–9 edges. Exactly minimal circuits have a median of 3 edges versus 6.5 for those with removable edges; in the worst case, the controller keeps 12 although 3 suffice. Slack thus appears mainly in larger endpoints, consistent with stopping on a flat accuracy curve. Every matched frozen circuit exceeds the cap: the smallest has 68 edges and the median 416 (Figure 3b; Appendix Table 21).

Small does not mean independent. Minimality shows whether edges can be removed, not whether those remaining act independently. We ablate all $\binom{k}{2}$ pairs and compare each pair's joint effect with the sum of its individual effects. Across $169{,}476$ ablations over 48 runs, every circuit has an interacting pair. The median number of interacting partners per edge is 7.4 across circuits, including in the 3-edge Gemma circuits. Condensation preserves these dependencies but makes exhaustive tests practical. Matched frozen circuits would require a median $55\times$ as many pair ablations, and only 15 of 32 cells have projected runtime under one day (Appendix G.1). Exact single-edge patching is also $3.6\times$ faster on the smaller graph (Appendices G.1 and J).

5.4 The Circuit Describes the Model We Started From

By construction, C_1 preserves the target behavior after condensation. We next ask whether its edges were already important in the original model. We rank the original edges with EAP-IG and select a frozen circuit by the same rule. It contains a median 100% of C_1's edges versus 8% for an equal-size random set, with above-baseline overlap in all 32 cells. Including every C_1 edge requires a median 13% of the ranked list (Appendix N). Condensation therefore concentrates behavior into edges that frozen discovery already ranks highly. Because their weights moved, overlap alone cannot show that the computation is unchanged. Across 96 runs, C_1 selects the same answer as the original model on a median 96.5% of sealed-test examples, with median token-level KL 0.215. On IOI/GPT-2, its KL is 0.056 versus 0.50 for the matched frozen circuit (Appendix H.4). At equal edge count, C_1 predicts original-model errors with median AUROC 0.894 versus 0.645, leading in 10 of 11 evaluable cells, all on agreement or induction (Appendix I).

5.5 Controls

The pruning path matters. We test three shorter routes at C_1's final edge count, allowing up to 2,000 healing steps per cut. None matches C_1 at endpoints of 16 edges or fewer: C_1 retains 0.965 accuracy versus 0.450 for the five-stage cut, 0.026 for the single cut, and 0.001 for the random mask (Table 2). The random mask matches none of the 32 cells. The single cut matches it in 15, but those

Table 3: **Controls**, in discussion order. Each row changes one ingredient; *Reference* gives its comparison. Outcomes are medians over runs. The target-test row reports IOI (other targets shrink 2–8×). PR is the EAP-IG participation ratio after versus before healing (Appendix F).

Control	What changes	Cells	Seeds	Outcome	Reference
Shorter path (Tab. 2)	route to the endpoint	32	1	0/7 match at $k \leq 16$	C_1 at equal size
Random order	edge ranking	32	3	$50\times$ larger	C_1 endpoint
Reverse order	edge ranking	32	1	0 edges removed	full graph
No-prune heal	pruning removed	32	3	PR ratio 0.99	before healing
Target test	behavior condensed	32	3	target $16\times$, rest $1.0\times$	untouched model
Sibling swap	head identity	32	3	30/32 to chance	intact circuit

cells retain $12\times$ more of the original graph than cells where it fails. Above 200 edges, the five-stage cut ties C_1: it wins 6 of 15 cells, C_1 wins 8, and 1 ties (median difference -0.001, sign test $p = 0.79$). All three know C_1's edge count in advance; the controller must discover where to stop.

Attribution ranking is necessary for deep pruning. With the controller fixed, we use random or reverse edge rankings. Random ordering returns circuits a median $50\times$ larger than C_1 and no closer than $2.6\times$. Reverse ordering accepts no cuts, returning the starting graph at a median $64\times$ the size of C_1 across 32 cells. Both preserve accuracy only by keeping most of the graph (Appendix F).

Healing alone does not concentrate attribution. With every edge open, we heal the model and measure the EAP-IG participation ratio; higher values mean attribution is spread across more edges. Its median end-to-start ratio is 0.99, and 92 of 96 runs remain within 15% of their starting value. The reduction therefore requires pruning and healing together (Appendix F.1).

Concentration remains tied to the target behavior. A generic sparsity effect would also simplify unrelated behaviors. We condense each behavior, then apply the same frozen search to all four at matched accuracy in the adapted and original models. Only the target moves: across eight models and three seeds, IOI becomes a median $16\times$ easier to localize ($8\times$–$128\times$), induction $8\times$, agreement $4\times$, and docstring $2\times$. All twelve off-target directions remain at $1.0\times$ (Appendix F).

The circuit depends on which heads feed the output. A small circuit could still be arbitrary if another head in the same layer worked as well. We replace each selected head feeding the final output with another from its layer. Median accuracy falls by 0.62, leaving 30 of 32 circuits no better than chance, including all seven endpoints with at most 16 edges. Adding a random edge leaves accuracy unchanged, so replacing the selected heads causes the collapse (Appendix F.2).

6 Limitations

We test four token-level behaviors on fixed checkpoints up to 4B parameters and three seeds, so results may not transfer to other checkpoints, distributions, or graph granularities. C_1 describes the adapted model and closely, but not necessarily exactly, tracks the original. Edge counts depend on interchange ablation and exclude computation within retained components and the LoRA adapter. Exact minimality is feasible for only 19 endpoints. Three seeds test stability, not uniqueness, while flat accuracy curves can make the fixed schedule stop above the minimum. The IOI circuit is sufficient without reconstructing the full published mechanism. Condensation requires training and raises unseen WikiText-2 perplexity by a median 9% (Appendix M).

7 Conclusion

Circuit Condensation turns circuit findability from a fixed-model property into a training target. Across 32 settings, it returns circuits $8.1\times$ smaller than the strongest frozen baseline in 30. The resulting circuits make exhaustive subset and pair tests feasible while remaining concentrated in pathways that frozen discovery ranks highly in the original model. This makes causal analysis cheaper to repeat and expands what can be verified directly.

Reproducibility Statement

Appendix A documents the model checkpoints and graph construction. Appendices B and C give the controller settings, per-run configurations, prompt generation, data splits, scoring, and selection rule. Each run uses one 40GB or 80GB NVIDIA A100 GPU and records its seed, launch settings, pruning history, final mask, LoRA adapter, and validation and test outputs. We manually verified the reported results against the saved JSON outputs and individual runs. Code and data-generation scripts will be released publicly upon acceptance.

AI Use Statement

We used generative AI tools to assist with code implementation and debugging, experiment orchestration, consistency checks, figure and table preparation, literature searches, suggestions on experimental controls and interpretation, and suggestions for drafting and revising prose to improve clarity, flow, and professional presentation. The authors designed the experiments, generated and evaluated all outputs, interpreted the results, verified the experiments and reported findings, checked citations against original sources, and wrote and approved the final manuscript. Experimental examples were produced by deterministic scripts rather than directly by an LLM, and model outputs were scored without LLM grading. AI-assisted code was manually reviewed and tested, and reported results were manually verified against saved JSON outputs and individual runs. We take responsibility for all text, code, analyses, and conclusions.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4895–4901, 2023. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298/. arXiv:2305.13245.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, et al. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Alaa Anani, Tobias Lorenz, Bernt Schiele, Mario Fritz, and Jonas Fischer. Certified circuits: Stability guarantees for mechanistic circuits. *arXiv preprint arXiv:2602.22968*, 2026.

Aryaman Arora, Zhengxuan Wu, Jacob Steinhardt, and Sarah Schwettmann. Language model circuits are sparse in the neuron basis. In *International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=OrviwFWcN4. Spotlight; arXiv:2601.22594.

Karan Bali, Jack Stanley, Praneet Suresh, and Danilo Bzdok. Quantifying LLM attention-head stability: Implications for circuit universality. *arXiv preprint arXiv:2602.16740*, 2026.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.16778.

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, 2022. URL https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing. AI Alignment Forum.

Hang Chen, Jiaying Zhu, Hongyang Chen, Hongxu Liu, Xinyu Yang, and Wenya Wang. Navigating by old maps: The pitfalls of static mechanistic localization in LLM post-training. *arXiv preprint arXiv:2605.06076*, 2026a.

Xi Chen, Mingyu Jin, Jingcheng Niu, Yutong Yin, Jinman Zhao, Bangwei Guo, Dimitris N. Metaxas, Zhaoran Wang, Yutao Yue, and Gerald Penn. All circuits lead to rome: Rethinking

functional anisotropy in circuit and sheaf discovery for LLMs. In *International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, 2026b. arXiv:2605.12671.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

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 (NeurIPS)*, volume 36, 2023. URL https://openreview.net/forum?id=89ia77nZ8u. arXiv:2304.14997.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.08600.

Florent Draye, Anson Lei, Hsiao-Ru Pan, Ingmar Posner, and Bernhard Schölkopf. Intrinsically interpretable attention via sparse post-training. *arXiv preprint arXiv:2512.05865*, 2025.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.11944.

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

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

Philip J. Fleming and John J. Wallace. How not to lie with statistics: the correct way to summarize benchmark results. *Communications of the ACM*, 29(3):218–221, 1986. doi: 10.1145/5666.5673.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1803.03635.

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616, 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html. arXiv:1805.04770.

Leo Gao, Achyuta Rajaram, Jacob Coxon, Soham V. Govande, Bowen Baker, and Dan Mossing. Weight-sparse transformers have interpretable circuits. *arXiv preprint arXiv:2511.13653*, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 9574–9586, 2021. arXiv:2106.02997.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Gemma Team. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, and Yonatan Belinkov. Position-aware automatic circuit discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2792–2817, 2025. doi: 10.18653/v1/2025.acl-long.141. URL https://aclanthology.org/2025.acl-long.141/.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. arXiv:1506.02626.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2403.17806; introduces EAP-IG.

Stefan Heimersheim and Jett Janiak. A circuit for Python docstrings in a 4-layer attention-only transformer. AI Alignment Forum, 2023. URL https://www.alignmentforum.org/posts/u6KXXmKFbXfWzoAXn/a-circuit-for-python-docstrings-in-a-4-layer-attention-only.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Aliyah R. Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Y. Odisho, Peter R. Carroll, and Bin Yu. Efficient automated circuit discovery in transformers using contextual decomposition. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/31044. arXiv:2407.00886.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Michael Li and Nishant Subramani. How much do circuits tell us? measuring the consistency and specificity of language model circuits. *arXiv preprint arXiv:2605.08348*, 2026.

Yueyan Li, Wenhao Gao, Caixia Yuan, and Xiaojie Wang. Fine-tuning is subgraph search: A new lens on learning dynamics. *arXiv preprint arXiv:2502.06106*, 2025.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics (TACL)*, 4: 521–535, 2016. doi: 10.1162/tacl_a_00115. arXiv:1611.01368.

Ziming Liu, Eric Gan, and Max Tegmark. Seeing is believing: Brain-inspired modular training for mechanistic interpretability. *Entropy*, 26(1):41, 2024. doi: 10.3390/e26010041. URL https://doi.org/10.3390/e26010041.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1711.05101.

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 (ICLR)*, 2025. arXiv:2403.19647.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic interpretability as statistical estimation: A variance analysis. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2510.00845.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. ROME; arXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023. MEMIT; arXiv:2210.07229.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. Introduces the WikiText-2 and WikiText-103 datasets; arXiv:1609.07843.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Meta AI, 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2407.08734.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html. arXiv:2504.13151.

Khumaisa Nur'aini, Ayu Purwarianti, Alham Fikri Aji, and Derry Wijaya. Beyond transfer accuracy: Mechanism-guided controlled adaptation for low-resource languages. *Findings of EMNLP*, 2026. arXiv:2601.08146.

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

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. arXiv:2209.11895.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1525–1534, 2016. doi: 10.18653/v1/P16-1144.

Swapnil Parekh. CIRCUS: Circuit consensus under uncertainty via stability ensembles. *arXiv preprint arXiv:2603.00523*, 2026.

Yunxiang Peng, Mengmeng Ma, Ziyu Yao, and Xi Peng. Inside-out: Measuring generalization in vision transformers through inner workings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38936–38946, 2026. Highlight; arXiv:2604.08192.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=8sKcAWOf2D.

Nikhil Prakash, Donghao Ren, Dominik Moritz, and Yannick Assogba. Constructive circuit amplification: Improving math reasoning in LLMs via targeted sub-network updates. In *ICLR 2026 Workshop on Trustworthy AI*, 2026. URL https://openreview.net/forum?id=zmkkBuz8ME. arXiv:2512.16914.

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.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.18290.

Daking Rai, Mor Geva, and Ziyu Yao. Data-driven circuit discovery for interpretability of language models. *arXiv preprint arXiv:2605.09129*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063. arXiv:2104.09864.

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, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/. arXiv:2310.10348.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2407.10827.

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 (ICLR)*, 2023. arXiv:2211.00593.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

An Yang et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Naiyu Yin, Dennis Wei, Tian Gao, Amit Dhurandhar, Karthikeyan Natesan Ramamurthy, and Yue Yu. Scalable circuit learning for interpreting large language models. In *Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.16939. arXiv:2606.16939.

Lei Yu, Jingcheng Niu, Zining Zhu, Xi Chen, and Gerald Penn. Sheaf discovery with joint computation graph pruning and flexible granularity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8822–8837, 2025. doi: 10.18653/v1/2025.emnlp-main.446. URL https://aclanthology.org/2025.emnlp-main.446/.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 12360–12371, 2019. arXiv:1910.07467.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.16042.

Honglin Zhang, Qianyue Hao, Fengli Xu, and Yong Li. Reinforcement learning fine-tuning enhances activation intensity and diversity in the internal circuitry of LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.21044.

Lin Zhang, Wenshuo Dong, Zhuoran Zhang, Shu Yang, Lijie Hu, Ninghao Liu, Pan Zhou, and Di Wang. EAP-GP: Mitigating saturation effect in gradient-based automated circuit identification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://nips.cc/virtual/2025/poster/116294. arXiv:2502.06852.

A Models, Graphs, and Head Decomposition

We test four behaviors on eight models: GPT-2 small, Llama-3.2-1B/3B-Instruct, Gemma-3-1B/4B-it, and Qwen3-0.6B/1.7B/4B. The checkpoints are openai-community/gpt2, meta-llama/Llama-3.2-1B-Instruct, meta-llama/Llama-3.2-3B-Instruct, google/gemma-3-1b-it, google/gemma-3-4b-it, Qwen/Qwen3-0.6B, Qwen/Qwen3-1.7B, and Qwen/Qwen3-4B. Each attention head is a separate graph writer; each MLP is one writer. The candidate graph grows from $2{,}041$ edges on GPT-2 to $43{,}993$ on Qwen3-4B (22 times larger). Table 20 reports every endpoint. We check the graph decomposition against each model's unmodified forward pass before using it for circuit experiments.

A.1 Head Decomposition Audit

For each architecture, we split the native attention output before the residual addition and apply the model's own output projection to each head slice. This preserves grouped-query attention and rotary position embeddings (Ainslie et al., 2023; Su et al., 2024), architecture-specific head dimensions, and any LoRA module attached to the projection. Gemma-3 applies root mean square layer normalization (RMSNorm) after summing the attention output (Zhang & Sennrich, 2019). We therefore compute its shared position-wise scale from the complete sum and apply it to each head contribution, so the scaled contributions add to the native normalized output. We test every architecture with both zero and nonzero adapters. Per-layer head sums match the native block within 6×10^{-7} relative error, fully open end-to-end logits match within 6×10^{-4}, and random masks and EAP-IG passes produce no non-finite values.

B Controller Configuration and Pseudocode

Algorithm 1 gives the controller used in every adaptive run. It ranks active edges, cuts the lowest-ranked ones, trains the adapter, and restores the previous state if accuracy or capability fails. We report the smallest accepted circuit; later rounds only show where further cuts fail (Table 4).

B.1 Run Configuration

Table 4 lists the settings shared across the grid. Controls change them only where stated.

Table 4: **Controller and run configuration.** Values are shared across all 96 runs and 32 cells unless a control states otherwise.

Setting	Value
Data	
Data per task and model	2000 train / 2000 validation (`eval_big.jsonl`) / 1000 sealed test
Validation read per controller decision	full 2000-example split
Paper seeds	11, 22, 33
Ranking and pruning	
Edge-ranking method	EAP-IG, re-ranked after every accepted cut
Integrated-gradient steps	5 interpolation points
Ranking / validation batches per round	8 / 250
Initial proposed cut	30% of active edges
Healing	
Healing objective	$\mathrm{KL}(p_{\text{orig}} \| p_{\text{masked}})$ on clean inputs
Healing steps per proposed cut	up to 500; checked every 40 steps and stopped once circuit accuracy clears the floor by a buffer (EMA of recent per-cut drops, seeded at $0.5\times$ the tolerance, floored at 0.005)
Optimizer / learning rate	AdamW (Loshchilov & Hutter, 2019) / 10^{-4}
Trainable parameters	LoRA adapter only; base model frozen
LoRA configuration	rank 16, $\alpha = 32$, dropout 0, no bias terms
Acceptance	
Cut on rejection	halved, floored at 0.05 of active edges
Re-heal attempts once the cut reaches its floor	3
Overshoot rounds past the endpoint	4
Mid-heal abort check	every 40 steps
Adapter weight decay	0.01
Healing / evaluation batch size	16 / 8
Edge floor (smallest returnable circuit)	1
Gate on the frozen arm (C_0)	inert (calibration is degenerate when no weight moves)
Capability-probe batches	50
Numerical precision	`bfloat16`
Accept accuracy tolerance	within 0.05 of the full-circuit validation baseline, tested on a one-sided 95% lower bound ($z = 1.645$)
Capability probe	320 prompts across five domains
Capability gate / mid-heal abort	perplexity ratio ≤ 1.05 / > 2.0
Reporting	
Arm selection rule	smallest circuit within 0.05 of *that arm's own* full-graph validation accuracy
Reported endpoint	smallest accepted faithful circuit, scored once on sealed test
Hardware	one NVIDIA A100 GPU per run (40GB or 80GB)

B.2 Controller Pseudocode

Algorithm 1 expands the loop shown in Figure 1.

Algorithm 1 Circuit Condensation controller

Require: Frozen model M; train pairs $\mathcal{D}_{\text{tr}}$; validation data $\mathcal{D}_{\text{val}}$; capability probe $\mathcal{D}_{\text{cap}}$
Require: Open mask z; initial cut fraction q; tolerances $\epsilon = \kappa = 0.05$
Initialize LoRA parameters θ; save (z, θ) as the accepted state
repeat
    Score active edges with EAP-IG on $\mathcal{D}_{\text{tr}}$
    Propose z' by removing the lowest-ranked fraction q of active edges
    Optimize θ' for up to 500 steps with z' fixed to minimize $\mathrm{KL}(p_M \| p_{M,z',\theta'})$
    Evaluate the masked circuit and full adapted model on $\mathcal{D}_{\text{val}}$
    Measure the adapter-on/off perplexity ratio on $\mathcal{D}_{\text{cap}}$
    if both target accuracies are within ϵ of baseline and the ratio is $\leq 1 + \kappa$ **then**
        Accept and save (z', θ'); continue from the smaller graph
    else
        Restore the accepted (z, θ); reduce q
until no smaller proposal can be accepted
return the smallest accepted mask and adapter, C_1

C Selection Protocol and Data Provenance

Clean and corrupted pairs. When an edge is closed, it reads an activation from a corrupted prompt. Each clean prompt therefore has a token-aligned counterpart that changes the answer while preserving the task's structure.

IOI. Clean prompts place the indirect object and subject in either order (ABBA and BABA templates). The corruption keeps the first clause but replaces the subject's second occurrence, the one performing the action, with a third name. The resulting prompt names three people and no longer implies the clean answer.

Agreement. We change the head noun's number and every slot that agrees with it. We also change the attractor so it stays opposite in number to the head; the correct verb changes, but the attractor structure does not. Prompts use three frames from the agreement literature.

Induction. The bigram $A\,B$ appears four times before a final cue A (`--repeats 4`). A 6-token prefix and 6-token fillers separate repetitions, drawn from 50 single-token words. The corruption replaces every B with a third word C, keeping the repeated pattern while changing its target.

Docstring. We rename parameters to change the in-order copy target without changing the function shape. All substituted words are single tokens under the model's tokenizer; the agreement and docstring generators discard pairs whose tokenized lengths differ. The generators are `scripts/prepare_{ioi,agreement,induction,docstring}.py`, and `scripts/prepare_bigval.py` makes the 2000-example validation split. We pass split sizes explicitly because the scripts default to smaller samples.

Splits and capability probe. We generate prompts separately for each tokenizer because substituted names and answer choices must remain single tokens. Within a task–model combination, the splits draw from the same template and lexical pools but share no clean prompts, and every arm uses the same data. Generation uses seed 0 for training, 2 for the sealed test split, and 777 for the 2000-example validation split, which excludes prompts from the existing splits. The fixed capability probe contains 320 prompts: 80 each from docstring code, IOI-style text, and GSM8K math (Cobbe et al., 2021), plus 40 each from general prose and a history task. Task-shaped prompts come from training splits. The probe is sampled once with seed 0 and reused across runs, and every run checks that it overlaps with neither validation nor test data. WikiText-2 and LAMBADA are separate evaluations that never enter the controller. The probe builder is `scripts/build_probe_corpus.py`.

Frozen-estimator implementation audit. EAP, EAP-IG, and EAP-GP score the edge gates in our graph. EAP uses the published attribution equation. EAP-IG holds the activation difference fixed

and averages gradients along an interpolation path in gate space, rather than the embedding space used by Hanna et al. (2024). EAP-GP applies the published GradPath updates in gate-coefficient space. No reference implementation of EAP-GP was released, so we follow its equations.

We check the shared ranking pipeline on a 325-edge GPT-2/IOI block graph with 16 examples. EAP-IG scores correlate with exact single-edge activation-patching effects at Spearman 0.545; a random ranking scores -0.055. This tests whether our pipeline tracks edge effects. It does not establish numerical equivalence to the authors' implementations.

Each `knee.json` records the full-circuit baseline, pruning rounds, selected endpoint, and sealed-test score. Frozen baselines are stored as curves over edge budgets. We report seeds 11, 22, and 33. An earlier pilot seed is excluded because its capability probe overlapped with its test data.

C.1 Selection and Evaluation Data

Table 5 breaks the headline reduction into four tasks. Each row includes all 24 model–seed runs.

Table 5: **Circuit-size reduction relative to frozen EAP-IG, by task.** Each task has 24 runs (eight models, three seeds). We select the smallest circuit within 0.05 of each arm's own full-graph validation accuracy (§4). Reductions are geometric means of per-run ratios; the maximum is the largest single-run reduction.

Task	Runs	Ratios available	Reduction (geom. mean)	C_1 fewer edges	Max reduction
IOI	24	24	17.7×	24/24	209.2×
Agreement	24	24	5.1×	21/24	30.6×
Induction	24	24	18.2×	24/24	316.0×
Docstring	24	24	2.7×	19/24	35.5×
Overall	96	96	8.1×	**88/96**	316.0×

Table 6 records which data are used to rank edges, select circuits, and report each quantity.

Each arm is evaluated against its own full graph. This could favor C_1 if adaptation lowered full-graph accuracy. The adapted-minus-original difference is 0.000 at the validation median, although the adapted graph scores lower in 42 of 96 runs and higher in 24. Using the *original* graph's accuracy minus 0.05 as an absolute bar, all 96 C_1 endpoints still qualify on validation and 91 qualify on sealed test. Thus, the arm-relative rule does not account for the reported reduction.

Table 6: **Data used for selection and reporting.** One rule determines the primary sizes. The analyses of §5.3 re-evaluate those circuits on the sealed test split, using a subsample when they require thousands of masked circuits. We verify each sample size from the stored accuracy denominator rather than the configuration.

Quantity	Ranked on	Selected on	Reported on	Examples
Primary sizes, every arm	train	validation (own full graph -0.05)	sealed test	2000 / 1000
Matched frozen circuit (§5.3)	train	matched to C_1 test endpoint	sealed test	1000
Subset enumeration	–	circuits fixed above	sealed-test subsample	200
Single-edge ablations	–	circuits fixed above	sealed-test subsample	96
Whole-set ablation, pairs, roles, specificity	–	circuits fixed above	sealed-test subsample	200

Frozen EAP and EAP-IG curves and C_1 endpoints are scored on the same 1000 sealed-test examples. When an analysis matches a frozen circuit to C_1's test accuracy, it uses that full split. Edge rankings still use training data only.

C.2 Per-Task Answer Sets

All four tasks use restricted-choice accuracy at one answer position, requiring the correct token to outrank a task-defined set of alternatives. IOI and agreement use two-way minimal pairs: the indirect object against the subject, and a verb against the same verb in the wrong number. Induction selects from a fixed pool of plausible copy targets filtered to valid tokens for each model, with the reciprocal of the pool size as its chance baseline. Docstring completion compares the four single-token argument names in the signature, with the correct third argument required to outrank the others.

D Headline Statistics and Per-Task Reductions

The size comparison uses 96 runs, but the three seeds for a task–model pair are related. The statistics below account for that dependence.

Seeds for one task–model pair share a dataset, base model, and graph. We therefore summarize each of the 32 pairs by its seed median. For reduction ratios, we report the geometric mean (Fleming & Wallace, 1986). The arithmetic mean is $29.6\times$ with interval $[11.7, 52.6]$; only 17 of 96 runs reach that mean. The geometric mean is less driven by those large ratios and agrees with the per-task summary because each task contributes equally many runs.

Confidence intervals use a cluster bootstrap over task–model pairs with 20,000 resamples. Every run has a selected circuit under the rule in §4, so all 32 pairs enter the test (Table 5). With eight pairs per task, the smallest attainable sign-test value is $p = 0.0078$. We correct the five comparisons in Table 1 with Holm–Bonferroni; all remain significant at 0.05, with largest adjusted $p = 2.6 \times 10^{-6}$. The task-level intervals are $[8.0, 32.9]$ for IOI, $[2.5, 9.7]$ for agreement, $[6.3, 61.1]$ for induction, and $[1.1, 6.7]$ for docstring. Treating all 96 runs as independent would give $p \approx 10^{-29}$ and overstate the evidence.

For the matched frozen circuit in §5.3, we match C_1's sealed-test accuracy. Its validation score was checked repeatedly during healing and could be optimistic out of sample.

Like every arm, C_0 is selected as its smallest circuit within 0.05 of its own full-graph validation accuracy. It stops within a median 0.03 validation accuracy of C_1. The per-task C_1–C_0 gap is $22.7\times$ on IOI, $4.5\times$ on agreement, $12.4\times$ on induction, and $2.4\times$ on docstring.

The overall $8.1\times$ reduction might be driven by cells where the capability gate rarely binds. Excluding Gemma leaves $5.8\times$, with C_1 smaller in 22 of 24 cells. Restricting to IOI and induction, whose own accuracy sets the endpoint, gives $18.0\times$ over 16 of 16 cells. Applying both restrictions gives $13.6\times$ over 12 of 12 (§6).

We also apply the common EAP-IG sweep to each task-SFT model. SFT alone concentrates the circuit: its selected circuits are about half the size of the unmodified model's and smaller in 24 of 32 cells. C_1 is still a geometric-mean $4.7\times$ smaller than SFT followed by discovery and wins in 28 of 32 cells. SFT wins only on agreement and docstring cells with weaker endpoint accuracy (Table 20).

On sealed test, 91 of 96 C_1 endpoints meet the tolerance. Both C_1 and EAP-IG meet their own tolerances in 71 runs; C_1 is smaller in 64 of those.

The run logs make validation reuse measurable. The controller makes a median of 32 validation decisions per run, counting accepted rounds, rejected proposals, aborted heals, and re-heal probes (interquartile range 25.25–37, maximum 52). The sealed test split is read once per endpoint and never enters those decisions.

D.1 Fixed-Performance Discovery Payoff

We repeat circuit selection at 90% of each arm's full-graph accuracy instead of the main 0.05 tolerance (Table 7).

Table 7: **Edges needed to recover 90% of full-model accuracy** (32 cells, three seeds). We compare C_1 with the smallest frozen alternative among EAP, EAP-IG, random ranking, ACDC, and C_0. ACDC covers 8 cells on the two smallest graphs, where runs take 3.4–5.9 hours; we do not run it on larger graphs. It uses the attribution arms' answer-slot KL metric. C_0 is best in 78 of 96 runs and ACDC in one. Ratios above $1\times$ favor condensation. C_1 loses eight runs, all on agreement or docstring, where the capability gate stops it early (§6).

Task	Runs	Median reduction	Wins	$\geq 3\times$
IOI	24	$14.2\times$	24/24	21/24
Agreement	24	$3.8\times$	21/24	16/24
Induction	24	$8.7\times$	24/24	21/24
Docstring	24	$1.7\times$	19/24	8/24
Overall	96	$5.2\times$	88/96	66/96

E Endpoints, Faithfulness, and Capability Retention

Table 8 reports sealed-test KL from each circuit to its adapted full model, alongside the tolerance pass counts against the original model's accuracy. Five of 96 endpoints miss the tolerance, by a median 0.009. Agreement has the highest KL and the lowest original-model accuracy (0.764 at the test median).

Table 8: **Sealed-test faithfulness across the head grid** (96 runs, medians with interquartile ranges). Agreement is the only task where circuit accuracy exceeds original-model accuracy at the median (0.777 versus 0.764). Interchange ablation can sharpen the margin by removing distractor pathways outside the circuit, which also explains logit-difference recoveries above 1.

Task	Median KL	KL IQR	Median circuit acc.	Median original acc.	Sealed pass
IOI	0.048	[0.045, 0.063]	0.978	0.996	24/24
Induction	0.081	[0.060, 0.124]	0.951	0.965	19/24
Docstring	0.089	[0.074, 0.138]	0.994	1.000	24/24
Agreement	0.254	[0.145, 0.343]	0.777	0.764	24/24
All	0.088	–	–	–	91/96

E.1 General-Capability Retention

With every gate open, the adapted models have a median 9% higher WikiText-2 perplexity (Merity et al., 2017) than the originals; the worst ratio is $1.7\times$. LAMBADA accuracy (Paperno et al., 2016) falls by 5–7 points (Table 9). The controller's 1.05 gate uses a separate mixed-domain probe. If that threshold were applied to WikiText-2, 18–19 of 32 cells would exceed it.

Table 9: **General-capability retention of C_1.** We report the WikiText-2 perplexity ratio (adapted/original), per-token KL from the original, and change in LAMBADA last-token accuracy. The mixed-domain gate corpus is separate from both datasets. Values are medians across 32 cells for each seed. Seed 22 has 31 cells, giving 95 of 96 runs in total.

Seed	Median ratio	Worst ratio	Median KL	Median ΔLAMBADA	Cells > 1.05
11	1.088	1.672	0.278	-0.068	19/32
22	1.091	1.742	0.255	-0.057	19/31
33	1.087	1.621	0.262	-0.052	18/32

F Control Protocols and Per-Cell Results

The controls change one part of the procedure at a time (Table 3). Tables 11 and 12 give the healing-only and head-swap results by cell.

Healing alone does not concentrate. We keep every edge open and run healing with the same capability gate. The EAP-IG participation ratio measures how many edges share the attribution mass. Its median end-to-start ratio is 0.99; 92 of 96 runs stay within 15% of their starting value. The four exceptions are IOI runs on Qwen3 models, with the lowest ratio at 0.76.

Attribution ranking is necessary. We replace EAP-IG with random ordering or cut the highest-ranked edges first. Random ordering returns circuits a median $50\times$ larger than C_1 across 32 cells and three seeds; even its closest run is $2.6\times$ larger (Table 10). Reverse ordering, run once per cell, accepts no cuts and retains the full graph.

Table 10: **Ranking controls.** Only the edge order changes. Values are seed medians except for deterministic anti-ranking, which retains the full graph. Random and anti-ranked accuracies are measured on validation, while EAP-IG accuracy is measured on the sealed test split. These controls do not use a C_1 endpoint.

Task	Model	EAP-IG (ours) edges @ acc	Random edges @ acc	Anti-ranked edges @ acc
IOI	GPT-2	**58** @ 0.94	1484 @ 0.94	2041 @ 0.98 (full)
Agreement	Llama-3.2-1B	**263** @ 0.71	7644 @ 0.60	8993 @ 0.70 (full)
Induction	Gemma-3-1B	**3** @ 0.90	1197 @ 0.92	3537 @ 0.89 (full)
Docstring	Llama-3.2-1B	**519** @ 1.00	7128 @ 1.00	8993 @ 1.00 (full)

Target specificity across all four behaviors. After condensing one target, we run frozen discovery for every behavior in both the adapted and original models, using the same three seeds. The targeted behavior becomes $8\times$ easier to localize for induction, $4\times$ for agreement, and $2\times$ for docstring, alongside the IOI result in §5.5. The 12 off-target direction medians stay at $1.0\times$ across 288 measurements. Individual cells do move: 58 measurements exceed $1\times$, mostly on agreement, where the non-monotone accuracy curve makes budget selection less stable (§6).

The circuit is load-bearing and head-specific. Closing every edge leaves the adapter at chance, so it cannot carry the task alone. To test the selected heads, we re-point each retained head→logits edge to another head in the same layer. Accuracy falls to chance or below in 30 of 32 circuits (Table 12). Adding a random edge leaves accuracy unchanged, which ties the drop to the head substitutions.

The all-edges-off reference can fall below guessing because every edge then reads a corrupted activation. Against that stricter reference, 27 of 32 swapped circuits finish within 0.13. The two above-chance cases are IOI/Llama-3.2-1B (0.58) and agreement/Gemma-3-1B (0.605), both well below intact accuracy; these are two-choice tasks with chance level 0.50.

F.1 Healing-Only Control

Table 11 shows representative healing-only runs; the grid-wide median ratio is 0.99 (§5.5).

Table 11: **Healing-only control at head granularity.** Representative cells are shown, with EAP-IG participation ratios averaged over seeds 11/22/33. Higher PR means attribution is spread across more edges. The grid-wide median ratio is 0.99 across 32 cells. Runs start from fresh checkpoints rather than a C_1 endpoint.

Task	Model	PR before	PR after (no-prune heal)	Ratio
IOI	GPT-2	50.7	50.5	1.00
IOI	Llama-3.2-1B	331.0	328.3	0.99
IOI	Qwen3-4B	1017.7	876.1	0.86
Agreement	Llama-3.2-1B	267.9	261.7	0.98
Agreement	Llama-3.2-3B	479.3	475.2	0.99
Induction	Gemma-3-1B	47.1	46.7	0.99
Docstring	Qwen3-1.7B	97.5	97.0	0.99

F.2 Conduit and Head-Specificity Control

Table 12 lists the head-swap result for each cell (§5.5).

Table 12: **Sibling-swap control** (head granularity, all 32 cells). Values are three-seed medians on 200 examples. *Swapped* counts re-pointed head→logits edges. *All off* and *Kept* give accuracy with every edge closed and with the circuit intact. Chance is 0.50 for the two-choice tasks, 0.25 for docstring, and the reciprocal of the induction pool size. *To chance* marks swapped circuits no better than guessing.

Task	Model	Edges	Swapped	Kept	All off	Chance	Swapped acc.	To chance
IOI	Gemma-3-4B	4	2	1.000	0.555	0.50	0.550	yes
	Llama-3.2-3B	10	3	0.985	0.510	0.50	0.490	yes
	Gemma-3-1B	12	5	0.985	0.510	0.50	0.435	yes
	Llama-3.2-1B	18	7	0.980	0.515	0.50	0.580	no
	GPT-2	58	17	0.970	0.485	0.50	0.475	yes
	Qwen3-1.7B	94	23	0.975	0.510	0.50	0.560	yes
	Qwen3-4B	135	34	0.965	0.525	0.50	0.495	yes
	Qwen3-0.6B	169	34	0.975	0.520	0.50	0.515	yes
Agreement	Gemma-3-1B	125	4	0.815	0.225	0.50	0.605	no
	Gemma-3-4B	180	13	0.710	0.375	0.50	0.390	yes
	Qwen3-4B	190	18	0.735	0.190	0.50	0.400	yes
	Qwen3-1.7B	205	13	0.805	0.255	0.50	0.235	yes
	Llama-3.2-1B	263	31	0.700	0.290	0.50	0.525	yes
	Llama-3.2-3B	314	22	0.695	0.280	0.50	0.515	yes
	GPT-2	453	72	0.815	0.170	0.50	0.450	yes
	Qwen3-0.6B	1054	138	0.850	0.150	0.50	0.180	yes
Induction	Gemma-3-1B	3	1	0.925	0.000	0.05	0.000	yes
	Gemma-3-4B	4	2	0.925	0.000	0.05	0.000	yes
	Qwen3-4B	88	33	0.985	0.000	0.05	0.000	yes
	Llama-3.2-3B	91	25	0.985	0.000	0.05	0.000	yes
	Llama-3.2-1B	104	38	0.970	0.000	0.05	0.000	yes
	GPT-2	160	50	0.930	0.000	0.05	0.000	yes
	Qwen3-1.7B	228	57	1.000	0.000	0.05	0.000	yes
	Qwen3-0.6B	404	76	0.980	0.000	0.05	0.000	yes
Docstring	Gemma-3-1B	3	1	0.970	0.000	0.25	0.000	yes
	Gemma-3-4B	7	3	0.985	0.000	0.25	0.000	yes
	GPT-2	144	34	0.415	0.000	0.25	0.000	yes
	Qwen3-1.7B	475	80	0.995	0.000	0.25	0.000	yes
	Llama-3.2-1B	519	111	1.000	0.000	0.25	0.000	yes
	Qwen3-4B	608	121	1.000	0.000	0.25	0.000	yes
	Llama-3.2-3B	697	119	1.000	0.000	0.25	0.000	yes
	Qwen3-0.6B	1877	157	1.000	0.000	0.25	0.000	yes

F.3 Joint Gate and Adapter Training

The joint-training arm removes the prune-and-heal loop. It trains gates and the LoRA adapter together for 3000 steps under an L_0 sparsity penalty, using the same head-level graph, splits, restricted-choice metrics, and binary edge masks. Base-model weights stay frozen; the LoRA parameters train. We record the norm of the LoRA B matrices, initialized at zero, to confirm that the adapter changed.

We choose the smallest admissible configuration separately for each cell. It must have at least one edge, clear the *original* model's validation accuracy minus 0.05, and stay within the controller's 1.05 capability-probe ratio. We use the original-model bar because joint training can lower its own gates-open accuracy; a bar relative to that damaged model would be easier to clear. Sealed-test accuracy is reported after selection, not used to choose the configuration (Table 13).

Four pilot cells score 16 configurations each, covering two preservation weights and eight sparsity weights. The other 28 cells score five sparsity weights at the capability-preserving preservation weight; the remaining planned configurations were not run. This limited sweep may miss better joint-training circuits. Per-cell tuning favors this arm relative to C_1's fixed controller settings. The observed 5 of 32 wins describe the configurations tested here, not a bound on what a wider search could find.

Table 13 gives each cell's result and the number of admissible configurations among those scored. A zero means none of the tested configurations passed all three criteria; it does not rule out untested settings.

Table 13: **Joint gates and adapter across all 32 cells.** C_1/W3 above 1 means the baseline returns fewer edges; below 1, the controller does. Dashes mean no tested configuration passes the accuracy and capability criteria. The final column gives admissible configurations out of those scored.

Task	Model	W3 edges	C_1 edges	C_1/W3	Admissible configs
IOI	GPT-2	5	58	11.6×	2 of 16
	Llama-3.2-1B	22	18	0.8×	2 of 5
	Llama-3.2-3B	none	10	—	0 of 5
	Gemma-3-1B	3	12	4.0×	3 of 5
	Gemma-3-4B	3	4	1.3×	4 of 5
	Qwen3-0.6B	none	169	—	0 of 5
	Qwen3-1.7B	none	94	—	0 of 5
	Qwen3-4B	none	135	—	0 of 5
Agreement	GPT-2	none	453	—	0 of 5
	Llama-3.2-1B	21	263	12.5×	2 of 16
	Llama-3.2-3B	none	314	—	0 of 5
	Gemma-3-1B	674	125	0.2×	1 of 5
	Gemma-3-4B	none	180	—	0 of 5
	Qwen3-0.6B	none	1054	—	0 of 5
	Qwen3-1.7B	104	205	2.0×	3 of 5
	Qwen3-4B	633	190	0.3×	1 of 5
Induction	GPT-2	none	160	—	0 of 5
	Llama-3.2-1B	none	104	—	0 of 5
	Llama-3.2-3B	none	91	—	0 of 5
	Gemma-3-1B	213	3	0.0×	2 of 16
	Gemma-3-4B	167	4	0.0×	2 of 5
	Qwen3-0.6B	none	404	—	0 of 5
	Qwen3-1.7B	none	228	—	0 of 5
	Qwen3-4B	none	88	—	0 of 5
Docstring	GPT-2	none	144	—	0 of 5
	Llama-3.2-1B	none	519	—	0 of 16
	Llama-3.2-3B	none	697	—	0 of 5
	Gemma-3-1B	128	3	0.0×	3 of 5
	Gemma-3-4B	47	7	0.1×	3 of 5
	Qwen3-0.6B	none	1877	—	0 of 5
	Qwen3-1.7B	none	475	—	0 of 5
	Qwen3-4B	none	608	—	0 of 5

F.4 Seed Stability Detail

Table 14 compares circuit edges across seeds. Edge Jaccard divides the overlap by the union; containment asks what share of the smaller circuit lies in the larger. The null draws two random subsets of the observed sizes from the full graph. On checked cells, 2,000 simulated draws per pair agree with the analytic null within 0.001.

Table 14: **Seed stability of C_1 across all 32 cells** (head granularity, seeds 11/22/33). Values are per-task medians over seed pairs. $\times$null is the observed edge Jaccard divided by its expectation for independent random subsets of the same sizes. High containment and above-null Jaccard indicate consistent, often nested circuits across seeds.

Task	Containment	Edge Jaccard	$\times$ null
IOI	0.93	0.53	$205\times$
Agreement	0.89	0.68	$59\times$
Induction	0.95	0.83	$243\times$
Docstring	0.94	0.76	$45\times$

G Verification Detail and Analysis Cost

Table 15 times exact single-edge patching on IOI/GPT-2. It measures the saving from testing fewer edges after condensation and excludes the cost of producing the circuit.

Table 15: **Exact-patching cost on IOI/GPT-2.** Forward-pass and wall-clock reductions after restricting candidates to the condensed graph.

Candidate set	Condensed	Original	Passes	Time	Speedup
Head edges	70	256	1280→350	137.02s→37.66s	$3.6\times$

Measured on an earlier 70-edge IOI/GPT-2 endpoint; the reported endpoint is 58 edges (Table 20), so this understates the saving.

G.1 Total Compute and Break-Even

Each of the 96 head-level runs uses one 40GB or 80GB NVIDIA A100. Median runtime is 2.95 hours, ranging from 0.33 hours on GPT-2 to 17.5 on Qwen3-4B. Healing takes about 11% of the time; repeated EAP-IG ranking and gated evaluation take most of the rest. Exact single-edge patching on the condensed graph is $3.6\times$ faster per analysis (Table 15), so many repeated analyses are needed to recover the training cost.

The larger gain is in analyses that scale faster than linearly with edge count. Exhaustive subset and pair tests would require a median $55\times$ more ablations on matched frozen circuits (interquartile range 7–$759\times$). Only 15 of 32 frozen cells have a projected all-pairs runtime under one day (§5.3).

H Legibility and Original-Model Checks

We compare the condensed IOI circuit with the head roles mapped by Wang et al. (2023). Across seeds 11/22/33, it contains a median 24 heads (21–27), including 17 with published roles (17–19). It covers every stage of the published algorithm except the previous-token stage. The matched-accuracy frozen circuit contains 61 heads (61–82), with 25 named in all three seeds. The condensed circuit has a higher share of named heads but omits 9 of the 26 published heads, including three of eight backup name movers. The frozen circuit retains all but one of those eight.

H.1 Interchange Intervention Detail

Role counts identify heads by name. Interchange interventions (Geiger et al., 2021; 2022) ask whether changing the variable assigned to those heads changes the answer as Wang et al.'s algorithm predicts (Table 16).

Table 16: **Interchange intervention accuracy under Wang et al.'s IOI algorithm** (GPT-2, three seeds, 200 prompt pairs each). Every arm uses the same hypothesis and edge indices. We assign the output variable to named name movers and the intermediate variable to subject-inhibition heads feeding them. Higher accuracy means the intervention more often produces the predicted answer. "Random" uses unnamed heads; "Sham" tests a task without a published role map. The coarse all-in-circuit variant saturates and serves only as a control.

Aligned variable	C_1 condensed	Frozen (matched acc.)	Unmodified model
Output (name-mover → logits)	0.995	0.985	0.995
Intermediate (subject inhibition)	**0.930**	0.735	0.645
per-seed range	0.910–0.945	0.715–0.740	0.640–0.650
Random-edge control (intermediate)	0.03	0.04	0.01
Sham hypothesis (different task)	0.00	0.01	0.01

Patching the named name-mover outputs scarcely separates the arms: the condensed circuit, frozen circuit, and unmodified model score 0.995, 0.985, and 0.995. The subject-inhibition intervention is more discriminating (Table 16). Circuit size may still affect the comparison. Cutting the frozen circuit to C_1's edge count leaves it at chance in two of three seeds. In the third, it remains functional at 0.885 task accuracy, and C_1 scores 0.965 versus its 0.845 on the intervention.

The next three checks compare the circuit with the original model. They come from an earlier, coarser replication that treats an attention or MLP block as one component. The main results use head-level graphs, so these block-level values do not enter the headline statistics.

H.2 Block-Level Endpoint Reproduction

We reran 23 block-level cells from scratch with the same controller and seed. The median change in smallest-faithful edge count is zero; most endpoints reproduce exactly. One IOI/Qwen3-4B rerun moves from 18 to 67 edges at matched accuracy, showing that the stopping point can vary. This check is separate from the head-level, three-seed comparison in Appendix F.4.

H.3 Block-Level Logit-Difference Recovery

For comparison with EAP-style evaluations, Table 17 reports normalized logit-difference recovery for the 32 block-level cells and the head-level IOI anchor: $(m(\text{circuit}) - m(\text{ablated}))/(m(\text{full}) - m(\text{ablated}))$. Here m is the contrastive logit margin for IOI and agreement. The earlier block-level induction and docstring runs use the answer logit because their restricted pools contain more than one foil. Those values do not enter the current head-level headline. A value of 1 reproduces the full model's margin, while values above 1 mean interchange ablation sharpens it. Median recovery is 0.95 (range 0.53–1.25) on the sealed test split without retraining.

H.4 Original-Model Faithfulness

Because condensation changes weights, a circuit can fit the adapted model without reproducing the original model. The healing objective encourages agreement through $\mathrm{KL}(\text{orig} \,\|\, \text{masked})$, and the following checks measure how close the two remain.

The current head-level grid. Across all 96 current head-level runs, the circuit selects the same answer as the unmodified model on a median 96.5% of sealed-test examples, with a median token-level KL of 0.215. These are the grid-wide values reported in §5.4. On the IOI/GPT-2 anchor, the per-seed KL to the original model is 0.045–0.068 (median 0.056).

The block-level replication. Across the 32 block-level cells, the circuit matches the original model's task choice on a median 95.5% of sealed-test examples, with a median token-level KL of 0.196. Agreement has the lowest choice consistency at 0.74–0.95. Docstring/GPT-2 is also weak because its base model scores only 0.59. For comparison, the head-level IOI anchor has KL 0.056 to the original GPT-2, against 0.50 for its matched frozen circuit. Table 17 reports both sets of results. The block-level values are supporting checks rather than a grid-wide claim about the current head endpoints.

Table 17: **Original-model faithfulness and logit-difference recovery.** Results combine the block-level replication with the head-level IOI anchor. Match(M) is the fraction of sealed-test examples where the circuit and original model select the same candidate. KL($M\|\hat{C}$) and KL($M'\|\hat{C}$) compare the circuit with the original and adapted full models. LD recovery is $(m(\text{circuit}) - m(\text{ablated}))/(m(\text{full}) - m(\text{ablated}))$, where 1 reproduces the full model's margin. The head-level IOI row gives three-seed medians and has no LD value because that evaluation exists only at block granularity.

Task	Model	Edges	Acc.	Match(M)	KL($M\|\hat{C}$)	KL($M'\|\hat{C}$)	LD rec.
IOI	GPT-2	13	0.96	0.95	0.079	0.025	0.58
	Llama-3.2-1B	15	0.98	0.99	0.217	0.047	0.95
	Llama-3.2-3B	12	0.955	0.95	0.067	0.040	0.58
	Gemma-3-1B	4	0.995	0.99	0.118	0.046	1.23
	Gemma-3-4B	4	0.995	0.99	0.072	0.043	0.91
	Qwen3-0.6B	71	0.98	0.96	0.157	0.030	1.03
	Qwen3-1.7B	29	0.945	0.94	0.153	0.066	0.57
	Qwen3-4B	67	0.95	0.96	0.089	0.033	0.81
Agreement	GPT-2	227	0.82	0.95	0.050	0.053	0.98
	Llama-3.2-1B	12	0.79	0.76	0.324	0.226	1.18
	Llama-3.2-3B	37	0.64	0.78	0.340	0.255	0.88
	Gemma-3-1B	39	0.75	0.79	0.533	0.369	0.90
	Gemma-3-4B	82	0.555	0.74	0.361	0.202	0.65
	Qwen3-0.6B	377	0.86	0.91	0.242	0.251	0.97
	Qwen3-1.7B	116	0.79	0.80	0.327	0.218	1.08
	Qwen3-4B	98	0.665	0.78	0.357	0.305	1.06
Induction	GPT-2	65	0.965	0.96	0.169	0.097	0.86
	Llama-3.2-1B	52	0.94	0.94	0.159	0.045	1.02
	Llama-3.2-3B	53	0.97	0.96	0.185	0.064	1.17
	Gemma-3-1B	4	0.93	0.84	0.250	0.126	0.79
	Gemma-3-4B	4	0.95	0.93	0.359	0.113	0.98
	Qwen3-0.6B	193	0.965	0.95	0.255	0.064	0.94
	Qwen3-1.7B	161	0.99	0.99	0.291	0.048	0.98
	Qwen3-4B	91	0.985	0.98	0.310	0.073	1.25
Docstring	GPT-2	24	0.415	0.59	0.227	0.154	1.10
	Llama-3.2-1B	50	1.0	1.00	0.165	0.081	0.92
	Llama-3.2-3B	95	1.0	1.00	0.149	0.119	0.85
	Gemma-3-1B	4	1.0	1.00	0.157	0.079	0.83
	Gemma-3-4B	4	0.945	0.94	0.270	0.171	0.94
	Qwen3-0.6B	278	0.99	0.99	0.133	0.095	1.00
	Qwen3-1.7B	208	1.0	1.00	0.178	0.111	1.18
	Qwen3-4B	114	0.99	0.99	0.207	0.172	1.10
IOI (head)	GPT-2	58	0.950	0.965	0.056	0.025	—

I Error Prediction from Circuit Margins

Circuit-shift monitoring has used head-level signals to detect distribution shift in vision transformers (Peng et al., 2026). Here we test whether a circuit's restricted-choice margin ranks errors made by the original model. At matched edge count, median AUROC is 0.894 for C_1 and 0.645 for the frozen circuit; C_1 leads in 10 of 11 cells (Table 18). IOI has too few errors (0–2%), and docstring has three example-specific foils rather than one fixed contrast, so neither is scored.

At its full faithful size, the frozen circuit often scores higher but uses 1.3–500× more edges. On induction/Llama-3.2-1B, it reaches 0.907 versus 0.898 for C_1, using 600 rather than 104 edges. Agreement/GPT-2 is the one matched-size exception, with the frozen circuit ahead by 0.002.

Table 18: **Predicting original-model errors in 11 cells.** AUROC uses each circuit's restricted-choice margin to rank the original model's errors. Values are per-cell medians over three seeds. The C_1 and frozen columns use the same edge budget k; the final column uses the full frozen circuit. Agreement/Qwen3-0.6B ties at the shown precision, with C_1 ahead by less than 0.001. The original model's own margin defines the error labels and would score perfectly. IOI and docstring lack enough suitable errors or a fixed contrast, respectively.

Model	Err. rate	k	C_1 (matched)	Frozen (matched)	Frozen (full)
Agreement					
GPT-2	19%	453	0.983	**0.985**	0.992 (600e)
Llama-3.2-1B	35%	263	**0.943**	0.927	0.965 (1400e)
Llama-3.2-3B	34%	314	**0.936**	0.635	1.000 (20329e)
Gemma-3-1B	22%	125	**0.829**	0.639	0.931 (600e)
Gemma-3-4B	37%	180	**0.823**	0.675	0.981 (2000e)
Qwen3-0.6B	21%	1054	0.931	0.931	1.000 (13833e)
Qwen3-1.7B	24%	205	**0.877**	0.621	1.000 (13833e)
Qwen3-4B	26%	190	**0.894**	0.645	1.000 (43993e)
Induction					
Llama-3.2-1B	5%	104	**0.898**	0.690	0.907 (600e)
Gemma-3-1B	13%	3	**0.690**	0.586	0.807 (600e)
Gemma-3-4B	5%	4	**0.796**	0.333	0.834 (2000e)
Median (11 cells)			**0.894**	0.645	0.981

The margin also supports selective prediction: keeping the most-confident 70% of inputs raises accuracy from 0.781 to 0.911, versus 0.844 for the matched frozen circuit.

J Additional Payoff Detail

The remaining checks ask how much can be learned from the smaller graph at a fixed inspection budget and how its size and interactions vary across cells.

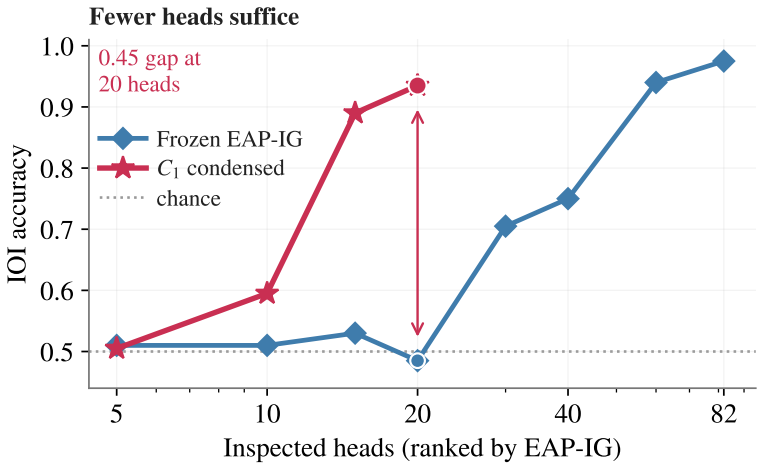


Figure 4: **Fixed-budget IOI legibility.** GPT-2 accuracy after retaining each arm's top EAP-IG-ranked heads, median over three seeds.

A smaller set of heads carries the behavior. With only the 20 highest-attribution IOI heads, the condensed circuit reaches 0.94 accuracy. The frozen circuit remains at chance and needs 60 heads

to match it (Figure 4). On induction, the same 66-head budget gives 0.955 for C_1 versus 0.60 for frozen discovery. Condensation does not increase the share of induction heads with identified roles; it makes fewer of them sufficient. The IOI interchange test is consistent with this result, although circuit size may affect that comparison (Appendix H.1).

Table 19: **Condensed and frozen IOI endpoints** (GPT-2, matched accuracy 0.950). Values are medians over seeds 11/22/33. Recall is the share of Wang's 26 heads in a circuit, and precision is the share of its heads with a published role. For both arms, KL is $\mathrm{KL}(\mathrm{orig} \,\|\, \mathrm{circuit})$ against the unmodified model. Since entries are per-seed medians, the reported F_1 need not equal the value computed from median precision and recall. The frozen circuit is re-matched for each seed, giving its 256–400 edge range.

	C_1	Frozen
Edges	**58**	256
Heads	**24**	61
KL to original	**0.056**	0.50
No published role	**7**	36
Precision	**0.71**	0.41
Recall	0.65	0.96
F_1	0.72	0.57

Circuit size does not track graph size in this grid. Qwen3-4B has 22 times as many candidate edges as GPT-2, yet its condensed endpoints are of similar size (Figure 5). A descriptive log–log fit has slope $+0.10$ (95% interval $[-0.21, +0.41]$), versus one for proportional growth. The pooled rank association is weak (Spearman $+0.08$, $p = 0.50$). We do not infer a scaling law from these eight models.

Table 20: **Condensed endpoints at head granularity, all 32 cells.** Median edges over seeds 11/22/33 with the seed range beneath, and median accuracy on the sealed 1000-example test split. "Full" is the complete head-level graph for that model.

		IOI		Agreement		Induction		Docstring	
Model	Full	Edges	Acc	Edges	Acc	Edges	Acc	Edges	Acc
GPT-2 small	2041	58 (54–71)	0.94	453 (317–595)	0.79	160 (148–160)	0.91	144 (118–175)	0.44
Llama-3.2-1B	8993	18 (17–27)	0.97	263 (205–279)	0.71	104 (104–119)	0.97	519 (480–519)	1.00
Llama-3.2-3B	20329	10 (8–267)	0.98	314 (254–316)	0.72	91 (77–97)	0.97	697 (545–720)	1.00
Gemma-3-1B	3537	12 (4–16)	0.99	125 (51–129)	0.83	3 (3–3)	0.90	3 (3–4)	0.97
Gemma-3-4B	10745	4 (3–6)	1.00	180 (176–226)	0.71	4 (4–6)	0.90	7 (6–7)	0.98
Qwen3-0.6B	13833	169 (163–253)	0.98	1054 (1054–1054)	0.84	404 (404–417)	0.94	1877 (1628–1976)	1.00
Qwen3-1.7B	13833	94 (65–114)	0.98	205 (178–221)	0.80	228 (226–240)	0.98	475 (391–559)	1.00
Qwen3-4B	43993	135 (119–317)	0.96	190 (169–242)	0.76	88 (82–104)	0.98	608 (608–608)	0.99

Within-family patterns differ by task. Qwen3 endpoints shrink with model size on agreement ($1{,}054 \to 205 \to 190$ edges) and induction ($404 \to 228 \to 88$, at accuracies 0.94, 0.98, 0.98), but not on IOI or docstring. Llama goes the other way on two tasks; agreement grows from 263 edges at 1B to 314 at 3B. The retained share falls from a median 7.5% in GPT-2 to 0.37% in Qwen3-4B, but the association across 32 task–model cells is not significant (Spearman -0.24, $p = 0.18$). Agreement accuracy also ranges from 0.71 to 0.84, so these sizes are only approximately performance-matched.

Why we report interacting partners per edge. Most strong pair effects are sub-additive: 87% satisfy $d_{ij} < d_i + d_j$. One edge can consume much of the available logit margin, leaving less effect for another to add. The fraction of pairs crossing a fixed 0.05 interaction threshold also falls sharply as circuit size grows (Spearman -0.89). We report interacting partners per edge rather than compare those fractions across differently sized circuits.

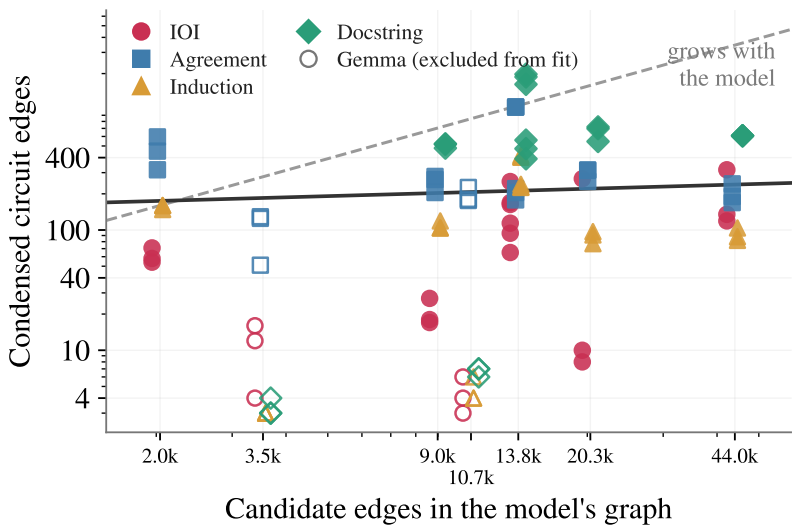


Figure 5: **Condensed circuit size against full graph size.** One point per run (93 total; low-accuracy docstring/GPT-2 is excluded). On log–log axes, the solid line fits the non-Gemma runs and the dashed line shows proportional growth. Hollow Gemma points are excluded from the fit because several endpoints have only 3–4 edges.

K Scope: Ablation Convention and Where the Result Is Weakest

We use interchange ablation for every circuit: a closed edge reads the activation from the corrupted prompt. Mean and zero ablation can change accuracy substantially. Under mean ablation, docstring circuits on Gemma-3-1B at $K = 64$ and Llama-3.2-3B at $K = 512$ score 0.00, while docstring/Llama-3.2-1B still scores 0.92 at $K = 512$. Matched-accuracy frozen circuits show similar sensitivity when they close a comparable fraction of edges. Frozen circuits that withstand other rules generally keep most of the graph. This pattern is consistent with sensitivity increasing as more edges are ablated; Miller et al. (2024) also report sensitivity to the intervention convention. Our size comparisons use the same convention in every arm. The reported edge count does not assert that the circuit works when all other activity is set to zero.

Table 21: **Verification checks by task** (§5.3). Values are medians over eight models and three seeds. *Accuracy after cutting* compares removal of all C_1 edges with an equal-size random removal. *Load-bearing alone* is the share of edges that matter individually. *No smaller subset* counts enumerable cells with no smaller faithful subset among all 2^k possibilities.

	Accuracy after cutting		Load-bearing alone		No smaller
Task	C_1	random edges	C_1	frozen	subset
IOI	0.52	0.98	15%	0%	2 of 7
Agreement	0.22	0.62	4%	0%	none enumerable
Induction	0.00	0.96	12%	0%	6 of 6
Docstring	0.00	0.97	3%	0%	3 of 6

K.1 Where the Result Is Weakest

The condensed IOI circuit omits 9 of the 26 heads mapped by Wang et al. (2023); it is sufficient for the task, but not a full reconstruction of their mechanism. Agreement accuracy ranges from 0.71 to 0.84, and docstring/GPT-2 reaches only 0.44; we exclude that cell from the payoff analyses in Sections 5.3 and 5.4. C_1 is also larger than frozen EAP-IG on docstring/Qwen3-0.6B and agreement/GPT-2. Several Gemma runs reach or approach the controller's edge floor, so we cannot tell how much further they might compress.

L Adapter Placement and Complexity

Edge counts exclude the LoRA adapter, so Table 22 reports its size for one IOI endpoint per family. Every run uses rank 16, $\alpha = 32$, and no dropout. GPT-2 adapts fused attention and MLP projections; the other families adapt seven attention and MLP projection matrices per layer. The adapter spans the model but adds only 0.9–1.9% as many parameters as the base model. Its updates are uneven: the top quarter of layers carries 38–61% of the squared Frobenius norm of BA, and the largest individual updates occur in MLP projections in every family. These measurements locate weight changes; the ablation and faithfulness tests assess what the retained edges carry.

Table 22: **LoRA accounting for one representative endpoint per family** (IOI, seed 22). "Top-quartile share" is the fraction of total squared update norm carried by the top quarter of layers.

Model	Adapted matrices	LoRA params	% of base	Top-quartile share
GPT-2 small	48	2.36M	1.90%	37.6%
Llama-3.2-1B	112	11.27M	0.91%	50.0%
Gemma-3-1B	182	13.05M	1.30%	47.9%
Qwen3-1.7B	196	17.43M	1.01%	61.3%

M Scope and Boundary Conditions

What circuit size measures. C_1 describes the adapted model; its outputs closely track, but do not exactly match, the original (§5.4). Edge counts use interchange ablation throughout and exclude computation within retained heads and MLPs and the LoRA adapter. Both condensed and comparably sparse frozen circuits can fail under other ablation rules (Miller et al., 2024) (Appendices K and L). Edge and node counts also differ: the per-task medians are 38 edges over 28 nodes on IOI, 98 over 58 on induction, 234 over 62 on agreement, and 497 over 150 on docstring. Even the 3-edge Gemma-3-1B endpoints span five nodes. On IOI, the circuit is sufficient for the behavior without recovering every part of the published mechanism (Appendix K.1).

How endpoints are selected. The tolerance, cut schedule, 500-step healing ceiling, and LoRA rank were fixed before the grid, so each returned circuit meets the same validation criteria. These settings do not guarantee the smallest faithful circuit: one 12-edge endpoint contains a faithful 3-edge subset, and we did not sweep the initial cut fraction or its decay. The ordering persists under a 90%-accuracy selection rule (Appendix D.1); shorter pruning routes fail at endpoints of 16 edges or fewer despite four times one round's healing budget (Table 2). Joint gate-and-adapter training returns admissible circuits in 12 of 32 cells, with 5 smaller than C_1; the controller returns a valid circuit across the grid without per-cell sparsity tuning (Appendix F.3).

When the capability constraint binds. On docstring, accuracy stays at 1.000 until the capability gate stops pruning; on agreement, it can improve as the circuit shrinks. Those endpoints are valid circuits, not measured global minima. The gate rarely rejects Gemma cuts because healing improves probe perplexity, compared with 15–20% of proposals in the other families (Appendix E.1). Excluding Gemma still leaves a $5.8\times$ average reduction, with C_1 smaller in 22 of 24 cells. We did not calibrate the gate separately by model, and the generic-text probe misses some nearby task damage: harder IOI prompts degrade in three of four families.

Cost and transfer. Condensation takes more computation than frozen discovery; the cost is most useful when the circuit is reused or enables an otherwise infeasible analysis (Appendix G.1).

Three seeds measure optimization stability: endpoints have median containment 89–95% per task and edge overlap $45\times$–$243\times$ above a matched-size null (Appendix F.4). We have not tested whether the same circuit appears under other checkpoints, input distributions, graph granularities, or pruning procedures; prior work reports sensitivity to these choices (Bali et al., 2026; Rai et al., 2026; Parekh, 2026).

N Containment in the Frozen Ranking

We rank the original model's edges with EAP-IG on training data, then select the smallest frozen circuit within 0.05 of its full-graph validation accuracy. We match edges by name across this circuit and C_1 and count how many C_1 edges it contains. For a random set of C_1's size, the expected overlap is $k_{\text{sel}}/|E|$: a frozen circuit holding one third of the graph would contain one third of a random edge set on average.

Median containment across 32 cells is 100%, versus 8% expected by chance; every cell exceeds its own chance level. In two cells, C_1 is larger than the frozen circuit, so full containment is impossible. At seed 11, agreement/GPT-2 has 595 versus 343 edges, and docstring/Qwen3-0.6B has 1,976 versus 798. Their containment scores are 0.57 and 0.39, still three to seven times chance. Even to include every C_1 edge, the frozen ranking extends no further than its top 13%.

The overlap shows that condensation uses edges ranked highly in the original model. It cannot establish that they perform the same computation after their weights change; the behavioral comparisons in §5.4 address that separate question.

O Post-Training Arms

Section 5.2 uses supervised fine-tuning as the post-training comparison. We also tried DPO, GRPO, and REINFORCE, but the short-answer tasks give the reinforcement arms little usable reward variation.

Each arm trains a LoRA adapter on the task using the same data splits and head-level graph as the other arms. Table 23 reports the EAP-IG participation ratio after training divided by its value before, so values below 1 indicate concentrated attribution. It also reports task accuracy and the general-probe perplexity ratio used by the controller's capability gate.

Table 23: **Post-training arms.** Ratios are after/before, with values below 1 indicating concentration. *Probe* is general-corpus perplexity relative to the original model, and the controller's limit is 1.05. DPO covers fewer cells because induction and docstring do not provide its required contrast answer.

Arm	Cells	PR ratio	Concentrates	Task acc.	Probe
SFT	32	0.76	$23/32$	$0.99 \to 1.00$	$1.37\times$
DPO (Rafailov et al., 2023)	16	0.51	$15/16$	$0.89 \to 1.00$	$12.9\times$
GRPO (Shao et al., 2024)	10	0.72	$10/10$	$0.76 \to 0.82$	$1.0\times$
REINFORCE (Williams, 1992)	10	1.88	$3/10$	$0.76 \to 0.49$	$6{,}336\times$

In 16 evaluated cells, the mean reward is 0.008 (at most 0.022): sampled answers rarely match the target. In the other four it averages 0.95, so almost every sample matches. Both regimes give little reward variation. Moreover, these rewards use exact match on generated text, while the paper's accuracy metric uses restricted choice. REINFORCE falls from 0.76 to 0.49 task accuracy and raises probe perplexity by $6{,}336\times$. GRPO stays inside the capability budget but improves accuracy only from 0.76 to 0.82 (Table 23).

These runs do not provide a matched reinforcement comparison. A different reward and evaluation objective would be needed before comparing RL post-training with condensation on these tasks.