REPRESENTATIONAL SIMPLICITY AND CIRCUIT SIZE DISSOCIATE IN A THRESHOLD-DEPENDENT WAY: A CONTROLLED TEST VIA ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse-autoencoder decomposability and concentrated feature attribution are increasingly treated as evidence that a model's computation is easier to reverse-engineer. Whether this representational and attributional cleanliness actually predicts a smaller or more tractable causal circuit remains an open question. We test this directly using adversarial training as a controlled instrument: it reliably reshapes internal representations, but this alone does not constitute a test of circuit size. We investigate this question through *reverse-engineering complexity*: the causal structure required to recover a model's behavior at a fixed level of faithfulness. To our knowledge, this is the first controlled empirical test of whether representational or attributional simplicity translates into causal simplicity at the circuit level. Starting from the same pretrained GPT-2 Small checkpoint, we apply matched standard and adversarial continual training, requiring both conditions to retain competence on indirect object identification and pass independent robustness verification before comparing mechanisms. We then compare the models along three complementary axes: sparse-autoencoder decomposability, SAE feature engagement in task attribution, and the size of faithful circuits recovered from the raw computational graph. The robust model is more SAE-decomposable and engages fewer SAE features in task attribution. Circuit size is regime-dependent: on competence-matched IOI, standard leads or ties below 85% faithfulness, but robust needs substantially fewer edges at high faithfulness (90%, 95%), a pattern established on the primary pair while representational trends generalize across a seven-point sweep and a second corpus.

1 INTRODUCTION

Mechanistic interpretability seeks to reverse-engineer the computations implemented by neural networks. Prior work has identified circuits for specific behaviors, including indirect object identification (IOI) and Greater-Than in GPT-2 Small (Wang et al., 2023; Hanna et al., 2023). Yet models that exhibit the same behavior need not implement it through mechanisms of comparable complexity. Their solutions may differ in how distributed, redundant, or causally entangled they are, and therefore in how difficult they are to recover.

Adversarial training provides a natural setting in which to study this variation. It can substantially alter the features and representations learned by a model (Tsipras et al., 2019; Engstrom et al., 2019), and has been associated with more interpretable gradients and saliency maps (Kim et al., 2019; Etmann et al., 2019). Recent work also links adversarial vulnerability to superposition and interference between overlapping features (Gorton & Lewis, 2025; Bereska et al., 2025; Stevinson et al., 2026). If robust training reduces this entanglement, it may make learned computations easier to isolate. Conversely, it may induce representational drift, redundancy, or compensatory pathways that make circuits harder to recover. Moreover, representational simplicity need not imply causal simplicity (Xu, 2026). Whether this representational simplicity translates into causal simplicity – a smaller or more tractable circuit – therefore remains an empirical question, and adversarial training offers a natural instrument to test it: a manipulation that reliably alters representations while task behavior can be held fixed.

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.

We study this question through *reverse-engineering complexity*: the causal structure required to recover a model's behavior at a fixed level of faithfulness. To our knowledge, this is the first controlled empirical test of whether representational or attributional simplicity translates into causal simplicity at the circuit level, rather than an inference from representational or attributional proxies alone. We define a reduction in reverse-engineering complexity as a smaller faithful circuit for a retained behavior, not as easier SAE reconstruction or more concentrated attribution. We test this with matched continual training from a shared checkpoint, competence and robustness qualification, and three complementary axes of simplicity, detailed in Section 2; this design separates changes in representation geometry from changes in causal organization and tests whether apparent simplicity is instead explained by redundant or compensatory computation. None of the literatures motivating this prediction test it at the circuit level.

Contributions.

- **Threshold choice determines whether representational simplicity predicts causal simplicity**: adversarial training increases SAE decomposability and reduces the number of active attribution features, though not their concentration once normalized by pool size; the faithful circuit recovered by edge attribution patching is not smaller at low-to-moderate faithfulness thresholds, but reaches high faithfulness with substantially fewer edges at high thresholds (90%, 95%).
- We corroborate this with α-ReQ, the decay rate of the residual-stream covariance spectrum, an SAE-free measure of representational simplicity: it rises with adversarial strength across a seven-point sweep, confirming that the representational trend is not an artifact of the SAE training procedure.

2 Experimental Design

2.1 Overview

We test whether adversarial training changes how readily a learned computation can be reverse engineered while controlling for initialization, data, training budget, and task performance. We continually train a standard and a family of adversarially trained GPT-2 Small models from the same pretrained checkpoint and on the same token stream. We first qualify candidate pairs by checking task retention and robustness under a multi-step attack. We then compare the selected standard and robust models along three progressively stronger axes: (i) how accurately matched sparse autoencoders (SAEs) reconstruct their internal activations, (ii) how many SAE features engage in task attribution, and (iii) how large a subgraph must be to reach a given faithfulness level. The first two axes measure decomposability and attribution structure; only the third directly measures reverse-engineering complexity.

Our primary matched-competence task is indirect object identification (IOI), for which all trained models retain near-ceiling performance. We additionally evaluate the Greater-Than (GT) task, which is not competence-matched between models.

2.2 Matched continual training

All models initialize from the official 124M-parameter GPT-2 Small checkpoint. Each model receives the same 1B-token stream in the same order and updates all parameters for 1,908 optimization steps. The primary comparison uses OpenWebText; Section 3.5 reports an identical standard-robust pair trained on FineWeb instead, to test whether the corpus drives the result. The standard arm minimizes next-token cross-entropy. The adversarial arms optimize

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{clean}} + \alpha\mathcal{L}_{\text{adv}}, \tag{1}$$

where $\mathcal{L}_{\text{adv}}$ is evaluated after a ten-step projected-gradient attack on token-embedding activations. Perturbations are applied after the token lookup and before positional embeddings. For every non-special token, the attack is constrained to an ℓ_2 ball with radius ϵ_{rel} times the mean non-special-token embedding norm in the micro-batch. We use random initialization, per-token normalized gradient ascent, and projection after every step; the special token is never perturbed. This continuous embedding-space construction follows the general setup of Xhonneux et al. (2024).

We sweep $\epsilon_{\text{rel}} \in \{0.05, 0.075, 0.10\}$ and $\alpha \in \{0.2, 0.5\}$, producing six adversarial candidates and one standard control. All runs use AdamW, a peak learning rate of 5×10^{-5} with 100 warmup steps and cosine decay, a global batch of 524,288 tokens, gradient clipping at 1.0, and bfloat16 computation. Appendix B gives the complete optimization and attack specification.

2.3 Tasks and qualification criteria

For IOI, we evaluate 1,000 fixed prompts balanced between ABBA and BABA templates (Wang et al., 2023). The behavioral score is the logit difference between the correct and incorrect names at the final position, and accuracy is the fraction of prompts with positive logit difference. The standard model must achieve at least 90% accuracy and each adversarial candidate at least 85%, floors that allow for ordinary continual-training drift while remaining far above the level indicating circuit absence (55.4% for a from-scratch control that failed to develop the behavior, Appendix B).

For robustness, we evaluate held-out OpenWebText under both a single-step attack and the ten-step attack used during training. The masking sanity check requires

$$|\mathcal{L}_{\text{PGD10}} - \mathcal{L}_{\text{single}}| \leq 0.10 \text{ nats}, \tag{2}$$

which guards against selecting a model that appears robust only under a weak attack (Athalye et al., 2018). Among candidates passing this check and the IOI floor, we rank robustness by attack susceptibility,

$$\Delta_{\text{adv}} = \mathcal{L}_{\text{PGD10}} - \mathcal{L}_{\text{clean}}, \tag{3}$$

where lower values indicate a smaller attack-induced loss increase. We select the qualified candidate with the lowest Δ_{adv}. The single-step versus PGD-10 gap is a masking diagnostic, not a robustness score; the non-robust standard control is therefore not required to pass it.

We also evaluate 1,000 GT prompts generated with the official YearDataset (Hanna et al., 2023). For a two-digit threshold YY, the probability-difference score is

$$\text{PD} = \sum_{y>YY} p(y) - \sum_{y \leq YY} p(y), \tag{4}$$

where the sums use the full-softmax probabilities of the 100 two-digit year tokens. We report paired differences from the standard model on identical prompts. GT is not a qualification gate because no established performance floor analogous to the IOI accuracy threshold exists for this task.

2.4 Axis 1: SAE decomposability

We train matched BatchTopK SAEs (Bussmann et al., 2024) on the residual stream entering transformer block 8, following standard practice of choosing a layer near the end of the network that contains diverse features without being specialized for final-layer output computation (Gao et al., 2025). Each SAE has 12,288 features, a 16-fold expansion over the 768-dimensional residual stream, and an average L_0 of 32 selected features per token. Training uses 300M post-training OpenWebText tokens disjoint from model training; evaluation uses a subsequent 10M-token partition. The standard and robust SAEs share the same architecture, sparsity target, data volume, optimizer, number of updates, and evaluation protocol. We train three seeds per condition for the principal reconstruction comparison.

The primary endpoint is mean squared reconstruction error in the original residual space on the held-out evaluation partition. We also report normalized-space training loss, fixed-threshold inference error, dead-feature counts, and the between-seed dispersion. The robust SAE's lower error at fixed dictionary size and sparsity indicates that its activation distribution is more compressible by this SAE family.

2.5 Axis 2: SAE feature engagement and selectivity

For each SAE (seed 42 primary; seeds 43 and 44 replicated), we attribute the IOI and GT scores to its features using node-level attribution patching (Syed et al., 2024; Marks et al., 2025). We splice the reconstruction $\hat{x}$ into the residual stream as

$$\tilde{x} = x + \hat{x} - \text{sg}(\hat{x}), \tag{5}$$

such that the forward pass is exactly the unmodified model while gradients flow through the SAE features. For feature i, target attribution is

$$A_i = \mathbb{E}_x \left[\sum_{t \in \mathcal{T}} \frac{\partial Y}{\partial f_{i,t}} \left(f_{i,t}^{\text{clean}} - f_{i,t}^{\text{corrupt}} \right) \right], \tag{6}$$

where Y is IOI logit difference or GT probability difference. We define the intervention position set $\mathcal{T}$ per task: $\mathcal{T}_{\text{IOI}} = \{S1, IO, S2, \text{END}\}$, the three name positions and the final position; $\mathcal{T}_{\text{GT}} = \{\text{start-year span}, \text{END}\}$, the token span stating the prompt's start year and the final position. Corrupted prompts swap the two IOI names or use the official GT counterfactual construction.

Our primary cross-model statistic is candidate pool size: the number of features active at task-relevant positions under an identical existence rule and identical prompts, measured before any ranking or attribution computation. As a secondary statistic, after sorting existing features by $|A_i|$, we report the smallest number required to account for 50%, 80%, and 90% of total absolute attribution; this concentration statistic is not scale-free with respect to pool size (Section 3.3), so we report it alongside pool size rather than as a stand-alone measure of task simplicity. As a complementary check on whether a feature's causal role is specific to the task or also shapes the model's broader output, we zero each pool feature on the same prompts and measure the shift in the full next-token distribution rather than the task readout alone (mean ablation-KL, Appendix F).

2.6 Axis 3: faithful circuit recovery

We recover task circuits with edge attribution patching (EAP-IG); a pilot of EAP-GP, a more recent variant designed to reduce gradient saturation, is reported in Appendix G (Syed et al., 2024; Zhang et al., 2025a). For each task and model, we define edge-native faithfulness at k retained edges as

$$F(k) = \frac{Y(\text{circuit}_k) - Y(\text{corrupt})}{Y(\text{clean}) - Y(\text{corrupt})}, \tag{7}$$

where circuit_k patches in the top-k edges by $|\text{score}|$ at their clean values and corrupt-patches every other edge to its value under the task's counterfactual distribution (name-swap for IOI, the official counterfactual construction for GT, Section 2.3); $Y(\text{clean})$ and $Y(\text{corrupt})$ are the task metric (IOI logit difference or GT probability difference) on the fully clean and fully corrupted graph, so $F = 0$ reproduces the fully corrupted baseline and $F = 1$ fully recovers clean performance. We distinguish two estimands on the same $F(k)$ curve. Partial-subgraph recoverability is the area under $F(k)$ across the full edge grid, including k where $F < 85\%$. Faithful-circuit size is the first grid k at which $F(k)$ reaches a target in the circuit regime $F \geq 85\%$; we report that crossing at several targets, not at a single canonical cutoff. The causal claim in this paper is about this estimand. AUC is reported as a descriptor of partial recovery, not as a circuit-size statistic. Conditional co-ablation tests whether apparently weak components become necessary after substitutes are removed, guarding against redundancy and self-repair (Gong et al., 2026; Hanna et al., 2024; Rushing & Nanda, 2024).

3 Results

3.1 Adversarial continual training yields a qualified robust comparison

All seven continually trained models retain IOI performance: accuracy ranges from 99.4% to 99.6%, compared with 99.7% for the original checkpoint. Four adversarial candidates also pass the masking sanity check. Among them, `ft-e100-a05` has the lowest attack susceptibility, reducing Δ_{adv} from 1.642 nats in the standard control to 0.362 while retaining 99.4% IOI accuracy. We therefore use `ft-clean` and `ft-e100-a05` as the principal standard-robust pair.

The near-constant IOI accuracy does not extend to GT. GT probability difference decreases smoothly as attack susceptibility falls across the sweep (Table 1). The selected robust model scores 0.442, compared with 0.671 for the standard model, with a paired prompt-level difference of -0.229 ± 0.010. By construction, PD lies in $[-1, 1]$, with values near zero indicating no informative signal; both models remain well above that floor, confirming the behavior persists substantially intact despite the competence gap between models on this task.

Table 1: Qualification sweep and secondary GT performance. The gap is $|\mathcal{L}_{\mathrm{PGD10}} - \mathcal{L}_{\mathrm{single}}|$; Δ_{adv} is the attack-induced loss increase, so lower is more robust. Asterisks mark adversarial candidates that pass both qualification gates.

Run	ϵ_{rel}	α	IOI	Gap	Δ_{adv}	GT PD
`ft-clean`	0	0	.996	.907	1.642	.671
`e050-a02`	.050	.2	.994	.153	.655	.609
`e050-a05`*	.050	.5	.995	.087	.494	.559
`e075-a02`	.075	.2	.994	.104	.556	.584
`e075-a05`*	.075	.5	.994	.057	.413	.500
`e100-a02`*	.100	.2	.995	.079	.495	.554
`e100-a05`*	**.100**	**.5**	**.994**	**.041**	**.362**	**.442**

3.2 Robust activations are more sparsely reconstructable

At equal dictionary size and sparsity, the SAE trained on robust-model activations reconstructs the held-out residual stream more accurately. Across three SAE seeds, original-space MSE is 1.1293 ± 0.0009 for the robust model and 1.1601 ± 0.0022 for the standard model, a 2.65% reduction with non-overlapping seed ranges. On position-matched evaluation blocks, the standard-minus-robust difference is 0.0323 ± 0.0007. The same ordering holds under fixed-threshold inference and in normalized activation space (Table 2), and at most one of the three seeds per condition has a single dead feature (Appendix E).

Table 2: Matched SAE reconstruction at the residual stream entering block 8. Lower is better. The original-space entry reports mean $\pm$ standard deviation across three SAE seeds; the remaining entries use seed 42.

Metric	Standard	Robust
Original-space MSE	1.1601 ± 0.0022	$\mathbf{1.1293 \pm 0.0009}$
Fixed-threshold MSE	1.1546	**1.1262**
Normalized-space loss	0.0877	**0.0768**
Dead features	0	0

3.3 Robust task attribution engages fewer SAE features, not more concentrated ones

Two effects distinguish the models, both consistent across all three SAE seeds and both tasks. First, under an identical existence rule and identical prompts, fewer robust-SAE features are active at task-relevant positions: 1,196 versus 1,612 candidates for IOI (a 26% reduction) and 360 versus 563 for GT (36%; Table 3). This measurement precedes any ranking or attribution computation.

Second, among active features, the robust SAE reaches each attribution-mass threshold with fewer of them in absolute terms: at 90% attribution mass, IOI requires 276 features versus 343, and GT requires 63 versus 90 (Table 3, Figure 1). Normalized by candidate pool size, this reverses: robust needs a larger share of its (smaller) pool to reach 90% mass on both tasks (IOI 23.1% versus 21.3%; GT 17.5% versus 16.0%), so the effect is that fewer features are active at all, not that attribution is more concentrated among them once pool size is accounted for. GT probability difference is lower under adversarial training (Table 1), so fewer GT features can reflect a weaker behavior rather than cleaner features. We read the GT counts as consistent with IOI, not as an independent matched comparison. Zeroing each pool feature and measuring the resulting shift in the model's general output distribution (mean ablation-KL) shows the same directional pattern, though it is not seed-robust and one of three IOI seeds reverses direction (Appendix F).

Table 3: Number of SAE features required to explain cumulative absolute target attribution. Reported values use the seed-42 SAE for each condition; the ordering holds at seeds 43 and 44 as well.

Model	Task	Candidates	50%	80%	90%
Standard	IOI	1,612	48	193	343
Robust	IOI	1,196	**43**	**163**	**276**
Standard	GT	563	14	54	90
Robust	GT	360	**8**	**33**	**63**

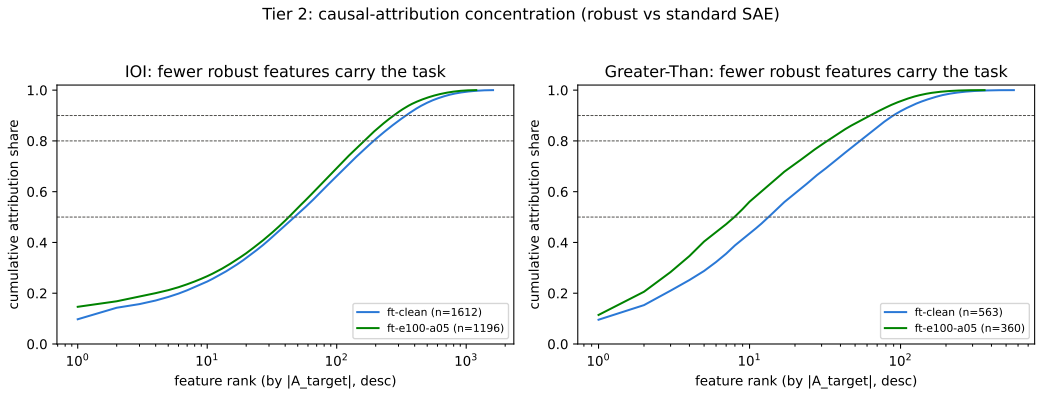


Figure 1: Cumulative attribution share by feature rank (descending $|A_i|$), IOI (left) and GT (right). The robust curve reaches each threshold with fewer features in absolute terms; normalized by candidate pool size this reverses (Section 3.3).

3.4 Circuit size on IOI depends on the faithfulness target

We recover faithful IOI circuits from the raw computational graph (156 nodes: 144 attention heads and 12 MLPs; 32,491 edges) using EAP-IG ($m = 5$ interpolation steps, Section 2.6), on 200 prompts (100 ABBA / 100 BABA, name-swap counterfactuals). Table 4 reports the first edge count reaching five faithfulness levels, rather than a single arbitrary threshold: at 70% standard leads; at 80–85% the models tie; at 90% robust needs far fewer edges (1,000 versus 4,000); at 95% the gap widens further (2,000 versus 16,000). Standard's curve plateaus after 85–90%, requiring a much larger circuit to close the remaining gap to near-complete faithfulness; robust's curve continues climbing steeply over the same range (Figure 2). Below 85% faithfulness, neither subgraph reproduces the behavior well enough to constitute a circuit; the comparison there is between two non-faithful partial subgraphs, not between circuit sizes (raw values in Appendix G, Table 12). At and above 85%, robust never needs more edges than standard, and needs substantially fewer at 90% and 95%.

Table 4: First edge count reaching each faithfulness threshold, IOI, EAP-IG. Both models are evaluated on the same geometric grid; a tie reflects grid resolution, not necessarily identical true crossing points. Left of the rule ($F < 85\%$): partial subgraphs, not faithful circuits. Right of the rule ($F \geq 85\%$): circuit-size comparison; 85% is a tie.

	70%	80%	85%	90%	95%
Standard	512	1,000	1,000	4,000	16,000
Robust	1,000	1,000	1,000	**1,000**	**2,000**

AUC (0.87 standard versus 0.60 robust) is the partial-subgraph recoverability estimand, as defined in Section 2.6; it does not speak to the edge count required to reach a faithful circuit at 90% or 95%. Conditional co-ablation, recovering backup components to account for self-repair at the node layer (Section 2.6), preserves the AUC ordering favoring standard at a much smaller scale (10–14 nodes, both models far below full faithfulness); it does not test the 90/95% edge-level crossings above.

Greater-Than moves the same direction at 95% (2,000 versus 8,000 edges). Because GT PD differs between models, we do not use that curve as a test of the main claim. Full detail, and the EAP-GP pilot, are in Appendix G.

3.5 Representational simplicity tracks robustness across the sweep

The qualification sweep (Section 3.1) produced seven continually trained models spanning $\epsilon_{\text{rel}} \in \{0.05, 0.075, 0.10\}$ and $\alpha \in \{0.2, 0.5\}$, not just the single selected robust model used above. We extend all three axes to all seven, and additionally measure α-ReQ, the power-law decay exponent of the residual-stream covariance spectrum (Appendix H), a representational-simplicity measure that does not depend on a trained SAE.

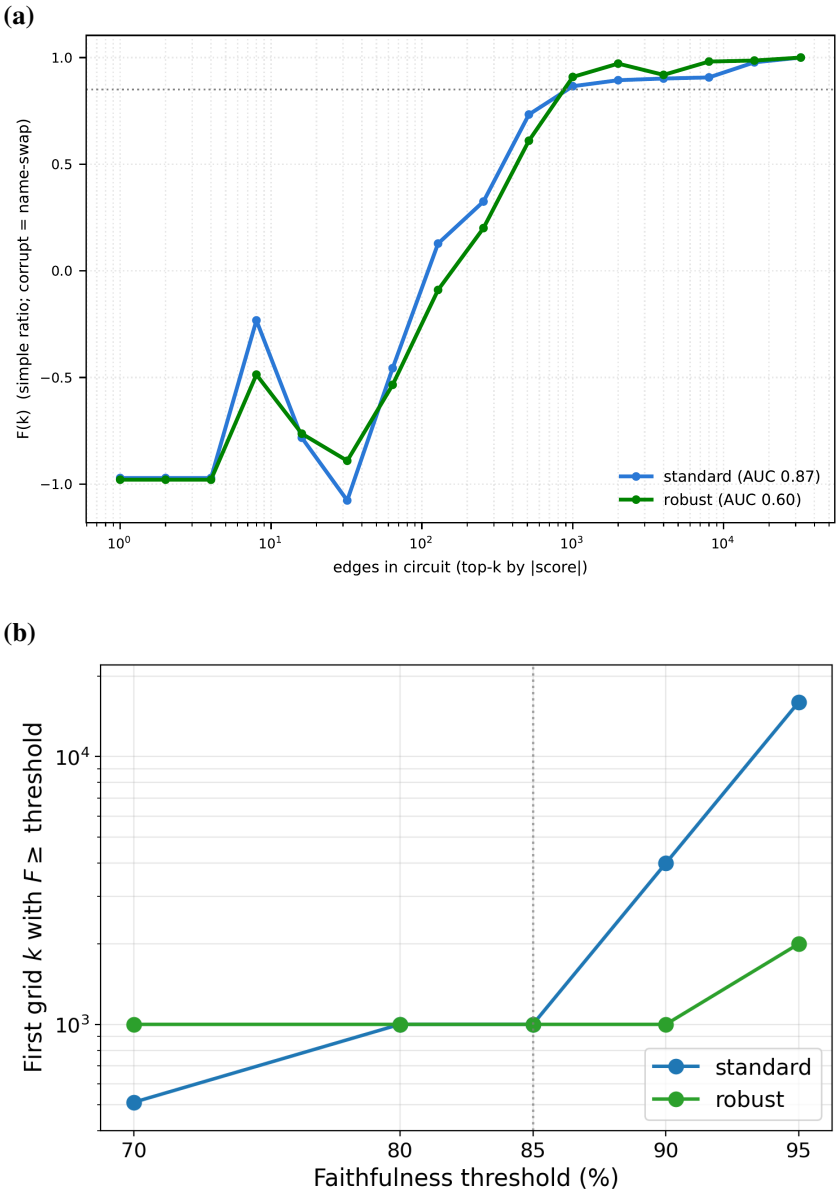


Figure 2: Two readings of the same IOI edge-native faithfulness curve, EAP-IG. **(a)** $F(k)$ against edges retained, log-k scale: standard and robust cross 85% faithfulness at the same edge count; standard leads by the largest margin in the steep rising region ($k \approx 64$–512); robust is locally ahead near $k = 32$ and across $k \approx 2000$–16000; net area under the curve favors standard (0.87 versus 0.60). **(b)** First grid point reaching each faithfulness threshold, evaluated on the geometric grid $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1000, 2000, 4000, 8000, 16000, 32491\}$; a reported crossing at k means the true crossing lies in $(k', k]$ for the preceding grid point k', not at k exactly. The dotted line marks 85%, the threshold used for the single-crossing comparison above; no threshold is a literature standard for this metric, which is why we report the full set of crossings rather than one value.

Across the sweep and FineWeb, SAE/α-ReQ simplicity increases with robustness while the 85% IOI crossing stays at or above the same grid point; 90/95% crossings were not computed for these models. The high-faithfulness gap in Section 3.4 is reported for the primary pair only.

Table 5: Representational and causal metrics across the full qualification sweep. Reconstruction MSE (inference-threshold mode) and IOI candidates are three-seed means (SAE training seeds 42/43/44); α-ReQ, participation ratio, and edges-to-85% are single deterministic measurements, since neither the α-ReQ probe (SAE-free, computed directly from the activation covariance) nor circuit recovery (run on the raw graph) depends on SAE training stochasticity. The edges-to-85% column is the single 85% gate used for qualification. Representational metrics move in the predicted direction as adversarial strength increases, with occasional non-monotonic steps between adjacent configurations of similar measured robustness (Table 1); circuit size does not track them at this threshold.

Model	ϵ_{rel}	α	Recon. MSE	IOI cand.	α-ReQ (model)	α-ReQ (SAE)	Part. ratio	Edges→85%
`ft-clean`	0	0	1.1528	1600.7	0.9230	0.8074	3.815	1,000
`e050-a02`	.050	.2	1.1315	1483	0.9465	0.8177	3.689	**2,000**
`e075-a02`	.075	.2	1.1277	1408	0.9546	0.8213	3.695	1,000
`e100-a02`	.100	.2	1.1242	1371	0.9609	0.8239	3.681	1,000
`e050-a05`	.050	.5	1.1432	1370	0.9644	0.8267	3.782	1,000
`e075-a05`	.075	.5	1.1301	1266	0.9762	0.8325	3.732	1,000
`e100-a05`	.100	.5	1.1260	1182	0.9865	0.8376	3.716	1,000

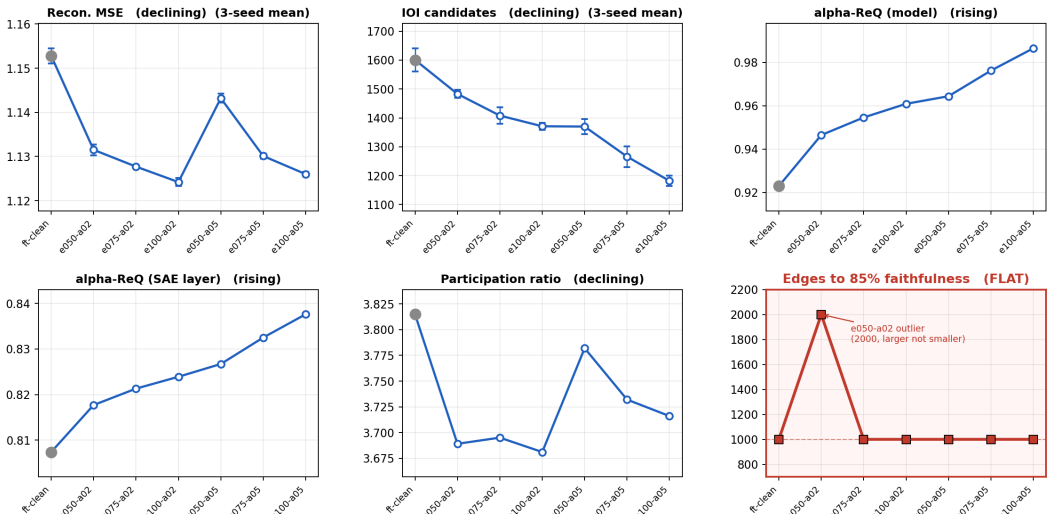


Figure 3: Representational metrics (panels 1–5) move in the predicted direction with adversarial strength across the seven-model sweep, ordered by measured Δ_{adv} (1.642 → 0.362); two adjacent configurations with near-identical measured robustness (`e100-a02`, `e050-a05`) produce small non-monotonic steps in reconstruction MSE and participation ratio. Panel 6 shows the single 85% qualification gate, flat at 1,000 edges except for one outlier (`e050-a02`, 2,000 edges – larger, not smaller). Reconstruction MSE (inference-threshold mode) and IOI candidates are three-seed means with ± 1 SD error bars (seeds 42/43/44); α-ReQ, participation ratio, and edges-to-85% are single deterministic measurements per model (Appendix H).

We separately test whether the representational trend depends on OpenWebText specifically, using one additional standard-robust pair continually trained identically to the primary comparison except for the pretraining corpus (FineWeb in place of OpenWebText; Appendix I). Every representational metric moves in the same direction as the primary result, and the 85% edge count is again flat (Table 6).

4 Conclusion

In this paper, we test whether adversarial training reduces reverse-engineering complexity, using it as a controlled instrument that reshapes representations while task competence is held fixed. To this end, we compare matched standard and adversarially trained GPT-2 Small models along three complementary axes: SAE decomposability, SAE feature engagement, and the size of faithful circuits recovered from the raw computational graph.

Table 6: Corpus ablation (inference-threshold mode MSE): standard vs. robust GPT-2 Small continually trained on FineWeb in place of OpenWebText, otherwise identical to the primary comparison.

Metric	Standard (`fw-clean`)	Robust (`fw-robust`)
Recon. MSE	1.089	**1.077**
IOI candidates	1549	**1126**
α-ReQ (model)	0.9288	**0.9943**
α-ReQ (SAE)	0.8121	**0.8434**
Edges$\rightarrow$85%	1,000	1,000

Our results reveal three key findings: (1) the robust model is more SAE-decomposable and engages fewer active SAE features, though not more concentrated ones once normalized by pool size, corroborated by α-ReQ, an SAE-free measure that rules out an SAE-specific artifact; (2) faithful circuit size is regime-dependent – standard leads or ties below 85% faithfulness, but robust needs substantially fewer edges at high faithfulness (90%, 95%); and (3) the representational trend generalizes across a seven-point adversarial-strength sweep and a second corpus, while the causal result is established only on the primary pair, making this a starting point for testing generalization across scale, architecture, and robustification method. Whether this regime-dependent pattern generalizes beyond the primary pair remains open; we do not draw broader claims about auditability or deployment from a single model pair on one task.

5 Related Work

Adversarial examples can exploit predictive but non-robust features, and adversarial training changes which features a model uses (Tsipras et al., 2019; Ilyas et al., 2019), with robustness linked to more human-aligned gradients, saliency maps, and representations (Ross & Doshi-Velez, 2018; Etmann et al., 2019; Engstrom et al., 2019; Wang et al., 2022); these results concern input sensitivity or representational geometry, not causal mechanism size. A closer precedent comes from vision: adversarially trained networks admit sparser attribution vectors (Chalasani et al., 2020) and exhibit feature purification (Allen-Zhu & Li, 2020), with gradients and saliency maps aligning more closely with human-interpretable structure (Kim et al., 2019) – predicting simpler features and sparser attributions, not a smaller causal graph, the prediction we test.

Superposition offers a sharper hypothesis, that models represent more features than dimensions by accepting interference (Elhage et al., 2022), recently linked to adversarial vulnerability (Gorton & Lewis, 2025; Stevinson et al., 2026; Gao et al., 2026; Zhang et al., 2025b; Bereska et al., 2025); SAEs measure this decomposition at scale (Huben et al., 2024; Gao et al., 2025), though reconstruction and sparsity alone do not guarantee causal features (Karvonen et al., 2025; Paulo & Belrose, 2026; Li et al., 2026), motivating our separate evaluation of decomposability and causal structure. On the causal side, manual and automated methods have identified and scaled circuit recovery for GPT-2 Small behaviors (Wang et al., 2023; Hanna et al., 2023; Conmy et al., 2023; Syed et al., 2024; Hanna et al., 2024), though recovered circuits remain conditional on the attribution method, threshold, and pruning procedure (Hanna et al., 2024), and ablations can trigger self-repair that understates causal dependence (McGrath et al., 2023; Rushing & Nanda, 2024).

The closest empirical studies apply mechanistic interpretability to adversarial vulnerability: García-Carrasco et al. (2024) localize components affected by adversarial examples in a fixed GPT-2 Small circuit, and Walkowiak et al. (2025) combine adversarial evaluation with EAP-IG circuits across languages; Bereska (2024) lists reverse-engineering robust models in a self-published proposal, not executed empirically. To our knowledge, we are the first to compare matched standard and adversarially trained language models through SAE decomposability, SAE feature engagement, and the size of established task circuits at fixed faithfulness. Appendix A discusses SAE evaluation, circuit-discovery variants, and self-repair in greater detail.

AI Use Disclosure

We used Claude (Anthropic) to assist with manuscript preparation.

Writing and polishing. Claude assisted with proofreading, polishing, and structuring the paper's sections.

Citations. Claude assisted with verifying citations against primary sources and identifying missing bibliography entries.

LaTeX and formatting. Claude assisted with LaTeX formatting, table and figure layout, and troubleshooting page-length and float-placement issues.

All experimental design, model training, data analysis, and interpretation of results were carried out by the authors. All claims, figures, and interpretations were reviewed and approved by the authors, who take full responsibility for the submission's content and accuracy.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning, 2020. URL https://arxiv.org/abs/2005.10190.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 2018. URL https://proceedings.mlr.press/v80/athalye18a.html.

Leonard Bereska. Mechanistic interpretability for adversarial robustness: A proposal. *Self-published*, August 2024. URL https://leonardbereska.github.io/blog/2024/mechrobustproposal/.

Leonard F. Bereska, Zoe Tzifa-Kratira, Reza Samavi, and Stratis Gavves. Superposition as lossy compression: Measure with sparse autoencoders and connect to adversarial vulnerability. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=qaNP6o5qvJ.

Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. Proper network interpretability helps adversarial robustness in classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1014–1023. PMLR, 2020. URL https://proceedings.mlr.press/v119/boopathy20a.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=d4dpOCqybL.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1383–1391. PMLR, 2020. URL https://proceedings.mlr.press/v119/chalasani20a.html.

David Chanin. Are sparse autoencoder benchmarks reliable?, 2026. URL https://arxiv.org/abs/2605.18229.

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations, 2019. URL https://arxiv.org/abs/1906.00945.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1823–1832. PMLR, 2019. URL https://proceedings.mlr.press/v97/etmann19a.html.

Jingchao Gao, Ziqing Lu, Raghu Mudumbai, Xiaodong Wu, Jirong Yi, Myung Cho, Catherine Xu, Hui Xie, and Weiyu Xu. Feature compression is the root cause of adversarial fragility in neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cbcce87f745072c819204529be843d16-Abstract-Conference.html.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Jorge García-Carrasco, Alejandro Maté, and Juan Trujillo. Detecting and understanding vulnerabilities in language models via mechanistic interpretability. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 385–393. International Joint Conferences on Artificial Intelligence Organization, 2024. doi: 10.24963/ijcai.2024/43. URL https://www.ijcai.org/proceedings/2024/0043.

Zhiren Gong, Zihao Zeng, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits, 2026. URL https://arxiv.org/abs/2607.01940.

Liv Gorton and Owen Lewis. Adversarial examples are not bugs, they are superposition, 2025. URL https://arxiv.org/abs/2508.17456.

Itamar Hadad, Guy Katz, and Shahaf Bassan. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b5afe13494c825089b1e3944fdaba212-Abstract-Conference.html.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=p4PckNQR8k.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=grXgesr5dT.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on*

Machine Learning, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264. PMLR, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Beomsu Kim, Junghoon Seo, and Taegyun Jeon. Bridging adversarial robustness and gradient interpretability. In *ICLR Workshop on Safe Machine Learning: Specification, Robustness, and Assurance*, 2019. URL https://arxiv.org/abs/1903.11626.

Aaron J. Li, Suraj Srinivas, Usha Bhalla, and Himabindu Lakkaraju. Evaluating adversarial robustness of concept representations in sparse autoencoders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5940–5957. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.279. URL https://aclanthology.org/2026.eacl-long.279/.

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

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

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

Gouki Minegishi, Hiroki Furuta, Yusuke Iwasawa, and Yutaka Matsuo. Rethinking evaluation of sparse autoencoders through the representation of polysemous words. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/380a0b16a7e6f8c5010f798c9f2d3c61-Abstract-Conference.html.

Adam Noack, Isaac Ahern, Dejing Dou, and Boyang Li. An empirical study on the relation between network interpretability and adversarial robustness. *SN Computer Science*, 2 (1):32, 2021. doi: 10.1007/s42979-020-00390-x. URL https://doi.org/10.1007/s42979-020-00390-x.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c1fe56b043848b211030c202764c6a7-Abstract-Conference.html.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11504. URL https://ojs.aaai.org/index.php/AAAI/article/view/11504.

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

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang. Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2172–2210. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.99. URL https://aclanthology.org/2026.acl-long.99/.

Edward Stevinson, Lucas Prieto, Melih Barsbey, and Tolga Birdal. Adversarial vulnerability from interference between features in superposition. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=fD6c4bD7sR.

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

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*, volume 37, 2024. doi: 10.52202/079017-1287. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/47c7edadfee365b394b2a3bd416048da-Abstract-Conference.html.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *The Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Paweł Walkowiak, Marek Klonowski, Marcin Oleksy, and Arkadiusz Janz. Unpacking robustness in inflectional languages: Adversarial evaluation and mechanistic insights, 2025. URL https://arxiv.org/abs/2505.07856.

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

Zifan Wang, Matt Fredrikson, and Anupam Datta. Robust models are more interpretable because attributions look normal. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22625–22651. PMLR, 2022. URL https://proceedings.mlr.press/v162/wang22e.html.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. Efficient adversarial training in LLMs with continuous attacks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0302fb83c62991efbccf0a003e4f5a92-Abstract-Conference.html.

Yongzhong Xu. Pattern selectivity is not task-causal structure, 2026. URL https://arxiv.org/abs/2606.05378.

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*, volume 38, 2025a. doi: 10.52202/085713-4748. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d029c97ee0db162c60f2ebc9cb93387e-Abstract-Conference.html.

Qi Zhang, Yifei Wang, Jingyi Cui, Xiang Pan, Qi Lei, Stefanie Jegelka, and Yisen Wang. Beyond interpretability: The gains of feature monosemanticity on model robustness. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/11822e84689e631615199db3b75cd0e4-Abstract-Conference.html.

A Extended Related Work and Methodological Context

Conventional interpretations of robust models. The connection between robustness and interpretability has most often been studied through gradients and visual representations. Input-gradient regularization can improve both resistance to transferred adversarial examples and the human legibility of gradients (Ross & Doshi-Velez, 2018). Robustness has also been connected to the alignment of saliency maps with perceptually relevant directions (Etmann et al., 2019; Kim et al., 2019), while robust image classifiers learn representations and attributions that are often more human-aligned (Engstrom et al., 2019; Wang et al., 2022). Conversely, training for robust interpretation can support robust classification (Boopathy et al., 2020; Noack et al., 2021). These findings provide evidence for a relationship between robustness and conventional interpretability, but each operationalizes interpretability through input-output sensitivity, attribution, or representation quality rather than through the causal organization of a learned computation.

Evaluating sparse autoencoders. SAEs can recover fine-grained language-model features and scale to large overcomplete dictionaries (Huben et al., 2024; Gao et al., 2025). Their evaluation nevertheless remains unsettled. SAEBench finds that improvements on reconstruction and sparsity proxies do not reliably transfer to practical interpretability tasks (Karvonen et al., 2025), while semantic evaluations based on polysemous words expose failures hidden by the usual reconstruction-sparsity frontier (Minegishi et al., 2025). Independent SAEs trained on the same model and data can learn substantially different dictionaries (Paulo & Belrose, 2026), motivating explicit measurement of feature consistency across runs (Song et al., 2026). A recent preprint further reports high noise or weak discriminability in several benchmark metrics (Chanin, 2026). SAE representations can themselves be adversarially fragile: small input perturbations can manipulate concept interpretations without materially changing the underlying language-model activations (Li et al., 2026). These limitations do not make SAEs uninformative, but they prevent reconstruction, sparsity, or automated feature labels from serving as stand-alone evidence of simpler computation.

Automated circuit discovery. Automated Circuit Discovery (ACDC) recursively removes edges using activation-patching effects and recovers compact subgraphs for established tasks (Conmy et al., 2023). Edge Attribution Patching (EAP) replaces repeated interventions with a gradient-based approximation, reducing circuit discovery to two forward passes and one backward pass (Syed et al., 2024). EAP with integrated gradients improves faithfulness relative to local-gradient EAP and demonstrates that overlap with a reference circuit can remain high even when the recovered circuit is behaviorally unfaithful (Hanna et al., 2024). EAP-GP modifies the integration path to mitigate saturation and further improve gradient-based circuit identification (Zhang et al., 2025a). Sparse feature circuits extend causal graph discovery from attention heads and MLPs to SAE features (Marks et al., 2025), while recent formal work defines circuit guarantees relative to explicit input domains and intervention semantics (Hadad et al., 2026). Together, these methods make circuit recovery scalable, but their outputs remain conditional on the graph granularity, patching baseline, task metric, attribution estimator, threshold, and search procedure. Our circuit-size comparison therefore concerns the smallest graph recovered by a fixed algorithm at a specified faithfulness threshold, not a global minimum over every possible circuit.

Circuit stability, selectivity, and self-repair. Circuit analyses can identify similar functional algorithms across training checkpoints and model scales even when individual components change (Tigges et al., 2024). Conversely, attention-pattern selectivity need not identify the components that causally implement a task (Xu, 2026). These results reinforce the distinction between observational structure and causal dependence. Causal interventions introduce their own complication: ablating a component can change the behavior of downstream components, producing Hydra-like compensation (McGrath et al., 2023). Self-repair occurs across language-model families and scales and can arise through several mechanisms, including LayerNorm scaling and reduced erasure of upstream contributions (Rushing & Nanda, 2024). Copy suppression provides a detailed example in which an attention-head motif explains both a component's normal function and part of the compensation observed after ablation (McDougall et al., 2024). Because one-at-a-time ablations can therefore underestimate distributed dependence, our causal analysis includes conditional co-ablation to test whether apparently weak components become important when potential substitutes are simultaneously removed.

Relation to adversarial mechanistic interpretability. García-Carrasco et al. (2024) propose a pipeline that first identifies a task circuit in a fixed GPT-2 Small model, then generates adversarial samples and localizes the affected components. Walkowiak et al. (2025) combine adversarial attacks with EAP-IG circuits to compare language and inflectional settings. The feature-level literature instead studies whether superposition or representation compression creates adversarial vulnerability (Gorton & Lewis, 2025; Bereska et al., 2025; Stevinson et al., 2026; Gao et al., 2026). Finally, Bereska (2024) lists reverse-engineering robust models as an explicit research objective in a self-published proposal, alongside investigating feature superposition and designing training methods that improve both robustness and interpretability; the proposal is not executed empirically. Our work carries out that objective: we experimentally compare matched standard and adversarially trained language models across three non-equivalent levels: sparse feature decomposition, SAE feature engagement, and fixed-faithfulness circuit recovery.

B Continual-training and attack details

B.1 Initialization and optimization

Every run initializes from `openai-community/gpt2`, revision `607a30d783dfa663caf39e06633721c8d4cfcd7e`. The checkpoint has 124M parameters, a 50,257-token vocabulary, and learned affine biases. We update all parameters rather than using parameter-efficient adapters. The training stream is drawn from the Hugging Face parquet conversion of OpenWebText at revision `433fe0f44ed7894fea29c08b3202aa348ccc6369`. The standard and adversarial runs receive the same examples in the same order.

Table 7: Continual-training hyperparameters shared by all seven runs.

Hyperparameter	Value
Training budget	exactly 1,000,000,000 tokens
Updates	1,907 full updates and one final partial update
Global batch	524,288 tokens per update
Micro-batch	16 sequences, with gradient accumulation
Optimizer	AdamW
Adam coefficients	$(\beta_1, \beta_2) = (0.9, 0.95)$
Peak learning rate	5×10^{-5}
Schedule	100-step warmup, cosine decay to 5×10^{-6}
Weight decay	0.1
Gradient clipping	global norm 1.0
Precision	bfloat16 autocast
Checkpoint interval	every 250 updates

The attack operates on the output of the token-embedding lookup before positional embeddings are added. In each micro-batch, its absolute radius is

$$\epsilon_{\text{abs}} = \epsilon_{\text{rel}} \frac{1}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} \|E(x_t)\|_2, \tag{8}$$

where $\mathcal{I}$ contains non-special-token positions. Each token perturbation is initialized at a random point inside its ℓ_2 ball. We then perform ten normalized-gradient ascent steps of size $2.5\epsilon_{\text{abs}}/10$, projecting every token independently after each step. Token 50256 is excluded from perturbation. Training and qualification use the same attack implementation and parameterization.

B.2 Why continual training is used

We initially attempted a matched comparison trained from scratch on 9.9B FineWeb tokens. The standard model reached only 55.4% IOI accuracy, compared with 99.7% for the official GPT-2 checkpoint on the same 1,000 prompts. Thus, the target behavior did not reliably emerge within the available from-scratch budget. A one-step embedding-adversarial variant also showed catastrophic overfitting: its loss was 1.20 under its training-time attack but 24.70 under PGD-10, a 23.5-nat discrepancy. These failures motivated continual training from a checkpoint that already possesses the target computations and the use of multi-step PGD during both training and verification.

C EVALUATION DATA AND PROVENANCE

Table 8 summarizes the OpenWebText partitions. Model-training data, robustness evaluation, SAE training, SAE evaluation, and the generic-language-modeling attribution control are positionally separated in the token stream. Because this OpenWebText conversion has no native document identifiers, we synthesize identifiers from parquet file and row indices and record hashes of the resulting lists. One document crosses each of two token-level partition boundaries; the token streams themselves remain disjoint.

Table 8: OpenWebText partitions used by the study.

Partition	Tokens	Provenance note
Continual training	1,000,000,000	973,424 logged document IDs
Robustness evaluation	100,000,000	88,637 reconstructed IDs
SAE training	300,000,000	post-training stream
SAE evaluation	10,000,000	subsequent held-out stream
Attribution control	256,000	1,000 sequences of length 256

The IOI set contains 1,000 prompts, split evenly between ABBA and BABA templates, generated with the official IOIDataset implementation at seed 42. Names are restricted to single GPT-2 tokens. The GT set contains 1,000 balanced prompts generated with the official YearDataset implementation at seed 42. To construct GT counterfactuals, we use the official "01-dataset" bad-sentence branch and require matching start-year token counts; 973 prompts pass this parity filter. The unused bad-sentence branch of the original generator was patched to avoid a batching crash when a sampled counterfactual year is represented by one BPE token. The sampling procedure, templates, and random-number stream are otherwise unchanged.

D QUALIFICATION SWEEP AND BEHAVIORAL TRADE-OFF

Qualification metrics are computed for every sweep cell rather than stopping after the first passing candidate. All adversarial runs clear the IOI retention floor, while four satisfy the masking sanity check. The standard model's large single-step versus PGD-10 gap is reported as a non-robust reference and is not interpreted as a failed adversarial-training run.

The GT results reveal a graded behavioral trade-off that IOI accuracy does not expose. Paired prompt-level differences from the standard model are -0.063 ± 0.005, -0.087 ± 0.006, -0.117 ± 0.007, -0.112 ± 0.006, -0.171 ± 0.008, and -0.229 ± 0.010 for `e050-a02`, `e075-a02`, `e100-a02`, `e050-a05`, `e075-a05`, and `e100-a05`, respectively. The original checkpoint scores 0.706 on this prompt set, the continually trained standard model scores 0.671, and the selected robust model scores 0.442. Because prompt sampling differs from the setup used to obtain previously published headline values, the paired within-set difference is the primary GT comparison.

E SPARSE-AUTOENCODER IMPLEMENTATION AND STATISTICS

We cache block-8 residual activations once per model as bfloat16 shards, recording a hash of the corresponding model weights. Each SAE is trained for 73,242 updates, one pass over 300M tokens, with batches of 4,096 tokens and Adam at learning rate 4×10^{-4} and $(\beta_1, \beta_2) = (0.9, 0.99)$. Training uses a roughly 500,000-token activation buffer, cross-shard shuffling, and refill when half the buffer has been consumed. No early stopping is used.

Inputs are normalized per token by $s = \sqrt{d}/\|x\|_2$ before encoding and returned to the original scale after decoding. BatchTopK selects 32 times the number of batch tokens activations across the batch. Decoder columns are constrained to unit norm by radially projecting their gradients before each update and renormalizing after it. AuxK uses 512 features, coefficient $1/32$, and defines a feature as dead if it has not been selected for 10M tokens. At most one dead feature per seed (Table 9).

For downstream inference, batch-relative selection is replaced by a fixed threshold. The threshold is the mean, across held-out evaluation batches, of the smallest positive activation selected by Batch-

TopK. At seed 42, the thresholds are 0.785 for the standard SAE and 0.741 for the robust SAE. This prevents the activation state of one prompt from depending on other prompts in the evaluation batch.

We quantify measurement uncertainty by pairing residual positions between models and clustering observations into 1,024-token sequence blocks. Across 9,756 blocks, standard-minus-robust reconstruction MSE is 0.0323 ± 0.0007 in original space and 0.0112 ± 0.0001 in normalized space. We separately retrain each SAE with seeds 42, 43, and 44. The original-space MSE ranges do not overlap: 1.1601 ± 0.0022 for the standard activations and 1.1293 ± 0.0009 for the robust activations. Reconstruction statistics, SAE feature engagement (Section 3.3), and the ablation-KL population comparison (Appendix F) use all three seeds.

Table 9 reports the per-seed values underlying the aggregate statistics above.

Table 9: Per-seed SAE statistics. Bold values indicate the robust condition.

Metric	clean s42	clean s43	clean s44	robust s42	robust s43	robust s44
MSE, original residual space (train-mode)	1.1624	1.1579	1.1601	**1.1301**	**1.1295**	**1.1284**
MSE, inference-threshold mode	1.1546	1.1511	1.1526	**1.1262**	**1.1265**	**1.1254**
Final training loss (normalized space)	0.0877	0.0902	0.0869	**0.0768**	**0.0770**	**0.0788**
Dead latents	0	1	0	0	0	1
Inference threshold T	0.785	0.783	0.787	**0.741**	**0.742**	**0.743**

Table 10 reports per-seed candidate counts and feature counts at the 90% attribution-mass threshold underlying Table 3 and Section 3.3. The robust SAE requires fewer candidates and fewer features at every seed, both tasks, with no exceptions.

Table 10: Per-seed SAE feature engagement. Values are candidate count / features for 90% attribution mass, per seed.

	`ft-clean` (s42/s43/s44)	`ft-e100-a05` robust (s42/s43/s44)
GT candidate features	563 / 561 / 533	**360 / 385 / 357**
GT features for 90%	90 / 78 / 85	**63 / 61 / 59**
IOI candidate features	1,612 / 1,634 / 1,556	**1,196 / 1,188 / 1,162**
IOI features for 90%	343 / 363 / 341	**276 / 280 / 262**

F Attribution implementation and selectivity controls

For IOI, target attribution is summed over the first subject, indirect object, second subject, and final readout positions. For GT, it is summed over the start-year span and final readout. A feature enters the candidate set if it activates at any intervention position on a clean or corrupted prompt. Each model-task analysis ranks its own candidate features; rankings are never merged across SAEs. The ablation pool includes all features tied at the top-100 boundary.

F.1 Non-discriminative generic-language-modeling control

On held-out OpenWebText, generic next-token attributions are $\sim 10^{-8}$ against task attributions $\sim 10^{-2}$. A selectivity ratio against that control is uninformative: the denominator is negligible in both models.

F.2 Ablation-KL population comparison

We zero-ablate each of the top-100 most-attributed features and measure $D_{\mathrm{KL}}(p \| p_i^{\mathrm{abl}})$ from the unspliced to the ablated output distribution, using a real intervention rather than a gradient-based estimate: KL divergence has zero first derivative at the unablated distribution, so a linearized (attribution-patching) estimate of this quantity is identically zero regardless of the true effect. For GT, ablation is applied at the readout position, where pool features are already reliably active. For IOI, an earlier readout-only version left 98 of the top 100 features permanently inactive, since IOI features act predominantly at the name positions; we instead ablate at the union of positions in $\mathcal{T}_{\mathrm{IOI}}$ where a given feature is active on that prompt, which resolves this and leaves all 100 features active for both models.

Because the two SAEs are trained independently, a feature index carries no correspondence across models, so we do not pair features by index or form a per-feature ratio against target attribution. Instead, for each model and task we treat the resulting 100 ablation-KL values as an unpaired population and compare the two models' populations directly, across all three SAE seeds. Table 11 reports the full per-seed detail. Values exclude control-set-dead features (zero effect by construction, since ablating an already-inactive feature is a no-op); dead counts are reported separately per cell. Medians are close between models on both tasks; the separation is concentrated in the 90th-percentile tail. We report this as a directional pattern only: with three seeds, per-seed means are not separated by standard deviation, and one of three IOI seeds reverses direction.

Table 11: Ablation-KL effect, robust vs. standard, excluding control-set-dead features. Mean/median in units of 10^{-3}.

Task	Model	Dead (s42/43/44)	Mean (excl.)	Median (excl.)	p90 (excl.)
GT	Standard	14/16/15	7.4 / 8.3 / 10.7	$\sim$1.0	10–14
GT	Robust	16/16/21	**7.0 / 6.0 / 9.3**	**0.6–1.1**	**10–12**
IOI	Standard	0/0/0	5.4 / 8.5 / 10.9	$\sim$0.9–1.0	18–20
IOI	Robust	0/0/0	6.4 / 6.2 / 7.2	$\sim$0.9	**9–15**

G CIRCUIT RECOVERY PROTOCOL

For each model-task pair, we score all 32,491 edges of the 156-node raw computational graph (144 attention heads, 12 MLPs) using EAP-IG ($m = 5$ interpolation steps) and, as a pilot, EAP-GP, on identical prompt sets and counterfactuals: 200 IOI prompts (100 ABBA / 100 BABA, name-swap counterfactuals) and 200 of the first 207 GT prompts scanned, surviving the start-year parity filter (collection stops at 200 passes, so this count is smaller than the full-set 973-of-1000 in Appendix C), "01-dataset" counterfactuals. Circuit size is the smallest edge count, from a geometric grid $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1000, 2000, 4000, 8000, 16000, 32491\}$ (a denser tail past 512, not a continued power-of-two sequence, to avoid skipping the 85% crossing), that reaches 85% of full-graph faithfulness, together with the trapezoidal area under the full faithfulness curve on a log-k axis as a finer-grained discriminator.

Table 12 reports the raw $F(k)$ values underlying Figure 2, IOI, EAP-IG, per Equation 7. Below $k = 1000$, several entries are negative, meaning that subgraph performs worse than the fully corrupted baseline; at $k = 128$ standard leads by its largest margin in the curve ($F = 0.13$ versus -0.09 robust), still well short of the 85% threshold. Negative values at small k likely reflect $|\text{score}|$ ranking admitting edges with negative effect; signed ranking is left to future work.

Table 12: Raw $F(k)$, EAP-IG. IOI underlies Figure 2. GT underlies the 95% sentence in Section 3.4; GT is not competence-matched.

k	1	2	4	8	16	32	64	128
IOI Standard	-0.97	-0.97	-0.97	-0.23	-0.78	-1.08	-0.46	0.13
IOI Robust	-0.98	-0.98	-0.98	-0.49	-0.76	-0.89	-0.53	-0.09
GT Standard	0.00	0.00	0.00	0.00	0.15	0.38	0.66	0.88
GT Robust	0.00	0.00	0.00	0.07	0.14	0.31	0.65	0.88

k	256	512	1000	2000	4000	8000	16000	32491
IOI Standard	0.32	0.73	0.87	0.89	0.90	0.91	0.98	1.00
IOI Robust	0.20	0.61	0.91	0.97	0.92	0.98	0.99	1.00
GT Standard	0.92	0.89	0.93	0.94	0.87	0.97	0.97	1.00
GT Robust	0.94	0.94	0.89	0.99	0.99	0.96	1.00	1.00

GT-standard's first grid point reaching 95% is $k = 8000$: $k = 2000$ (0.94) and $k = 4000$ (0.87, non-monotonic) both fall short.

EAP-GP is a recently proposed variant using integrated gradient paths intended to mitigate the saturation effect known to affect EAP-IG. No public reference implementation existed at the time of this work; our implementation follows the method description, with one identified approximation – a single shared gradient path computed at the input activations and reused for all downstream edges,

rather than one path per node, matching the roughly 5× per-step cost the source paper reports relative to EAP-IG. This approximation is not independently validated against the source paper's own reported results. EAP-GP did not outperform EAP-IG on our graph for either task or model: area under the faithfulness curve was negative for EAP-GP on IOI for both models (standard: -0.97; robust: -0.70), against positive AUC for EAP-IG (standard: 0.87; robust: 0.60). We select the primary method per task by comparing AUC between EAP-IG and EAP-GP before inspecting the standard-versus-robust comparison, and report EAP-IG throughout the main text on this basis. We do not draw conclusions about EAP-GP's general effectiveness from this result; a genuine implementation bug, rather than a property of the method on this graph, cannot be ruled out.

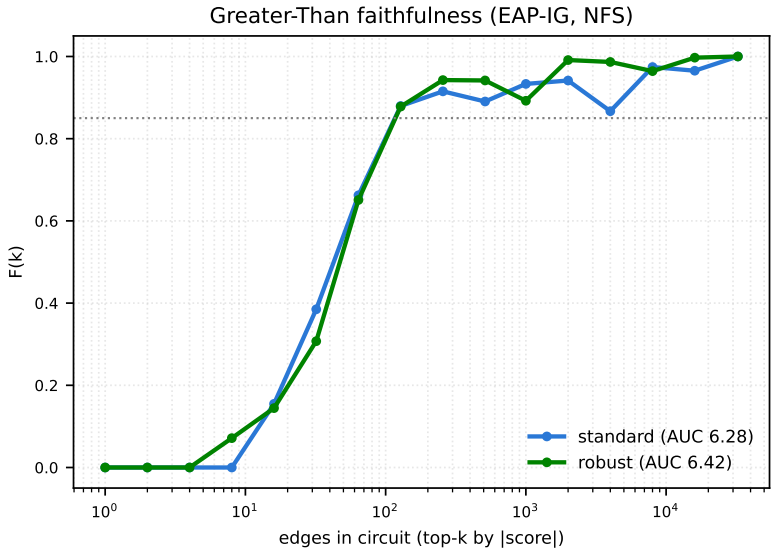


Figure 4: Greater-Than edge-native faithfulness, $F(k)$, EAP-IG (normalized-faithfulness-score, or NFS, normalization). AUC is an unnormalized trapezoid of $F(k)$ over $\ln k$ (maximum ≈ 10.4 if $F = 1$ throughout the grid); GT's curve stays near zero or above and rises early, while IOI's curve dips to $F \approx -1.08$ before climbing (Table 12), which is why GT and IOI AUCs sit on different numeric scales and are not directly comparable. Not competence-matched between models; not used as evidence for the main result (Section 3.4).

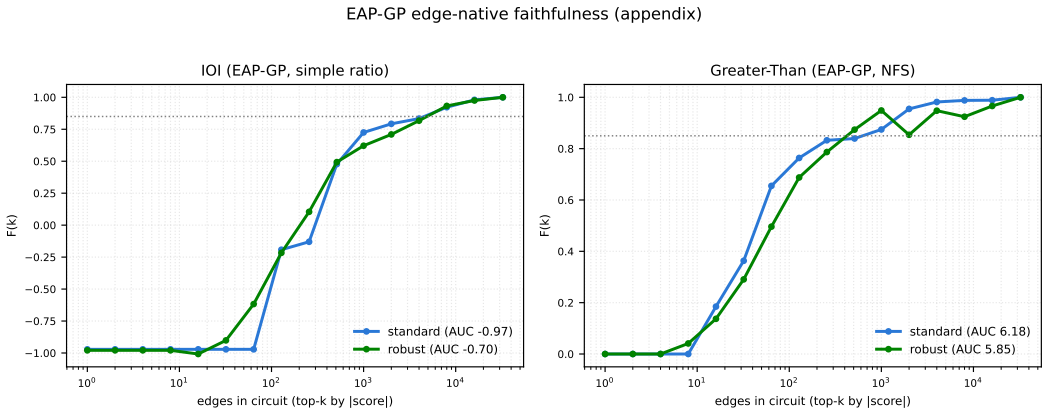


Figure 5: EAP-GP edge-native faithfulness, both tasks. Negative AUC on IOI for both models; EAP-IG is used for the headline result on this basis.

Top-k edge selection can yield a disconnected subgraph, which could in principle make a circuit look less faithful for reasons unrelated to the underlying attribution method's quality. As a check,

we rebuilt circuits with a connectivity-preserving greedy edge selector in place of plain top-k and recomputed both methods' faithfulness curves under it. The verdict does not change: EAP-IG beats EAP-GP under both builders, on both tasks and models. The greedy builder improves both methods' AUC, since connected circuits are generally more faithful, but it improves EAP-IG more than EAP-GP on IOI (+0.81 versus +0.48), so EAP-IG's advantage widens rather than narrows under the connectivity-preserving builder. The same holds on Greater-Than, our secondary check: EAP-IG's margin over EAP-GP moves from $+0.10/+0.56$ (standard/robust, top-k) to $+0.13/+0.54$ (greedy), essentially unchanged. Top-k disconnection is therefore not the source of EAP-GP's underperformance on either task.

Conditional co-ablation (CoAx) recovers backup components after primary-circuit ablation, following the scoring formula of Gong et al. (2026)[1], and tests whether apparently weak components become necessary once substitutes are removed, guarding against self-repair understating causal dependence. This step costs a fixed 295 forward passes for backup recovery plus 22 for the backup-validity gate, per model, task, and attribution method – constant by construction (a function of the candidate set size, not of which circuit is being recovered), so it does not itself discriminate between the standard and robust conditions; the resulting circuit's faithfulness curve does. Node-level faithfulness on the primary-plus-backup set remains far below 85% for both models (10–14 nodes). It is not a second 90/95% edge-level test.

H α-REQ: REPRESENTATION EIGENSPECTRUM DECAY

α-ReQ is a representational-simplicity measure independent of any trained SAE, used to test whether the dose-response in Section 3.5 is an artifact of the SAE family rather than a property of the underlying representation. For each residual-stream layer (every block's `resid_post`, plus the layer entering block 8 used for the SAE), we accumulate the centered activation covariance over a fixed evaluation token set, take its eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ ($d = 768$), and fit $\lambda_i \sim i^{-\alpha}$ by OLS on a log-log scale over ranks 6–691 (dropping the top-5 spikes, which reflect a small number of dominant directions rather than the bulk spectrum, and the bottom 10% floor, where eigenvalues approach numerical noise), applied identically to every model and layer. A more compressed representation decays faster, giving a larger α; we report this exponent at the model level (mean across layers) and at the SAE-hook layer specifically. Participation ratio, $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$, is a second summary of the same eigenvalue spectrum: unlike α, it requires no fitting or rank-range choice, so agreement between the two – both moving the same direction across the sweep – is evidence the α-ReQ trend is not an artifact of the fit range. HT-SR weight-spectrum alpha (WeightWatcher) was also measured and found flat across the full sweep (2.923–2.934) – the continual-training update does not move the weight spectrum measurably, so this metric is not used further.

I CORPUS ABLATION: FINEWEB

The corpus-ablation pair uses the same continual-training recipe, budget, and qualification gate as the primary standard-robust pair (Section 2.2–2.3), with FineWeb substituted for OpenWebText as the training and SAE corpus. SAE architecture, training budget, and evaluation protocol match Axis 1 exactly. Both arms pass the IOI qualification floor and the masking sanity check before comparison.

J REPRODUCIBILITY CHECKS AND LIMITATIONS

All experiments use seed 42 for model initialization where applicable, data order, PGD initialization, and primary SAE training; SAE reconstruction and attribution analyses are additionally replicated at seeds 43 and 44. The implementation is covered by 80 CPU tests on small models, including token accounting, attack constraints, value identity of the straight-through splice, attribution-patching linearity against finite differences, BatchTopK selection semantics, KL-control computation, and provenance checks.

[1] Propositions 2–3.

The main limitations are the use of one base architecture, one adversarial training regime, and one analyzed residual-stream site. SAE reconstruction and SAE feature engagement use three seeds; the ablation-KL population comparison (Appendix F) also uses three seeds but is not seed-robust (one of three IOI seeds reverses direction); circuit recovery (Section 3.4) uses a 200-prompt subsample, smaller than the 1,000-prompt sets used elsewhere, reflecting its higher per-prompt cost. The k-grid used for circuit size can rule out a large reduction but not distinguish modest differences inside the same bin (Section 3.4). On Greater-Than the robust model scores lower than the standard model on the same prompts (PD 0.442 vs. 0.671; paired difference -0.229 ± 0.010). We therefore do not treat GT circuit size as a competence-matched test. OpenWebText identifiers are synthesized rather than native. Our EAP-GP implementation follows the method description in the absence of a public reference implementation, with one identified approximation not independently validated against the source paper's reported results.