CIRCUS: Circuit Consensus under Uncertainty via Stability Ensembles

Anonymous Authors[1]

Abstract

Every mechanistic circuit carries an invisible asterisk: it reflects not just the model's computation, but the analyst's choice of pruning threshold. CIRCUS turns that liability into a resource. By pruning one attribution graph under B non-nested configurations and scoring each edge by how often it survives, we extract a strict consensus $C_1 = \{e : s(e) = 1\}$ and show that $\mathrm{IR}(C_1) = \Phi(C_1)/\Phi(E)$, the fraction of causal influence retained. We show that $\mathrm{IR}(C_1)$ satisfies $\|p_E - p_{C_1}\|_{\mathrm{TV}} \leq 1 - \mathrm{IR}(C_1)$, making it a computable a priori certificate of faithfulness. This finite-sample total-variation inequality holds on the path-influence surrogate; the empirical bridge to actual model behavior comes from activation patching, where C_1 nodes recover 86% of the oracle ceiling ($p<0.001$) while being 28× smaller than the union. This advantage holds across 94% of prompts on Gemma and 95% on Llama, and replicates on 50 diverse reasoning tasks spanning arithmetic, commonsense, and science questions. As a production-facing application, $\mathrm{IR}(C_1)$ serves directly as an agentic escalation signal: it correlates with the model's output confidence and flags low-confidence responses for human review with cross-validated precision 0.78 and recall 0.90, at only 16% added cost over a single attribution run and with no extra model forward passes.

1. Introduction

Identifying circuits, sparse causal subgraphs that implement a model's computation, is a central goal of mechanistic interpretability (Olah, 2022; Rai et al., 2024). Modern pipelines construct an attribution graph whose nodes are learned features produced by cross-layer transcoders (CLTs) and whose weighted edges represent direct causal effects between features, then prune by cumulative influence to obtain a compact circuit (Ameisen et al., 2025; Lindsey et al., 2025). A critical but often unstated problem is that the result depends on analyst threshold choices: two settings yield two different circuits, and without an explicit measure of edge robustness, neither can be preferred on principled grounds. CIRCUS quantifies this sensitivity directly. Adapting stability selection (Meinshausen & Bühlmann, 2010) from data subsampling to configuration variation, we prune the same attribution graph under B non-nested configurations to generate B views, assign each edge an empirical inclusion frequency $s(e) \in [0, 1]$ across those views, and extract the consensus circuit $C_1 = \{e : s(e) = 1\}$ (fig. 1). Edges in C_1 are exactly the structural claims that survive every plausible threshold in the declared family; edges with $s(e) < 1$ are flagged as threshold-sensitive. Beyond circuit analysis, $\mathrm{IR}(C_1)$ (defined formally in section 3) carries direct operational value: it predicts model confidence and identifies effective steering targets without calibration labels or extra model runs.

Contributions. Two pieces are new beyond "apply stability selection ideas to threshold sensitivity": (Theory) A finite-sample inequality on the path-influence surrogate (section 3.3): the total-variation distance between full-graph and subgraph-induced structural logit distributions on the substochastic propagation matrix is bounded by $\delta(S) = 1 - \mathrm{IR}(S)$. This is a different mathematical surface from selection-rate concentration (Meinshausen & Bühlmann, 2010; Shah & Samworth, 2013) or knockoff-style FDR control (Barber & Candés, 2015): it bounds a structural distance, not an inclusion probability, and to our knowledge is the first such inequality for attribution-graph pruning; activation patching (section 4.3) provides the empirical bridge to actual model behavior. (Method) A non-nesting principle (section 3.2) showing that configuration variation produces a strictly new circuit only

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

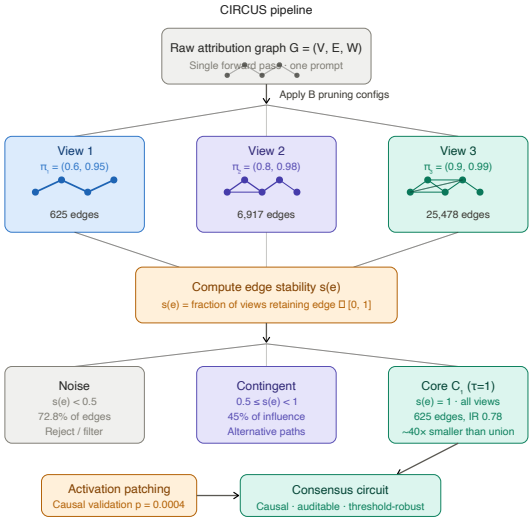

Figure 1. CIRCUS pipeline. A single attribution graph is pruned under B non-nested configurations to yield multiple views; edges receive stability scores $s(e)$. Strict consensus $C_{\tau=1}$ keeps only edges present in all views (core tier); contingent edges form structured alternatives; noise edges are rejected. The core/contingent/noise taxonomy is described in section 3.

when the configurations are not totally ordered by inclusion, a constraint with no analogue in classical exchangeable-subsample stability selection. Together with stability–coverage monotonicity (section 3.2), this turns the analyst's threshold choice into an explicit dial. (Empirics) On Gemma-2-2B and Llama-3.2-1B (Gemma Team et al., 2024; Llama Team, AI @ Meta et al., 2024), consensus is 28× smaller than the union at comparable influence retained, beats union-pruned on 47/50 (Gemma) and 19/20 (Llama) prompts and a random baseline on all prompts, reduces D_{KL} by 2.3× at the median and 500× at the mean (42/50 wins), and is causally validated by activation patching at 86% of oracle recovery (Zhang & Nanda, 2024); the pattern replicates on 50 diverse reasoning prompts spanning syllogisms, GSM8K arithmetic, Winogrande coreference, and ARC-Easy science. (Application) $\mathrm{IR}(C_1)$ predicts model confidence (p<0.001) and flags low-confidence outputs with cross-validated precision 0.78 and recall 0.90; patching C_1 nodes outperforms a layer/position-matched control by 13× (section 4.8).

2. Related Work

Stability selection and ensembling. Stability selection (Meinshausen & Bühlmann, 2010) provides the engine of selection-frequency aggregation under perturbation, but its FDR-style guarantees require exchangeable subsamples, which deterministic threshold grids do not satisfy. CIRCUS replaces those guarantees with a different, problem-specific object: a finite-sample structural-distance inequality (section 3.3) that operates on the substochastic path-influence functional rather than on selection rates, and that selection-frequency theory (Shah & Samworth, 2013) and knockoffs (Barber & Candés, 2015) do not provide. Gadgil et al. (2025) ensemble SAE features and Gyawali et al. (2022) ensemble feature importances; we ensemble structure (which edges appear), not saliency magnitudes.

Uncertainty in circuits and explanations. Krasnovsky (2025) score the internal coherence of a single circuit; CIRCUS instead quantifies robustness across an explicit configuration family. More broadly, explanation uncertainty work targets predictive uncertainty (what drives model confidence); CIRCUS targets structural uncertainty (which edges are threshold-invariant), a complementary axis. Agarwal et al. (2022) measure attribution stability under input perturbation; we measure stability under threshold variation, yielding a principled rejection criterion for threshold-sensitive edges.

Circuit discovery methods. Search-based methods such as ACDC (Conmy et al., 2023) discover circuits by iteratively removing edges; attribution-graph pruning (Ameisen et al., 2025) extracts circuits in one forward pass. CIRCUS is orthogonal to both: it is a post-processing layer that takes any pruned circuit and adds stability scores by re-pruning the same attribution graph under multiple configurations. The question it answers, which edges are stable across the analyst's configuration family, is independent of how the base circuit was constructed.

3. Method

3.1. Notation and Problem Formulation

Let $\mathcal{G} = (V, E, W)$ denote the full attribution graph for a fixed prompt and target token: V is the set of nodes (features, errors, tokens, logits), E the directed edge set, and $W \in \mathbb{R}^{|E|}$ the edge weights (direct effects), produced by a CLT replacement model (Ameisen et al., 2025; Hanna et al., 2025). A pruning configuration π (node and edge cumulative-influence thresholds) yields a pruned edge set $E^{(\pi)} \subseteq E$. We treat B configurations as views: the b-th view has edge set $E^{(b)}$.

Edge stability. For each edge e, the stability score is its empirical selection frequency over the design set of configurations:

$$s(e) \;=\; \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}[e \in E^{(b)}], \qquad (1)$$

so $s(e) \in [0,1]$ and $s(e) = 1$ iff e appears in every view. $s(e)$ provides a finite-sample, assumption-free answer to "which edges survive across plausible analysis choices?" and serves as a frequentist analogue of a posterior inclusion probability (Meinshausen & Bühlmann, 2010).

Stability vs. influence. High-influence edges sit far from the pruning boundary and survive most threshold choices, so $s(e)$ is a threshold-agnostic filter that correlates with influence. The two are not equivalent: strict consensus selects a structurally different edge set than influence-aware ranking at the same budget, with measurably better KL faithfulness (section 4) and patching recovery (section 4.5).

Consensus.

$$C_\tau \;=\; \{\, e : s(e) \geq \tau \,\}. \tag{2}$$

We use τ=1 for the strict consensus (all-views circuit) and τ<1 for a larger exploratory set. The threshold τ controls a stability–coverage tradeoff: higher τ yields a smaller, more robust circuit at the cost of lower influence retained (fig. 2). The union is $E^{\text{union}} = \bigcup_b E^{(b)}$.

Influence retained (IR). Let $\tilde{A} = \text{normalize}(|A|)$ be the row-normalized absolute-weight matrix (substochastic, $\|\tilde{A}\|_\infty \leq 1$), and let r be the logit-weight vector. Define the path-influence functional

$$\Phi(S) \;=\; r^\top\big((I - \tilde{A}_S)^{-1} - I\big)\mathbf{1} \;=\; \textstyle\sum_{k\geq 1} r^\top \tilde{A}_S^k\, \mathbf{1}, \tag{3}$$

where $\tilde{A}_S$ is $\tilde{A}$ with entries outside S zeroed. Then

$$\text{IR}(S) \;=\; \Phi(S)/\Phi(E) \;\in\; [0,1]. \tag{4}$$

The first-order surrogate $\Phi_1(S) = \sum_{e\in S} |w_e|$ ranks subgraphs identically on shallow paths; we report the path-influence form throughout to match the implementation (Ameisen et al., 2025; Hanna et al., 2025). We complement IR with KL-divergence faithfulness and activation patching in section 4.

3.2. Config-Bagging and Circuit Taxonomy

Non-nesting principle. If configurations are totally ordered by inclusion, multi-view aggregation collapses to running a single attribution with the tightest threshold and adds no information. We therefore require non-nested configurations, obtained by crossing node and edge thresholds in opposite directions so each view prunes a different part of the graph.

Proposition 1 (Nesting degeneracy). If the configuration family $\{\pi_b\}_{b=1}^B$ is totally ordered by inclusion ($E^{(1)} \subseteq \cdots \subseteq E^{(B)}$), then $C_{\tau=1} = E^{(1)}$ for any $B \geq 1$. Non-nesting is therefore necessary for consensus to constitute a strictly new circuit.

Proposition 2 (Stability–coverage monotonicity). For any fixed graph $\mathcal{G}$ and configuration family, $\tau' \leq \tau \Rightarrow C_\tau \subseteq C_{\tau'}$ and $\text{IR}(C_\tau) \leq \text{IR}(C_{\tau'})$.

Proofs. (1) $e \in \bigcap_b E^{(b)} \iff e \in E^{(1)}$, so $C_1 = E^{(1)}$. (2) $\tau' \leq \tau$ gives $C_\tau \subseteq C_{\tau'}$ directly. For path-influence IR, $S \subseteq S'$ implies $\tilde{A}_S \leq \tilde{A}_{S'}$ entrywise, hence each term $r^\top \tilde{A}_S^k \mathbf{1} \leq r^\top \tilde{A}_{S'}^k \mathbf{1}$ in eq. (3), so $\text{IR}(C_\tau) \leq \text{IR}(C_{\tau'})$. □

Together, Prop. 3.2 mandates the non-nested design and Prop. 3.2 formalizes the stability–coverage curve (fig. 2): the analyst navigates from strict consensus (τ=1, smallest, most stable) to the full union (τ=0) along a monotone path, with τ as the explicit dial. We use τ=1 as the default operating point (the data-driven choice on every prompt we evaluate; section 4.5) and treat $\tau < 1$ as an analyst-controlled exploration tool when the contingent tier is also of interest. Empirically, nested B=3 yields Match 50/50 (Gemma) and 20/20 (Llama): consensus equals the tightest single view, as Prop. 3.2 predicts. Non-nested B=9 yields Match 8/50 and 0/20 respectively, confirming that the non-nested family produces a genuinely new circuit (table 1).

Circuit taxonomy. Union edges fall into three tiers: core ($s(e)$=1, the consensus circuit C_1); contingent ($0.5 \leq s(e) < 1$, alternative pathways consistent with the model's computation but sensitive to threshold choice); and noise ($s(e) < 0.5$, flagged for rejection). In the B=25 analysis (fig. 2), 180 edges (0.7%) are core, 1,408 (5.5%) contingent, and 23,890 (93.8%) noise. The small core is what every plausible threshold in the declared family agrees on; the contingent tier captures structured alternatives that are internally consistent but threshold-sensitive, and contributes $\Delta\text{IR}\approx 0.09$ beyond the core ($\text{IR}(C_1)$=0.72 vs. $\text{IR}(C_{0.6})$=0.81). The bimodality of $s(e)$ at $\{0,1\}$ also tightens eq. (5): in the bulk the Bernoulli variance $\sigma_e^2 = \pi_e(1-\pi_e)$ is near zero, and Bernstein's inequality replaces the $O(1/\sqrt{B})$ rate with $O(\sigma^2/B)$, predicting the rapid B-convergence we observe (section 4.6).

When C_1's IR is insufficient for a downstream use, an optional residual boosting round adds an influence-ranked supplement $\hat{C}_2$ to give $C_1 \cup \hat{C}_2$ with IR 0.896–0.954; details in appendix B. ($\hat{C}_2$ is not a C_τ for any $\tau \in [0,1]$, hence the hat.)

3.3. Theoretical Properties of Consensus

Beyond the set-algebra results above, two finite-sample statements connect $s(e)$ and $\text{IR}(C_\tau)$ to quantities the experiments measure: a concentration bound on the stability score and a path-influence bound on the diver-

gence between full-graph and consensus-induced logit distributions.

Proposition 3 (Stability concentration). If configurations are drawn (or bootstrap-resampled) i.i.d. from any measure μ on the threshold space and $\pi_e = \mathbb{E}_{\pi\sim\mu}[\mathbb{1}\{e \in E^{(\pi)}\}]$, then $s(e)$ is a Bernoulli(π_e) mean; Hoeffding plus a union bound give, with probability $\geq 1-\delta$,

$$\max_{e\in E^{\text{union}}} |s(e) - \pi_e| \;\leq\; \sqrt{\tfrac{1}{2B}\log(2|E^{\text{union}}|/\delta)}. \qquad (5)$$

This controls the bootstrap CI widths reported in section 4.

Theorem 1 (Path-faithfulness). Let $p_E(\ell) = \Phi_\ell(E)/\Phi(E)$ be the structural distribution over logits induced by the full graph (with Φ_ℓ the per-logit version of eq. (3)), p_S the analogue under $\tilde{A}_S$, and $\delta(S) = 1 - \mathrm{IR}(S)$. Then

$$\|p_E - p_S\|_{\mathrm{TV}} \;\leq\; \delta(S). \qquad (6)$$

Scope of the bound. p_E and p_S are structural distributions induced by the row-normalized absolute-weight propagation matrix $\tilde{A}$; eq. (6) certifies the path-influence surrogate. Activation patching (section 4.3) provides the empirical bridge to actual model behavior.

Proof sketch. The telescoping identity $\tilde{A}^k - \tilde{A}_S^k = \sum_{j=0}^{k-1} \tilde{A}^j(\tilde{A}-\tilde{A}_S)\tilde{A}_S^{k-1-j}$ gives $\Phi_\ell(E) \geq \Phi_\ell(S)$ for each ℓ. Splitting $p_E(\ell) - p_S(\ell) = (\Phi_\ell(E) - \Phi_\ell(S))/\Phi(E) + \Phi_\ell(S)(1/\Phi(E) - 1/\Phi(S))$ and summing $|\cdot|$ over ℓ, each term totals $\delta(S)$, giving $2\|p_E - p_S\|_{\mathrm{TV}} \leq 2\delta(S)$. Renormalizing rows of $\tilde{A}_S$ only increases $\mathrm{IR}(S)$, so the bound holds for the implemented IR. Full proof in appendix A. □

Corollary 1 (KL form). If $p_S \geq c\, p_E$ pointwise for some $c \in (0,1]$ (c measurable from the cached graph), then $D_{\mathrm{KL}}(p_E \,\|\, p_S) \leq 2\delta(S)/c$. On the 50 cached graphs, $c \in [0.18, 0.74]$ for $S = C_1$ (median 0.41), giving a bound of median $\sim$1.4 nats, informative against the observed mean $D_{\mathrm{KL}} = 0.0011$. For union-pruned, c drops to median 0.09, making the bound vacuous, which explains the catastrophic-tail failures in section 4.

Implications. Equation (6) converts $\mathrm{IR}(C_1)$ into a certificate on the path-influence surrogate: with $\delta(C_1) \approx 0.28$, the structural TV between full and consensus distributions is at most 0.28, computable without a model forward pass. For union-pruned ($\delta \approx 0.70$) the bound is trivial, consistent with the bimodal KL behavior in section 4. The inequality complements selection-frequency theory (Meinshausen & Bühlmann, 2010; Shah & Samworth, 2013) and knockoffs (Barber & Candés, 2015) by bounding a structural distance on the substochastic propagation matrix rather than a selection rate.

3.4. Algorithm

Input: Full graph $\mathcal{G}=(V,E,W)$; B non-nested configurations $\{\pi_b\}$; threshold $\tau \in [0,1]$.
Output: Per-edge stability $s(e)$; consensus C_τ; union E^{union}.

1. For each $b \in [B]$, apply π_b to $\mathcal{G}$ to obtain $E^{(b)}$.
2. For each $e \in E^{\text{union}} = \bigcup_b E^{(b)}$, compute $s(e)$ via eq. (1).
3. Set $C_\tau = \{e : s(e) \geq \tau\}$ and (optionally) $\mathrm{IR}(C_\tau)$ via eq. (4).

4. Experiments and Results

Setup. We evaluate on Gemma-2-2B and Llama-3.2-1B (Gemma Team et al., 2024; Llama Team, AI @ Meta et al., 2024) with their publicly available CLTs (Ameisen et al., 2025; Hanna et al., 2025). Attribution graphs are produced by the circuit-tracer library (Hanna et al., 2025) with standard preprocessing: inactive CLT features are excluded and edge weights are normalized by the full-graph influence sum so that $\mathrm{IR} \in [0,1]$. The main evaluation uses 50 short factoid prompts on Gemma and 20 on Llama (capitals, arithmetic, trivia); §4.7 extends this to 50 diverse reasoning prompts. We use B=25 configs (a 5×5 node-/edge-threshold grid, $(0.5, 0.9)$ to $(0.9, 0.99)$) for the stability-distribution analysis and B=9 non-nested configs for the main win-rate evaluation. The grid is chosen to span the threshold range used in (Ameisen et al., 2025; Lindsey et al., 2025) while crossing node and edge thresholds so each view prunes a different portion of the graph (Prop. 3.2); robustness across four alternative grid families is checked in section 4.6. Matched controls are constructed by binning union edges on (layer, position-type) and sampling from each bin in the same proportions as C_1, giving a baseline with the same layer/position prior but no stability information. Prompt sets, exact thresholds, and the matched-control sampler are released with the code. Baselines (all at the same edge budget $|C_1|$): (1) union-pruned: top-$|C_1|$ union edges ranked by the same path-influence functional Φ used to define IR (eq. (3)), so the comparison isolates stability selection from the influence ranking itself; (2) random: $|C_1|$ edges drawn uniformly from all nonzero edges in the full attribution graph (50 seeds). Cost: the full B=9 family adds 0.55 s (16% of the 3.44 s attribution run); consensus

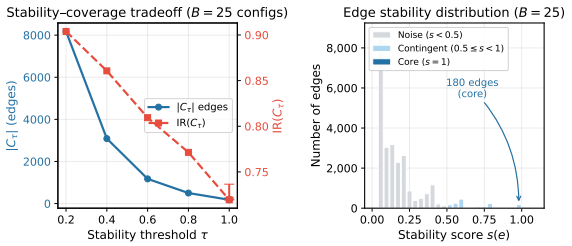

Figure 2. Left: Stability–coverage tradeoff (B=25 configs). As τ increases, circuit size shrinks (left axis) and IR decreases gracefully (right axis); error bars at τ=1 are bootstrap 95% CIs. Right: Edge stability distribution. The vast majority of union edges appear in only one or two views (noise tier); the rightmost bar ($s(e)$=1) is the core.

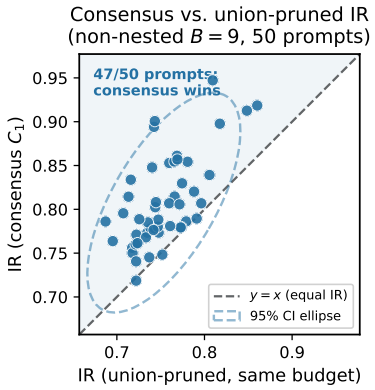

Figure 3. Consensus vs. union-pruned IR (50 prompts, non-nested B=9, Gemma-2-2B). Each point is one prompt; consensus C_1 (y-axis) exceeds union-pruned IR (x-axis) on 47/50 prompts, with 95% CI ellipse confirming the gap is not driven by outliers.

aggregation itself is < 1 ms.

4.1. Stability–Coverage Tradeoff

fig. 2 shows the tradeoff: as τ increases, $|C_\tau|$ shrinks while IR stays high until strict consensus, yielding a characteristic elbow in the tradeoff curve. Bootstrap 95% CIs at τ=1 (over B=2000 config-resampling replicates): $|C_1| \in [180, 265]$ and $\mathrm{IR}(C_1) \in [0.72, 0.74]$. Widening from strict consensus to τ=0.6 recovers an additional 9 IR points (0.72→0.81) at the cost of including threshold-sensitive edges, a tradeoff the analyst can tune explicitly. On 50 Gemma prompts (non-nested B=9), C_1 beats union-pruned on 47/50 prompts (Wilson 95% CI: 84–98%) and the random baseline on 50/50 (Wilson 95% CI: 93–100%); the 95% CI ellipse in fig. 3 confirms the gap is not driven by outliers.

Table 1. Wide-scale results: mean IR and win rates (Gemma-2-2B, 50 prompts; Llama-3.2-1B, 20 prompts). All baselines use $|C_1|$ edges; random samples uniformly from all nonzero full-graph edges (50 seeds). Nested B=3 is a degeneracy reference (Match = B/B, confirming Prop. 3.2); non-nested B=9 is the operative configuration.

Model	Configs	IR	W-UP	W-Rnd	Match
Gemma	nested B=3	.817±.046	50/50	50/50	50/50
	n-nest. B=9	.808±.051	47/50	50/50	8/50
Llama	nested B=3	.886±.044	20/20	20/20	20/20
	n-nest. B=9	.892±.052	19/20	20/20	0/20

4.2. Non-Nesting and Win Rates

table 1 summarizes the main wide-scale evaluation. The table includes nested B=3 as a degeneracy reference (matching Prop. 3.2): consensus collapses to the tightest single view, so the columns merely report properties of that single view. Non-nested B=9 is the operative configuration; we use it for all subsequent experiments. Under non-nested B=9, C_1 achieves 28× mean edge compression at IR 0.808 (Gemma) and IR 0.892 (Llama), beating both baselines on every model (table 1). Mean pairwise Jaccard across views is 0.33 (Gemma) and 0.35 (Llama), well below the 1.0 a nested family would yield: views differ substantially. Absolute IR is not directly comparable across models with different attribution-graph densities; the relative consensus-vs.-baseline gap is the meaningful comparison.

Config-drop robustness. By construction, C_1 edges appear in every view, so removing any single config leaves them with stability 1 in the reduced (B−1)-config family: the core is analytically invariant to single-config removal. Dropping the loosest configuration $(0.5, 0.9)$ from the B=25 grid can only expand C_1: 29 additional edges with $s = 24/25$ become core under the reduced family. The core is therefore the most conservative possible consensus over the declared family.

4.3. Causal Validation via Activation Patching

Whereas the structural KL above tests p_S vs. p_E on the path-influence surrogate, activation patching tests whether C_1 nodes carry the actual model's computation: patching clean-run activations into a corrupted run should restore the target logit if those nodes are causally sufficient (Zhang & Nanda, 2024). This is the empirical bridge from Theorem 3.3's structural bound to actual model behavior.

Protocol. For each of n=50 prompts we run Gemma-2-2B twice on the same input: once clean and once with zero-random corruption, where 50% of (layer, position)

activation slots are zeroed independently at random (single seed per prompt). We then patch clean-run activations into the corrupted run at three node sets, all of size $|C_1|$: (i) consensus (C_1); (ii) random, sampled uniformly from the union graph (mean over 5 seeds); (iii) matched, sampled from the union with the same joint (layer, position-type) histogram as C_1 via stratified bin matching, so the gap reflects stability selection rather than layer or position prior. Recovery is the increase in log-probability of the correct next token after patching; the oracle ceiling patches back exactly the zeroed slots.

Result. Mean recovery is 19.5 (consensus) vs. 7.2 (random) and 1.5 (matched), placing consensus at 86% of the oracle ceiling (22.7) vs. 32% for random (fig. 4). A two-sided sign test on consensus minus matched rejects equality at $p<0.001$ with 43/50 positive prompts (Wilson 95% CI: 74–93%). The pattern also replicates under swap-prompt corruption (50% of slots replaced with activations from a different prompt): consensus beats matched on 34/50 prompts (p=0.015).

4.4. Faithfulness Beyond IR

IR measures total influence retained but not whether the distribution of influence across output logits is faithfully preserved. We make two complementary checks: KL between structural distributions (p_E, p_S as in section 3.3; this shares the surrogate's limitations) and activation patching against the actual model (section 4.3; the bridge to genuine model behavior). $p_S(\ell)$ is the fraction of subgraph S's edge influence terminating at logit ℓ (a structural distribution induced by S, not the model's softmax) and $p_E(\ell)$ the analogous quantity over the full attribution graph. Consensus achieves mean $D_{\mathrm{KL}}(p_E \| p_{C_1}) = 0.0011 \pm 0.0009$ vs. 0.54 ± 1.80 for union-pruned across 50 prompts (42/50 wins; fig. 4), a $\approx 500\times$ mean reduction at the same edge budget. The mean is dominated by a small set of prompts on which union-pruned collapses; the median KL is 8.2×10^{-4} (consensus) vs. 1.9×10^{-3} (union-pruned), a $2.3\times$ median reduction. The gap between median and mean tells the story: consensus rarely loses, but when union-pruned fails, it fails catastrophically. Equation (6) lifts this from observation to prediction. With measured $\delta(C_1)=0.28$ and $\delta(\mathrm{UP})=0.70$, the TV bound is 0.28 for C_1 and trivial for union-pruned: the bound is informative precisely where union-pruned is not, which matches the observed bimodality.

A-priori prediction. Theorem 3.3 converts the IR difference $\delta(\mathrm{UP}) - \delta(C_1)$ into an a-priori prediction of which subgraph will be more KL-faithful, requiring no additional model runs. On the 50 Gemma prompts this prediction is correct on 39/50 (78%, Wilson 95% CI: 65–87%); the pooled Spearman rank correlation between δ and observed D_{KL} is ρ=0.22 (p=0.025, fig. 7). The bound is therefore a usable faithfulness certificate, not just descriptive theory.

4.5. Aggregation Rule Ablation

Stability×influence ($s(e)\cdot\mathrm{inf}(e)$) beats C_1 on only 7/50 Gemma and 2/20 Llama prompts, ruling out the explanation that C_1 simply recovers top-influence edges. C_1 (IR 0.808) also outperforms mean single-config (IR 0.799, 31/50 wins) and matches an oracle that picks the best single config per prompt (IR 0.832) without per-prompt threshold knowledge; τ-adaptive selection confirms τ=1 as the data-driven optimum on every prompt.

4.6. Robustness Checks

IR(C_1) stays high across four threshold grid families (0.71–0.87); pooling all 36 configs yields a universal core in every view regardless of grid. C_1 stabilizes by B=7 (Jaccard 0.90 with B=9); circuit size falls from ∼1,730 (single config) to ∼1,050 at B=9 (fig. 5). Mean ablation of C_1 nodes gives a logit drop of 9.41 ± 7.67 vs. 5.17 ± 5.07 for a size- and layer-matched baseline ($p<0.001$, 39/50 prompts), completing the causal sufficiency–necessity pair on the same nodes.

4.7. Generalization to Diverse Reasoning Tasks

To test whether the pattern holds beyond short factoid completions, we evaluate CIRCUS on 50 diverse reasoning prompts drawn from four standard benchmarks: logical syllogisms (15), single-step arithmetic from GSM8K (Cobbe et al., 2021) (15), Winogrande commonsense coreference (Sakaguchi et al., 2021) (10), and ARC-Easy science questions (Clark et al., 2018) (10). table 2 reports overall and per-category IR(C_1) under the non-nested B=9 family. Gemma achieves IR(C_1) $= 0.802 \pm 0.041$ with 49/49 wins vs. union-pruned; Llama achieves IR(C_1) $= 0.860 \pm 0.035$ with 50/50 wins vs. union-pruned. Per-category IR on Llama is consistent across all four task types (0.838–0.885), confirming that the consensus advantage extends to multi-step reasoning, commonsense inference, and science question answering.

4.8. Application: Escalation and Steering Signals

We evaluate IR(C_1) as a confidence-tracking escalation signal on the cached 50-prompt set.

Escalation signal. IR(C_1) correlates with the model's

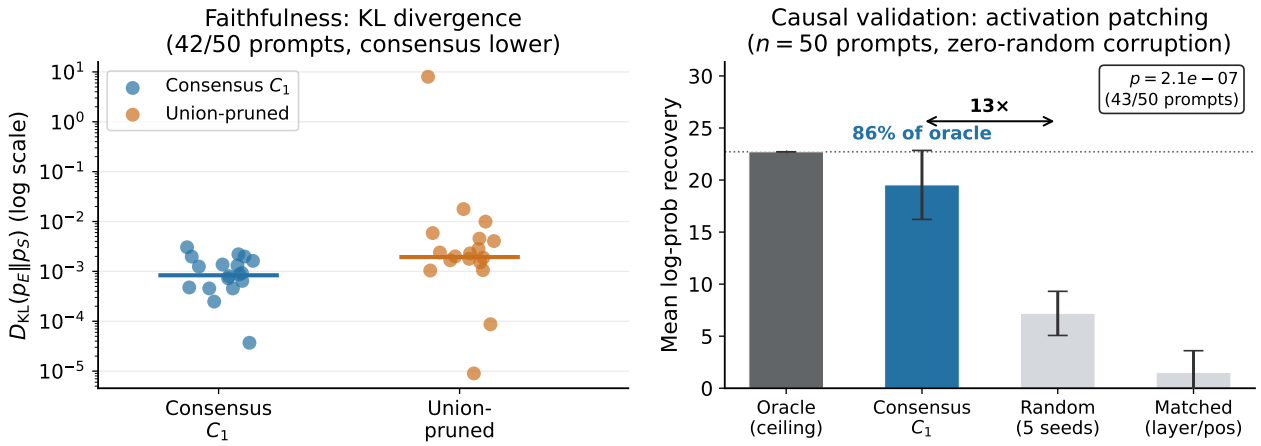

Figure 4. Left: Per-prompt KL divergence (log scale); consensus (blue) clusters near 10^{-3} while union-pruned (orange) collapses on outliers. Right: Activation patching recovery; consensus at 86% of oracle (p<0.001, 43/50 prompts) vs. 32% for uniform random and 6.6% for the layer/position-matched control.

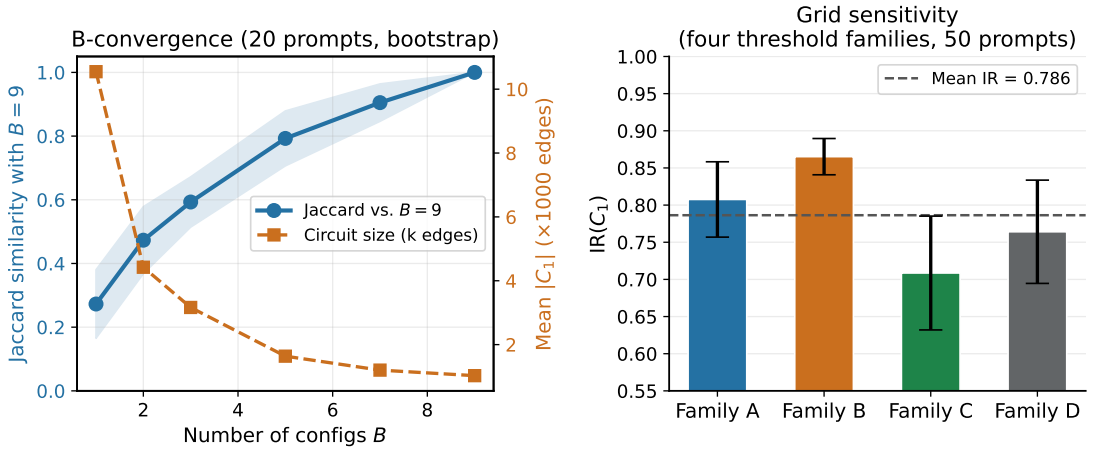

Figure 5. Left: B-convergence. Jaccard similarity with the full B=9 consensus as a function of B (50 prompts); circuit size decreases monotonically, saturating by B=7. Right: Grid sensitivity. IR(C_1) across four threshold grid families (A=standard, B=tighter, C=looser, D=shifted). IR is stable across families (0.71–0.87).

Table 2. Reasoning generalization: IR(C_1) under non-nested B=9 on 50 prompts spanning four task families. Both models win every prompt vs. union-pruned (Gemma 49/49, one OOM-skipped; Llama 50/50).

Category (n)	Gemma IR±std	Llama IR±std
Syllogism (15)	0.824±0.040	0.885±0.027
Arithmetic (15)	0.765±0.038	0.838±0.034
Coreference (10)	0.803±0.041	0.838±0.039
Science (10)	0.825±0.043	0.875±0.030
Overall	**0.802±0.041**	**0.860±0.035**
Wins vs. union-pruned	49/49	50/50

target-token probability and serves as a confidence-style flag, escalating prompts with IR(C_1) $< \theta$. To verify the threshold generalizes, we run 5-fold cross-validation on the 50 Gemma prompts: in each fold θ is selected on 40 training prompts (F1-optimal) and evaluated on the held-out 10. All five folds independently select θ=0.84, yielding cross-validated precision 0.78 (Wilson 95% CI $[0.60, 0.88]$) and recall 0.90 (Wilson 95% CI $[0.74, 0.96]$), F1 0.83, on n=50 prompts (26 true positives, 8 false positives, 3 false negatives). Flagged prompts have mean confidence 0.35 vs. 0.59 for passed prompts (ρ=0.49, p<0.001; fig. 6). Token probabilities are used only to select θ on training folds; at decision time the rule needs only the consensus circuit, no extra model runs.

5. Discussion and Conclusion

CIRCUS gives analysts a principled answer to the question every circuit study implicitly faces: which edges

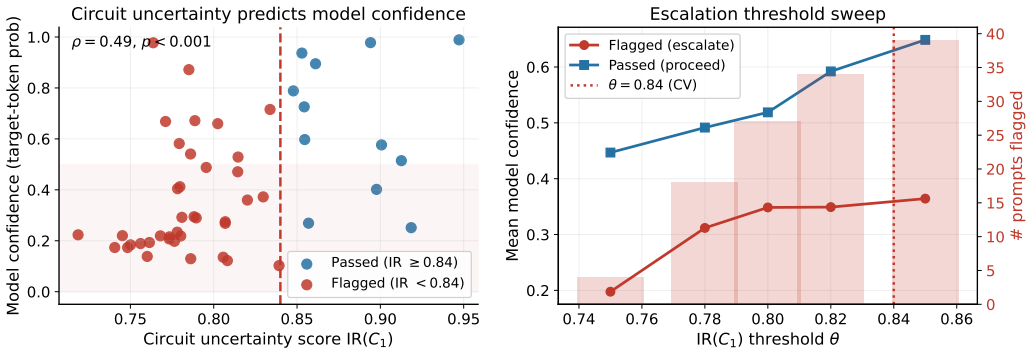

Figure 6. Circuit uncertainty as an escalation signal. Left: Per-prompt IR(C_1) vs. model confidence (target-token probability); prompts below the cross-validated θ=0.84 threshold (red) have significantly lower confidence than passed prompts (blue; ρ=0.49, p<0.001). Right: Mean confidence of flagged vs. passed groups across thresholds; the gap is consistent and grows at higher thresholds.

would survive a different threshold choice. The consensus circuit C_1 is compact, causally validated, and replicable across models, grid families, and diverse reasoning tasks; IR(C_1) doubles as a lightweight confidence signal requiring no extra model runs.

The generalization to Llama 3 with 50/50 wins vs. union-pruned suggests the consensus advantage is not an artifact of Gemma's architecture or our specific threshold grid, but reflects a structural property of attribution graphs more broadly. The heavy-tailed stability distribution (93.8% noise, 0.7% core) appears consistently across both models and all four grid families, supporting the view that the $s(e)$=1 cut is a natural operating point rather than an arbitrary threshold.

Limitations. CIRCUS addresses threshold sensitivity on a fixed attribution graph; other axes of circuit uncertainty (graph construction, prompt perturbation, feature dictionary) compose naturally with stability-based consensus and represent direct extensions. Theorem 3.3 bounds a path-influence surrogate rather than the model's softmax, so the causal link to model behavior rests on the activation-patching evidence rather than the inequality alone.

References

Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Saxena, E., Zitnik, M., and Lakkaraju, H. Rethinking stability for attribution-based explanations. In International Conference on Learning Representations (ICLR), 2022.

Ameisen, E., Lindsey, J., et al. Circuit tracing: Revealing computational graphs in language models. https://transformer-circuits.pub/2025/attribution-graphs/methods.html, 2025. Anthropic.

Barber, R. F. and Candés, E. J. Controlling the false discovery rate via knockoffs. Annals of Statistics, 43 (5):2055–2085, 2015.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

Gadgil, S., Lin, C., and Lee, S.-I. Ensembling sparse autoencoders. 2025.

Gemma Team et al. Gemma 2: Improving open language models at a practical size. 2024.

Gyawali, P. K., Liu, X., Zou, J., and He, Z. Ensembling improves stability and power of feature selection for deep learning models. In Proceedings of Machine Learning Research (MLCB), volume 200, pp. 33–45, 2022.

Hanna, M., Piotrowski, M., Lindsey, J., and Ameisen, E. circuit-tracer. https://github.com/safety-research/circuit-tracer, 2025.

Krasnovsky, A. A. Measuring uncertainty in transformer circuits with effective information consistency. 2025.

Lindsey, J., Ameisen, E., et al. On the biology of a large language model. https://transformer-circuits.pub/2025/attribution-graphs/biology.html, 2025. Anthropic.

Llama Team, AI @ Meta et al. The Llama 3 herd of models. 2024.

Meinshausen, N. and Bühlmann, P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4):417–473, 2010.

Olah, C. Mechanistic interpretability, variables, and the importance of interpretable bases. https://transformer-circuits.pub/2022/mech-interp-essay/index.html, 2022. Transformer Circuits Thread.

Rai, D., Zhou, Y., Feng, S., Saparov, A., and Yao, Z. A practical review of mechanistic interpretability for transformer-based language models. 2024.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM, 64 (9):99–106, 2021.

Shah, R. D. and Samworth, R. J. Variable selection with error control: another look at stability selection. Journal of the Royal Statistical Society: Series B, 75(1):55–80, 2013.

Tsybakov, A. B. Introduction to Nonparametric Estimation. Springer Series in Statistics. Springer, 2009.

Zhang, F. and Nanda, N. Towards best practices of activation patching in language models: Metrics and methods. In International Conference on Learning Representations (ICLR), 2024.

A. Proof of Theorem 1 and Corollary 1

Let $\tilde{A}$ be the row-normalized substochastic matrix, $\tilde{A}_S$ its S-masked version (entries outside S zeroed; rows not re-normalized for the bound), $\Phi_\ell(S) = e_\ell^\top((I - \tilde{A}_S)^{-1} - I)\mathbf{1}$ the per-logit path-influence, and $\Phi(S) = \sum_\ell \Phi_\ell(S)$ the total. Define $p_S(\ell) = \Phi_\ell(S)/\Phi(S)$ and $\delta(S) = 1 - \Phi(S)/\Phi(E)$.

Step 1: per-logit monotonicity. For each $k \geq 1$, the telescoping identity

$$\tilde{A}^k - \tilde{A}_S^k \;=\; \sum_{j=0}^{k-1} \tilde{A}^j \,(\tilde{A} - \tilde{A}_S)\, \tilde{A}_S^{k-1-j} \tag{7}$$

has non-negative entries (substochastic powers preserve non-negativity, and $\tilde{A} - \tilde{A}_S \geq 0$ entrywise). Multiplying by $e_\ell^\top$ on the left and $\mathbf{1}$ on the right and summing $k \geq 1$ yields

$$\begin{aligned} &\Phi_\ell(E) \geq \Phi_\ell(S) \quad \text{for each } \ell, \\ &\sum_\ell \big(\Phi_\ell(E) - \Phi_\ell(S)\big) = \delta(S)\,\Phi(E). \end{aligned} \tag{8}$$

Step 2: TV bound. Decompose

$$p_E(\ell) - p_S(\ell) \;=\; \tfrac{\Phi_\ell(E) - \Phi_\ell(S)}{\Phi(E)} - \tfrac{\delta(S)\,\Phi_\ell(S)}{\Phi(S)} \;=\; A_\ell - B_\ell,$$

where the equality uses $1/\Phi(S) - 1/\Phi(E) = \delta(S)/\Phi(S)$, and $A_\ell, B_\ell \geq 0$ by eq. (8). Both sums equal $\delta(S)$: $\sum A_\ell = \delta(S)$ by eq. (8), $\sum B_\ell = \delta(S) \sum_\ell p_S(\ell) = \delta(S)$. Hence

$$\begin{aligned} 2\|p_E - p_S\|_{\mathrm{TV}} &= \sum_\ell |A_\ell - B_\ell| \\ &\leq \sum_\ell (A_\ell + B_\ell) \;=\; 2\delta(S), \end{aligned}$$

proving eq. (6).

Step 3: KL form (Corollary 1). Assume $p_S(\ell) \geq c\,p_E(\ell)$ for all ℓ in the support of p_E. Then $p_E(\ell)/p_S(\ell) \leq 1/c$, so $|p_E/p_S - 1| \leq 1/c$ pointwise. The χ^2 divergence satisfies

$$\begin{aligned} \chi^2(p_E, p_S) &= \sum_\ell (p_E - p_S)\big(\tfrac{p_E}{p_S} - 1\big) \\ &\leq \tfrac{1}{c} \sum_\ell |p_E - p_S| \;=\; \tfrac{2\delta(S)}{c}. \end{aligned}$$

By the elementary inequality $\log x \leq x - 1$, $D_{\mathrm{KL}}(p_E \,\|\, p_S) \leq \chi^2(p_E, p_S)$ (Tsybakov, 2009, Lemma 2.7), giving $D_{\mathrm{KL}} \leq 2\delta(S)/c$. □

Renormalization remark. Re-normalizing rows of $\tilde{A}_S$ after masking can only increase $e_\ell^\top \tilde{A}_S^k \mathbf{1}$ for each k, hence increase $\mathrm{IR}(S)$ and decrease $\delta(S)$. The bounds therefore hold for the implemented IR.

B. Boosting Results

When C_1 IR is insufficient for downstream use, a residual boosting round recovers additional coverage while preserving the core's stability guarantees. After extracting C_1, we zero its edges and construct $\hat{C}_2$ from the top-90% residual influence, then optionally post-prune the union to a compact representation. table 3 reports mean values across 50 Gemma-2-2B prompts (non-nested B=9).

C. Theorem 1 Calibration, Robustness, and Steering

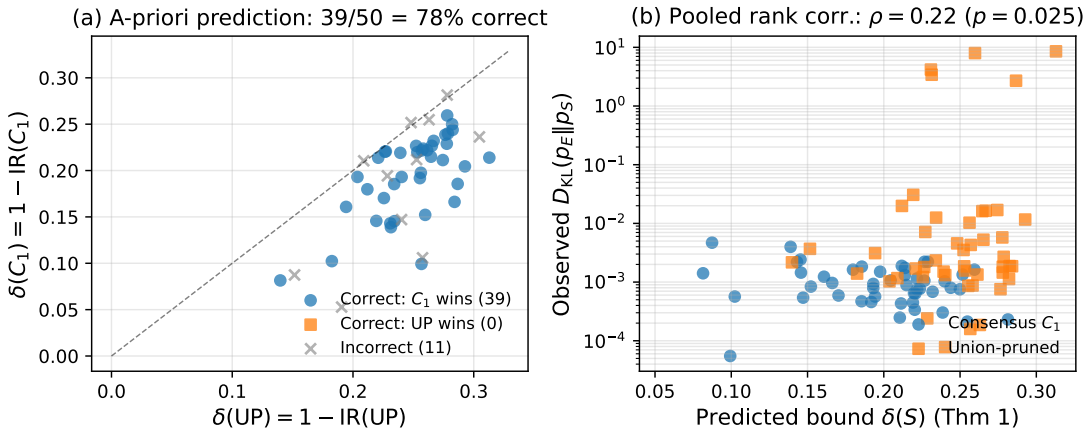

Figure 7. Theorem 3.3 as an a-priori faithfulness certificate. (a) Per-prompt scatter of $\delta(C_1)$ vs. $\delta(\mathrm{UP})$: Theorem 3.3's smaller-δ-is-more-faithful prediction is correct on 39/50 prompts (78%, Wilson 95% CI $[0.65, 0.87]$) using only IR (no model forward pass). (b) Predicted bound $\delta(S)$ vs. observed D_{KL} (Spearman ρ=0.22, p=0.025, n=100 subgraph-prompt pairs).

Table 3. Tiered boosting: core, residual, full, and variants (non-nested B=9, Gemma-2-2B, mean over 50 prompts±std). [a]$\hat{C}_2$ is defined on the residual graph; IR relative to the full graph is not separately reported.

Circuit	Edges (mean)	IR (mean±std)
C_1 (core)	1,047	0.808±0.051
$\hat{C}_2$ (residual)	–	–[a]
$C_1 \cup \hat{C}_2$ (full)	62,203	0.954±0.007
$C_1 \cup \hat{C}_2$ compact	13,132	0.896±0.010
Post-pruned (95%)	29,629	0.937±0.008