Certified Interventional Fidelity: Anytime-Valid, Adaptive Evaluation of Causal Claims in Mechanistic Interpretability

Amir Asiaee[1]

[1]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37232, USA

Abstract

Mechanistic interpretability often evaluates explanations by intervening on a model: swapping hidden states, patching activations, ablating components, or comparing a compressed model to the original one. These experiments are usually summarized by a point estimate, even though the evaluation may be monitored while it runs or adapted toward suspected failures. This makes it hard to tell whether a reported fidelity or patching effect is a stable causal claim or a consequence of finite sampling and evaluation choices.

We introduce **Certified Interventional Fidelity (CIF)**, a statistical layer for interventional interpretability evaluations. CIF first writes the quantity being reported as a causal estimand: an expectation of a bounded score over a stated input distribution and a stated intervention distribution. It then provides confidence intervals and anytime-valid confidence sequences for this estimand, including under adaptive intervention sampling via bounded mixture importance weighting. We instantiate CIF with Hoeffding-style sequences and variance-adaptive betting sequences, the latter reducing certification cost by 10–30$\times$ in our experiments. On MNIST abstractions and GPT-2 Small IOI circuits, CIF certifies high-fidelity claims, shows when apparent method differences are not statistically supported, and makes sensitivity to the intervention distribution explicit.

1 INTRODUCTION

Deep neural networks can match or exceed human performance, but understanding *how* they compute remains challenging. Mechanistic interpretability aims to provide explanations that are faithful simplifications of the internal computation of a trained model [Olah et al., 2020, Geiger et al., 2025]. A central move in recent interpretability has been the shift from observational probes to *interventional* evaluations: we modify internal states, paths, or mechanisms and ask whether the resulting behavior supports a proposed mechanistic explanation.

The explanations being evaluated usually fall into two broad classes. The first class consists of *abstraction or reduction fidelity claims*: a simpler object, such as a high-level causal model, compressed network, pruned network, or circuit, is claimed to preserve the relevant causal behavior of the original model. Causal abstraction and interchange intervention accuracy (IIA) are canonical examples [Geiger et al., 2021, 2025]; recent work on mechanism transformations places structured neural compression in the same interventional-fidelity setting [Asiaee, 2026]. The second class consists of *component-effect claims*: a set of heads, neurons, paths, or edges is claimed to causally support a behavior. Activation patching, path patching, causal tracing, circuit ablation, and many circuit-discovery evaluations belong to this class [Meng et al., 2022, Goldowsky-Dill et al., 2023, Zhang and Nanda, 2024]. Causal scrubbing and circuit-completeness evaluations sit near the boundary: depending on the reported score, they can be viewed either as testing a reduced explanation's fidelity or as measuring the effect of selected components [Chan et al., 2022].

The missing piece: statistical validity. Despite the causal framing, evaluation practice is often informal. A typical workflow is: sample a few thousand input pairs/prompt pairs; sample a few thousand interventions; report a single scalar metric. But interpretability workflows are inherently *sequential* and *adaptive*: we monitor results, adjust evaluation sets, and hunt for counterexamples. Without uncertainty quantification and without accounting for adaptivity, it is easy to overstate fidelity, mis-rank methods, or miss rare but important failure modes.

We propose **Certified Interventional Fidelity (CIF)**, a statistical layer for turning these interventional evaluations into

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:210–227.

certified claims. CIF is deliberately organized around the workflow an evaluator already follows: choose what population of inputs and interventions the claim is about, run the interventions, and report what can be concluded with uncertainty. The framework makes three design choices.

1. **Make the target explicit.** Every reported score is written as an expectation over a declared input distribution and intervention distribution. This separates the scientific question being evaluated from the sampling procedure used to estimate it.
2. **Report uncertainty that survives monitoring.** CIF uses anytime-valid confidence sequences, which remain valid if the evaluator checks the result repeatedly or stops when the evidence is strong enough [Howard et al., 2021, Waudby-Smith et al., 2024].
3. **Permit adaptive evaluation without changing the claim.** The sampler may focus over time on likely failures or high-impact interventions, while bounded mixture importance weighting preserves the original target estimand [Horvitz and Thompson, 1952, Ville, 1939].

Positioning relative to closest prior work. CIF connects three lines of work: causal abstraction and interventional interpretability [Geiger et al., 2025], recent diagnoses of unreliable mechanistic-interpretability evaluation [Méloux et al., 2025], and off-policy confidence sequences for adaptively sampled data [Karampatziakis et al., 2021]. The distinction is that CIF turns existing interventional interpretability metrics into explicit causal estimands and then supplies finite-sample, anytime-valid uncertainty guarantees for their evaluation. Extended related work is given in Appendix A.

Contributions. The central contribution is a reformulation: interventional evaluations of mechanistic explanations can be treated as sequential causal-inference problems. Abstraction/reduction fidelity claims and component-effect claims ask different scientific questions, but their empirical evaluations share the same statistical form: sample an input and an intervention, compute a bounded score, and estimate the population mean of that score. CIF attaches uncertainty guarantees to this population claim rather than to a method-specific point estimate.

- We express two broad classes of interpretability evaluations, abstraction/reduction fidelity and component-effect recovery, as bounded causal estimands over inputs and interventions.
- We give fixed-budget confidence intervals and anytime-valid confidence sequences for these estimands, using both transparent Hoeffding bounds and variance-adaptive betting sequences [Waudby-Smith and Ramdas, 2024].
- We introduce adaptive CIF, which supports failure-directed sampling while preserving the target estimand through bounded mixture importance weighting.
- We provide algorithms for certification, paired comparison, adaptive sampling, and one-sided stopping rules, together with a reporting checklist for applying CIF to new interventional evaluations.
- We evaluate CIF on MNIST neural abstractions and GPT-2 Small IOI circuits, showing that it can certify high-fidelity claims, identify statistically unsupported method differences, and expose sensitivity to the intervention distribution.

2 BACKGROUND AND NOTATION

2.1 A RUNNING NEURAL-NETWORK SCM

We use a small feedforward network as a running example. Write its computation as

$$H_1 = f_1(X), \qquad H_2 = f_2(H_1), \qquad Y = f_3(H_2),$$

where X is the random input drawn from an evaluation distribution $\mathcal{D}$, H_1 and H_2 are hidden activations, and Y is the output prediction or score vector. The corresponding deterministic SCM is $\mathcal{M} = (U, V, \mathcal{F})$: $U = \{X\}$ supplies the exogenous input, $V = \{H_1, H_2, Y\}$ contains the endogenous variables, and $\mathcal{F} = \{f_1, f_2, f_3\}$ contains the structural equations [Pearl, 2009, Geiger et al., 2021, Asiaee, 2026]. A realized input is denoted x. There is no extra randomness inside the network; for fixed x and a fixed intervention i, the resulting activations and output are deterministic.

In this view, an intervention changes one or more structural equations. Setting a neuron to zero is a hard intervention on a coordinate of H_2; replacing the computation of a group of units by an affine map is a soft intervention; patching an activation from a clean run into a corrupted run replaces the value of an internal variable during that forward pass. We write $\mathrm{State}(\mathcal{M}, x, i)$ for the realized assignment of the endogenous variables after running model $\mathcal{M}$ on input x under intervention i. It is a vector of values, not a distribution; it becomes random only when X and I are sampled.

2.2 CLASS I: ABSTRACTION AND REDUCTION FIDELITY

In the first class, the explanation is itself a simpler causal object. The low-level model $\mathcal{M}_L$ is the original network. The high-level model $\mathcal{M}_H$ is a candidate explanation produced by some abstraction or reduction method: for example a hand-written causal model, a pruned network, an affine compressed network, or a circuit treated as a reduced mechanism. CIF does not construct $\mathcal{M}_H$; it evaluates the candidate $\mathcal{M}_H$ together with the correspondence maps supplied by

the method. Two maps specify how the objects are supposed to correspond:

$$\tau : \mathrm{States}(\mathcal{M}_L) \to \mathrm{States}(\mathcal{M}_H), \qquad \omega : \mathcal{I}_H \to \mathcal{I}_L.$$

The state map τ translates low-level activations into high-level variables. In the running example, if $\mathcal{M}_H$ keeps a subset S of second-layer units, then $\tau(h_2) = h_{2,S}$. If $\mathcal{M}_H$ uses a learned affine compression, then $\tau(h_2) = Ah_2 + b$. The intervention map ω translates an intervention on the high-level variables into the corresponding intervention inside the original network. Thus τ maps states, while ω maps interventions. The form of ω is method-dependent and is part of the abstraction being evaluated.

The fidelity claim is interventional: $\mathcal{M}_H$ should match $\mathcal{M}_L$ not only on ordinary inputs, but under corresponding manipulations of their internal variables. Schematically,

$$\tau(\mathrm{State}(\mathcal{M}_L, X, \omega(I))) \approx \mathrm{State}(\mathcal{M}_H, X, I).$$

CIF evaluates this comparison through an output-level or score-level discrepancy, which is the quantity available in most mechanistic-interpretability experiments. The display is pointwise in realized inputs and interventions; the distributions enter when we average the resulting scores.

Example 1: interchange interventions for abstraction fidelity. Let X be a source input, X' a donor input, and $M \in \{0,1\}^k$ a random mask over the high-level coordinates of $\mathcal{M}_H$. A sampled high-level intervention is $I = (M, X')$. It replaces the selected abstract coordinates in the source run by their values from the donor run. The mapped low-level intervention $\omega(I)$ performs the corresponding swap in $\mathcal{M}_L$: for a subset abstraction it swaps the original neurons indexed by the retained coordinates; for a subspace or affine abstraction it applies the low-level operation specified by the abstraction map. The mask is therefore not an input mask; it indexes an internal intervention. CIF then asks whether these matched interventions produce similar outcomes in $\mathcal{M}_L$ and $\mathcal{M}_H$.

The intervention distribution Π is the sampling rule for these tests; a typical choice is $M_j \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(p)$ and $X' \sim \mathcal{D}_{\mathrm{donor}}$. Changing p changes the claim being certified: $p = 0.1$ asks for fidelity under mild coordinate swaps, whereas $p = 0.5$ asks for fidelity under a much more severe family of interventions.

2.3 CLASS II: COMPONENT-EFFECT AND RECOVERY CLAIMS

In the second class, the explanation is not a separate high-level model. Instead, it is a claim that some internal component, path, or circuit causally supports a behavior of the original network. The evaluation intervenes on $\mathcal{M}_L$ itself and measures a bounded effect. Activation patching, path patching, causal tracing, and many circuit-ablation studies have this form.

Example 2: activation patching for recovery. Let $X = (X_{\mathrm{clean}}, X_{\mathrm{corr}})$ be a clean/corrupted prompt pair, and let $I = S$ denote the component set being restored, such as a set of heads, neurons, paths, or residual-stream locations. The patched run follows the corrupted prompt except that the activations at S are copied from the clean run. If S is fixed, then Π is a point mass; if the evaluation samples many candidate components, Π records how those components are chosen. As in Example 1, Π is not a new causal semantics: it is the population of intervention tests over which the reported metric is averaged.

2.4 BOUNDED INTERVENTIONAL SCORES AND ESTIMANDS

The two classes differ scientifically, and CIF does not force them to use the same score. What they share is the statistical form: sample an input and an intervention, compute a bounded interventional score, and estimate its expectation.

Fidelity estimands. For abstraction/reduction fidelity, fix a bounded discrepancy $d : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ and define

$$Z \coloneqq d\big(\mathrm{Outcome}(\mathcal{M}_L, X, \omega(I)), \mathrm{Outcome}(\mathcal{M}_H, X, I)\big) \in [0,1]. \quad (1)$$

Then the interventional risk is $R \coloneqq \mathbb{E}_{(X,I)\sim\mathcal{D}\times\Pi}[Z]$, and fidelity is $F \coloneqq 1 - R$. Here small Z means small disagreement, so larger F means higher fidelity.

Recovery/effect estimands. For component-effect evaluations, the bounded score may already be oriented so that larger is better. For activation patching, let $g(\cdot)$ be a scalar behavior score, such as the indirect-object-identification (IOI) logit difference. Write $g_{\mathrm{clean}}(X)$, $g_{\mathrm{corr}}(X)$, and $g_{\mathrm{patch}}(X, I)$ for the score on the clean run, corrupted run, and patched run. A normalized patching recovery is

$$\Delta(I, X) \coloneqq \mathrm{clip}_{[0,1]}\left(\frac{g_{\mathrm{patch}}(X, I) - g_{\mathrm{corr}}(X)}{g_{\mathrm{clean}}(X) - g_{\mathrm{corr}}(X)}\right) \in [0,1].$$

Here $\Delta = 0$ means the patch recovers none of the clean behavior, while $\Delta = 1$ means full recovery. CIF certifies the direct mean $\mu \coloneqq \mathbb{E}_{(X,I)\sim\mathcal{D}\times\Pi}[\Delta]$. Appendix B gives additional method mappings in this notation.

3 CERTIFIED INTERVENTIONAL FIDELITY

Once an interventional metric has been chosen, the evaluation problem is statistical: estimate and certify its population value from a finite number of neural-network runs, while

avoiding overconfidence from monitoring, adaptation, and selection. CIF addresses four recurring tasks: **(P1)** finite-sample uncertainty quantification with guaranteed confidence intervals; **(P2)** sequential experimentation, where the evaluator may inspect results and stop early; **(P3)** adaptive failure-directed sampling, where later interventions focus on observed weaknesses; and **(P4)** selection and multiple comparisons across many candidate components, circuits, or abstractions.

We write all targets in a common bounded-mean form. Fix an evaluation design: an input distribution $\mathcal{D}$, a target intervention distribution Π, and a bounded interventional score $Y \in [0,1]$. CIF estimates and certifies the population mean $\theta := \mathbb{E}_{(X,I)\sim\mathcal{D}\times\Pi}[Y(X,I)]$. For abstraction fidelity, $Y = Z$ and $\theta = R$, with fidelity $F = 1 - R$. For recovery or component-effect metrics, $Y = \Delta$ and $\theta = \mu$. Thus the statistical machinery below is written for θ; translating back to F or μ only changes the final one-sided certification rule.

The section adds inferential machinery in the order it is needed in practice. Fixed-budget intervals address (P1) when the sample size is fixed before looking at results. Confidence sequences address (P2) by remaining valid under repeated monitoring and data-dependent stopping. Importance weighting addresses (P3) without changing the target distribution Π. Betting confidence sequences keep the same anytime guarantees but reduce certification cost when the observed variance is small. Paired inference and multiplicity control address (P4). Ordinary fixed-n central-limit-theorem intervals can be useful when n is large and committed in advance, but CIF emphasizes finite-sample, anytime-valid guarantees because interpretability evaluations are often monitored and adapted while they run.

3.1 FIXED-BUDGET CONFIDENCE INTERVALS

This is the right tool when the number of intervention tests n is fixed in advance. Draw $(X_t, I_t) \overset{\text{i.i.d.}}{\sim} \mathcal{D} \times \Pi$, compute $Y_t = Y(X_t, I_t) \in [0,1]$, and let $\widehat{\theta}_n := n^{-1}\sum_{t=1}^{n} Y_t$.

Proposition 1 (Hoeffding confidence interval for a bounded interventional mean)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$, $|\widehat{\theta}_n - \theta| \le \sqrt{\log(2/\delta)/(2n)}$.*

For fidelity, this interval is applied to R and transformed to $F = 1 - R$; for patching recovery, it is applied directly to μ. If Y_t is Bernoulli, as in 0–1 disagreement for IIA, exact binomial intervals can also be used.

Compute planning. To achieve half-width ε at confidence $1-\delta$, Hoeffding requires $n \ge \log(2/\delta)/(2\varepsilon^2)$. This is useful when the forward-pass budget is fixed in advance, but it does *not* justify checking repeatedly and stopping when the result looks favorable.

3.2 ANYTIME-VALID CONFIDENCE SEQUENCES (CSS)

Interpretability evaluations are often monitored while they run: after 100, 500, or 1,000 interventions, we may decide whether to stop, continue, or inspect failures. Fixed-n confidence intervals do not protect this workflow. A *confidence sequence* $(C_n)_{n\ge1}$ satisfies $\mathbb{P}(\forall n \ge 1 : \theta \in C_n) \ge 1-\delta$, so it remains valid under arbitrary stopping and repeated monitoring [Howard et al., 2021].

A simple construction spends the error probability over time. Let $\delta_n = 6\delta/(\pi^2 n^2)$ and $b_n = \sqrt{\log(2/\delta_n)/(2n)}$.

Theorem 1 (Anytime Hoeffding confidence sequence via spending)**.** *With the above spending schedule, with probability at least $1 - \sum_{n\ge1}\delta_n = 1-\delta$, for all $n \ge 1$ simultaneously, $\theta \in [\widehat{\theta}_n - b_n, \widehat{\theta}_n + b_n]$.*

Certification via one-sided bounds. Write $U_n = \widehat{\theta}_n + b_n$ for the upper confidence bound (UCB) and $L_n = \widehat{\theta}_n - b_n$ for the lower confidence bound (LCB). To certify fidelity $F \ge F_0$, apply the CS to R and stop when $1 - U_n(R) \ge F_0$. To certify recovery $\mu \ge \mu_0$, apply the CS to μ and stop when $L_n(\mu) \ge \mu_0$. The confidence-sequence guarantee makes either stopping rule valid even though the stopping time is data-dependent.

3.3 ADAPTIVE FAILURE-DIRECTED SAMPLING VIA BOUNDED IMPORTANCE SAMPLING

The previous subsections assume interventions are sampled from the target distribution Π. In practice, once failures appear, we may want to sample more interventions like them. This can find counterexamples faster, but a naive average under the adaptive sampler estimates the wrong population. CIF uses importance weighting to keep the estimand tied to the original Π.

Adaptive intervention policies. At time t, after observing the history $\mathcal{H}_{t-1}$, choose a proposal distribution $q_t(\cdot \mid \mathcal{H}_{t-1})$ over interventions and sample $I_t \sim q_t$. The proposal may depend arbitrarily on prior observations, enabling counterexample mining, but it must be chosen before observing the current score Y_t.

Mixture proposals to bound weights. Let Π be the *target* intervention distribution and $\tilde{q}_t$ any adaptive auxiliary proposal. Define the mixture $q_t := (1-\alpha)\,\Pi + \alpha\,\tilde{q}_t$ for $\alpha \in (0,1)$. Then the importance weight $w_t := \Pi(I_t)/q_t(I_t) \le 1/(1-\alpha) := W_{\max}$, so weights remain bounded even when $\tilde{q}_t$ is highly concentrated.

Weighted estimator. The Horvitz–Thompson-style estimator [Horvitz and Thompson, 1952] is $\widehat{\theta}_n^{\mathrm{IS}} := n^{-1}\sum_{t=1}^{n} w_t Y_t$.

Lemma 1 (Unbiasedness under adaptive proposals)**.** *For any adaptive sequence (q_t) with full support wherever $\Pi > 0$, $\mathbb{E}[\widehat{\theta}_n^{\mathrm{IS}}] = \theta$. Moreover, $M_n := \sum_{t=1}^{n}(w_t Y_t - \theta)$ is a martingale w.r.t. the history filtration.*

Theorem 2 (Anytime-valid CS under adaptive sampling)**.** *Assume $Y_t \in [0,1]$, $w_t \le W_{\max}$ almost surely, inputs $X_t \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, and the proposal q_t is $\mathcal{H}_{t-1}$-measurable. With $\delta_n = 6\delta/(\pi^2 n^2)$ and $b_n^{\mathrm{IS}} := W_{\max}\sqrt{\log(2/\delta_n)/(2n)}$, we have with probability at least $1-\delta$, for all $n \ge 1$, $\theta \in [\widehat{\theta}_n^{\mathrm{IS}} - b_n^{\mathrm{IS}}, \widehat{\theta}_n^{\mathrm{IS}} + b_n^{\mathrm{IS}}]$.*

Choosing the proposal. The mixture rate α controls an efficiency tradeoff. Larger α gives the auxiliary proposal more influence, which can expose rare failures sooner, but it also increases $W_{\max} = 1/(1-\alpha)$ and can widen worst-case bounds. This choice affects efficiency, not the estimand: as long as the weights are used, the target remains the original θ under Π. For transparency, adaptive evaluations should report α, describe the auxiliary proposal, and compare against an i.i.d. baseline when certification cost is a central claim.

3.4 VARIANCE-ADAPTIVE CONFIDENCE SEQUENCES VIA BETTING

The Hoeffding-style CSs above are simple and transparent, but conservative: the radius depends only on n and the range of observations, not on the empirical variance. For importance-weighted observations, this conservatism introduces a multiplicative $W_{\max}$ penalty that can make adaptive sampling *slower* than i.i.d. sampling. Betting CSs keep anytime validity while adapting to the observed variance.

Betting-based confidence sequences [Waudby-Smith and Ramdas, 2024] address this conservatism by adapting to the empirical variance of the data. To avoid overloading notation, let $B_t \in [0,1]$ denote the bounded observation passed to the betting tracker. The key idea is to construct a *capital process* for each candidate mean $m \in [0,1]$:

$$K_0(m) = 1, \qquad K_n(m) = \prod_{t=1}^{n}\big(1 + \lambda_t(B_t - m)\big), \quad (2)$$

where λ_t is a "bet" chosen based on past data. By Ville's inequality, if $\mathbb{E}[B_t \mid \mathcal{H}_{t-1}] = m$ under the candidate mean, then $\mathbb{P}(\exists\, n : K_n(m) \ge 1/\delta) \le \delta$, so the confidence set $C_n = \{m : K_n(m) < 1/\delta\}$ is an anytime-valid CS.

We use the *Online Newton Step* (ONS) betting strategy, which sets

$$\lambda_{n+1} = \mathrm{clip}_{[-c,c]}\big(\lambda_n + \eta\, g_n / A_n\big),$$
$$g_n = \tfrac{B_n - m}{1 + \lambda_n(B_n - m)}, \quad A_n = A_{n-1} + g_n^2,$$

with $\lambda_1 = 0$, $A_0 = 1$, $\eta = 2/(2 - \log 3) \approx 2.22$, and $c = 1/2$. In practice, we find C_n as an interval $[L_n, U_n]$ via bisection on (2), evaluating $\log K_n(m)$ in $O(n)$ per candidate.

Importance-weighted extension. For ordinary i.i.d. evaluation, set $B_t = Y_t$ and the betting CS targets θ. For adaptive CIF with weights $w_t \le W_{\max}$, set $B_t = \widetilde{Y}_t := w_t Y_t / W_{\max} \in [0,1]$, run the betting CS on B_t to obtain an interval for $\mathbb{E}[B_t] = \theta/W_{\max}$, and scale back. Because betting adapts to the empirical variance of B_t, the $W_{\max}$ factor no longer acts as a fixed multiplier on the radius; when $\mathrm{Var}(w_t Y_t)$ is small, the betting CS can be much tighter than Hoeffding.

Algorithm 1 only requires a streaming tracker: after each intervention, update with the observed score and optional importance weight, then query lower and upper confidence bounds. Hoeffding and betting CSs therefore serve as interchangeable evaluation layers.

3.5 COMPARING TWO EXPLANATIONS WITH PAIRED, ANYTIME-VALID INFERENCE

When comparing two abstractions, circuits, or component sets, evaluate both on the same sampled (X_t, I_t). The paired difference $D_t = Y_t^{(1)} - Y_t^{(2)} \in [-1,1]$ often has much lower variance than two separate estimates. The same CS machinery applies to $\mathbb{E}[D_t]$ after rescaling the range, and one can stop early when a one-sided CS excludes 0.

3.6 MULTIPLE COMPARISONS AND SELECTION

The previous guarantees are for a single pre-specified estimand. When scanning K components, circuits, or abstractions and selecting the best-looking result, uncorrected intervals are optimistic. Simple options are to: (i) allocate δ/K to each component (Bonferroni), (ii) use sequential "peeling" (stop sampling clearly suboptimal candidates), or (iii) use anytime-valid false discovery rate methods [Wang and Ramdas, 2022]. In the experiments we recommend reporting both naive point estimates and Bonferroni-corrected CSs to show how selection affects confidence.

Algorithms. Algorithm 1 summarizes the adaptive two-sided loop for a generic bounded score Y. Algorithm 2 gives one-sided certification; the user supplies whether larger or smaller values are favorable.

Proofs of the formal statements in this section are given in Appendix C. A short reporting checklist is provided in Appendix D.

4 EXPERIMENTS

We evaluate CIF in four settings: a controlled MNIST abstraction benchmark, a GPT-2 circuit evaluation on the IOI task, a sensitivity analysis over intervention distributions and metrics, and a coverage check under repeated monitoring. The first two settings test whether confidence sequences can

Algorithm 1 CIF with adaptive sampling (two-sided CS)

Require: target distribution Π; mixture rate α; confidence δ; target half-width ε; adaptive proposal builder $\tilde{q}_t(\cdot \mid \mathcal{H}_{t-1})$

$n \leftarrow 0, \widehat{\theta} \leftarrow 0$
while true **do**
  $n \leftarrow n + 1$
  build $\tilde{q}_n(\cdot \mid \mathcal{H}_{n-1})$ (e.g., AtP*/gradients or past failures)
  $q_n \leftarrow (1-\alpha)\Pi + \alpha\tilde{q}_n$
  sample $(X_n, I_n) \sim \mathcal{D} \times q_n$, compute $Y_n \in [0,1]$
  $w_n \leftarrow \Pi(I_n)/q_n(I_n)$
  update $\widehat{\theta} \leftarrow \widehat{\theta} + \frac{1}{n}\big(w_n Y_n - \widehat{\theta}\big)$
  set $\delta_n \leftarrow 6\delta/(\pi^2 n^2)$ and $b_n^{\mathrm{IS}} \leftarrow \frac{1}{1-\alpha}\sqrt{\frac{\log(2/\delta_n)}{2n}}$
  if $b_n^{\mathrm{IS}} \le \varepsilon$ **then**
    break
  end if
end while
return certified interval $[\widehat{\theta} - b_n^{\mathrm{IS}}, \widehat{\theta} + b_n^{\mathrm{IS}}]$

Algorithm 2 CIF-Certify: one-sided certification

Require: target distribution Π; confidence δ; target θ_0; direction larger/smaller-is-better

run CIF to maintain $[\widehat{\theta}_n - b_n, \widehat{\theta}_n + b_n]$
if larger-is-better and $\widehat{\theta}_n - b_n \ge \theta_0$ **then**
  return Certified
end if
if smaller-is-better and $\widehat{\theta}_n + b_n \le \theta_0$ **then**
  return Certified
end if
if budget exhausted **then**
  return Not certified
end if

certify useful interventional claims at realistic forward-pass budgets; the latter two check that the reported conclusions are stable to design choices and valid under the sequential workflows CIF is meant to support.

4.1 E1: CERTIFIED IIA FOR PRUNED AND TRANSFORMED ABSTRACTIONS

The first experiment evaluates abstraction fidelity under interchange interventions. The goal is not to declare one compression method universally best, but to ask which claims can be certified once uncertainty is reported.

Setup. We use a three-layer MLP ($784 \to 512 \to 512 \to 10$) trained on MNIST to $> 98\%$ test accuracy, following the architecture used in prior structured-mechanism work [Asiaee, 2026]. We construct abstractions at the second hidden layer using four methods: (i) hard interventions via structured pruning (zeroing coordinates and removing the corresponding columns from the downstream weight matrix), (ii) soft interventions via affine mechanism replacement (merging neurons using learned affine maps and compiling the result into a smaller dense network), (iii) a variance-based pruning baseline (removing coordinates with lowest activation variance across the training set), and (iv) a random pruning baseline. For each abstraction, we retain $k \in \{64, 128, 256\}$ of the 512 hidden units (additional pruning levels are reported in Appendix E), yielding 12 reported abstraction pairs. We evaluate under interchange interventions: for each trial we draw a pair (x, x') of MNIST images and a binary swap mask $m \in \{0,1\}^k$ (each coordinate swapped independently with probability $p = 0.5$), then compare the predictions of $\mathcal{M}_L$ and $\mathcal{M}_H$ under the corresponding interventions [Geiger et al., 2021].

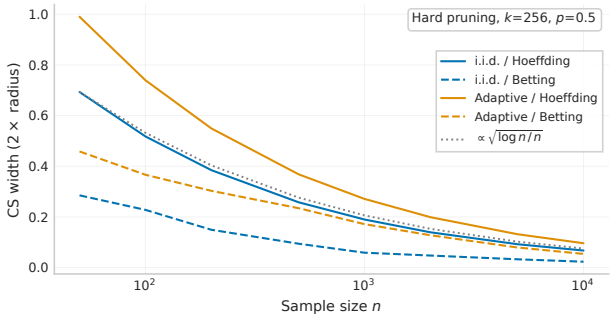

Figure 1: Confidence-sequence width versus forward passes for hard pruning at $k = 256$, $p = 0.5$. Solid curves use Hoeffding CSs; dashed curves use betting CSs. Blue denotes i.i.d. sampling, orange denotes adaptive mixture sampling with $\alpha = 0.3$, and the gray dotted curve shows a $\sqrt{\log n / n}$ reference rate.

Metrics. The main score is the disagreement indicator $Z \in \{0,1\}$, so $F = 1 - \mathbb{E}[Z]$ is interchange intervention accuracy. We also use bounded KL and L_2 discrepancies in the sensitivity analysis (E3). For each method and compression level, we report: (a) confidence sequences for F, and (b) the stopping time, measured in forward passes, to certify $F \ge F_0$ for $F_0 \in \{0.90, 0.95, 0.99\}$.

Adaptive proposal. For the adaptive runs, the auxiliary proposal places more mass on coordinates with larger downstream weight norm, $\tilde{q}_t(j) \propto \|W_{:,j}\|_2$. We mix this proposal with the target Bernoulli-mask distribution using $\alpha = 0.3$. Appendix F compares this static proposal with a history-updated failure proposal and sweeps α.

Results. Figure 1 illustrates the main inferential pattern. Hoeffding widths depend only on sample size and range, whereas betting intervals shrink with the empirical variance of the observed scores.

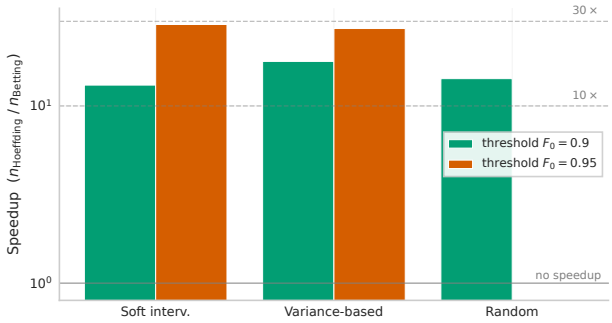

Figure 2: Certification-cost ratio $n_{\text{Hoeffding}}/n_{\text{Betting}}$ for each method at $k=256$, $p=0.1$, under i.i.d. sampling. Each run stops when the lower confidence bound on F exceeds the target F_0. Hard pruning is omitted because it never certifies; dotted reference lines mark $10\times$ and $30\times$.

At the aggressive swap probability $p = 0.5$, no non-identity abstraction reaches the certification threshold $F \geq 0.90$ (Table 1). This is a useful negative result: under large coordinate swaps, point estimates alone would still rank methods, but CIF shows that none supports a high-fidelity claim at this threshold. A paired comparison between soft interventions and random pruning at $k = 256$ further shows why interval choice matters: the estimated difference is -0.033, with Hoeffding radius 0.067 and betting radius 0.016.

At the milder swap probability $p = 0.1$, several abstractions become certifiable (Table 2). Variance-based pruning at $k = 256$ certifies $F \geq 0.90$ in 64 forward passes with betting, compared with 1,141 under Hoeffding. Soft interventions at $k = 256$ show the same pattern: certification requires 93 versus 1,219 samples for $F \geq 0.90$, and 342 versus 9,849 samples for $F \geq 0.95$. Figure 2 summarizes these reductions at $k = 256$.

Adaptive sampling is not uniformly faster for certification. In this high-fidelity regime, the adaptive mixture roughly doubles the sample count for variance-based pruning at $k = 256$ (131 samples versus 64 under i.i.d. betting). Appendix F shows the complementary case: adaptive proposals are more useful for finding failures than for certifying already high fidelities.

Tables 1 and 2 give the corresponding fidelity estimates and stopping times.

4.2 E2: CERTIFIED PATCHING EFFECTS AND CIRCUIT COMPLETENESS IN TRANSFORMERS

The second experiment evaluates component-effect claims in a transformer. Here each intervention requires a full GPT-2 Small forward pass, so certification cost is central.

Method	$k=64$	$k=128$	$k=256$
Hard pruning	0.105	0.104	0.105
Soft interv.	0.462	0.483	0.559
Variance-based	0.364	0.493	**0.689**
Random	0.278	0.401	0.521

Table 1: Estimated fidelity $\widehat{F}$ at $p=0.5$ ($n=5{,}000$, Hoeffding CS, $1-\delta=0.95$; half-width 0.046). No configuration certifies $F \geq 0.90$. Bold indicates the highest fidelity in the row block.

Method	k	$\widehat{F}$	n ($F \geq 0.90$)		Speedup
			Hoeff.	Bet.	
Soft int.	64	0.929	—	1,813	∞
Soft int.	128	0.956	3,393	131	25.9×
Soft int.	256	0.984	1,219	93	13.1×
Var.-based	64	0.633	—	—	—
Var.-based	128	0.917	—	4,725	∞
Var.-based	256	**0.990**	1,141	**64**	**17.8**×
Random	256	0.954	3,549	249	14.3×
$F \geq 0.95$ *certification:*					
Soft int.	256	0.984	9,849	342	28.8×
Var.-based	256	0.990	6,888	**252**	**27.3**×

Table 2: Certification at $p = 0.1$ (n up to 10,000, i.i.d., $1-\delta = 0.95$). "—" means the threshold is not certified within budget. Bold indicates the fastest certification time for each threshold.

Setup. We use GPT-2 Small (12 layers, 12 heads, 768-dimensional residual stream; ∼124M parameters) on the Indirect Object Identification (IOI) task [Wang et al., 2023]. In this task, the model must complete sentences of the form "When Mary and John went to the store, John gave a drink to" with the indirect object ("Mary"). The behavior score is the IOI logit difference $g = \text{logit(IO)} - \text{logit(S)}$, where IO is the indirect object and S is the subject. We evaluate circuits discovered via ACDC [Conmy et al., 2023], attribution patching [Syed et al., 2024], and AtP* [Kramár et al., 2024], as well as the hand-identified IOI circuit from Wang et al. [2023]. Patching interventions restore a set of nodes S from clean to corrupted prompts, following best practices [Zhang and Nanda, 2024, Hanna et al., 2024]. We use the clipped recovery score Δ from Section 2.4; CIF certifies $\mu = \mathbb{E}[\Delta]$, where larger values mean more recovery of the clean behavior.

Results. Figure 3 shows that all evaluated circuits have high recovery estimates, $\hat{\mu} \in [0.959, 0.972]$. The distinction is inferential: at $n = 2{,}000$, Hoeffding radii are about 0.070, while betting radii are about 0.005. Thus the same point estimates lead to much sharper lower confidence bounds under betting.

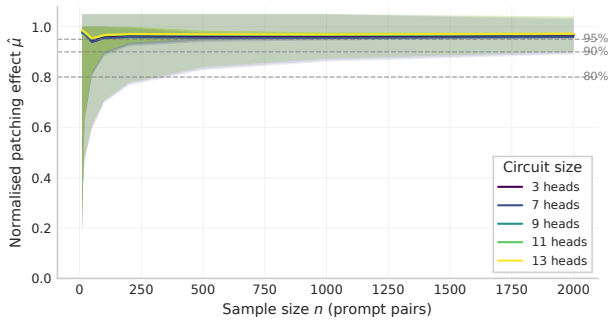

Figure 3: Confidence sequences for patching recovery μ on the IOI task with GPT-2 Small. Each curve corresponds to a circuit of increasing size (3–13 heads). Lighter bands show Hoeffding CSs; darker bands show betting CSs ($1-\delta = 0.95$).

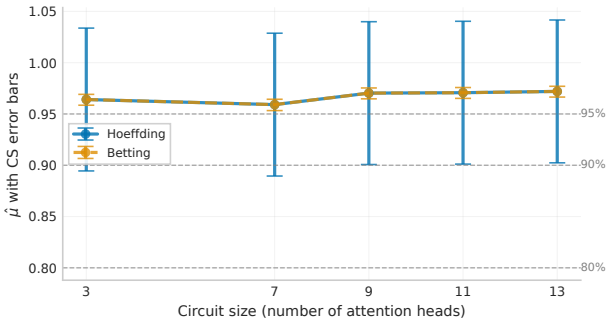

Figure 4: Point estimate $\hat{\mu}$ and confidence sequence at $n = 2{,}000$ as a function of IOI circuit size. Solid curves show Hoeffding CSs; dashed curves show betting CSs.

Circuit	$\hat{\mu}$	n ($\mu \geq 0.90$)		n ($\mu \geq 0.95$)	
		Hoeff.	Bet.	Hoeff.	Bet.
3 heads	0.964	—	110	—	840
7 heads	0.959	—	118	—	1,545
9 heads	0.970	1,973	104	—	413
11 heads	0.971	1,954	104	—	369
13 heads	0.972	1,875	**102**	—	**357**

Table 3: Certification of IOI circuits (GPT-2 Small, i.i.d., $1-\delta = 0.95$). "—" means the threshold is not certified within $n = 2{,}000$.

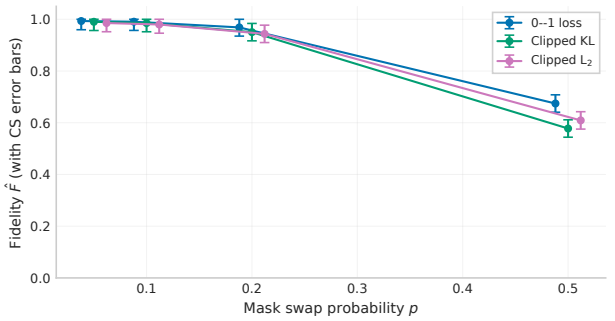

Figure 5: Sensitivity of certified fidelity to swap probability p and discrepancy metric d. At $p \leq 0.2$, the confidence sequences overlap; at $p = 0.5$, they separate.

Table 3 reports the resulting stopping times. The full 13-head circuit certifies $\mu \geq 0.90$ at $n = 102$ and $\mu \geq 0.95$ at $n = 357$ under betting; under Hoeffding, the corresponding requirements are $n = 1{,}875$ and more than the 2,000-sample budget. Even the 3-head name-mover circuit certifies $\mu \geq 0.90$ in 110 samples under betting.

Figure 4 plots recovery against circuit size at $n = 2{,}000$. Under betting, the lower confidence bound ranges from 0.953 for the 7-head circuit to 0.966 for the 13-head circuit, suggesting diminishing returns after the core name-mover heads are included.

4.3 E3: SENSITIVITY TO INTERVENTION DISTRIBUTION AND METRIC CHOICE

CIF certifies a metric under a declared intervention distribution. This experiment checks how much the conclusion changes when that design choice changes.

Setup. We fix a model and circuit and evaluate under multiple Π: different corruption baselines (for patching), different mask strengths (for interchange), and different behavior metrics (probability vs logit difference vs KL) [Zhang and Nanda, 2024]. For each setting, we report confidence sequences and check whether point-estimate differences survive uncertainty quantification.

Results. Figure 5 shows point estimates and confidence sequences across four swap probabilities $p \in \{0.05, 0.1, 0.2, 0.5\}$ and three discrepancy metrics (0–1 loss, clipped KL, clipped L_2). At mild perturbation ($p \leq 0.2$), the three metrics yield overlapping CSs: $\widehat{F} \in [0.94, 0.99]$ with CS width 0.067, so metric choice has no statistically detectable effect. At $p = 0.5$, the metrics separate: the 0–1 loss gives $\widehat{F} = 0.674$ while the clipped KL gives $\widehat{F} = 0.578$, and their CSs do *not* overlap (gap ≈ 0.03). At aggressive perturbation levels, the choice of discrepancy function can therefore change qualitative conclusions; CIF makes that dependence visible by reporting intervals rather than only point estimates.

4.4 E4: COVERAGE VALIDATION

Finally, we validate coverage under i.i.d. sampling, adaptive sampling, and an aggressive peeking protocol (500 runs per configuration; full details in Appendix G). Hoeffding CSs achieve empirical coverage 1.000 in all configurations. Betting CSs range from 0.930 to 1.000, always at or above

the nominal $1 - \delta$ while using less excess coverage.

5 DISCUSSION AND LIMITATIONS

Choosing the target distribution Π. CIF makes explicit that fidelity is defined relative to a choice of intervention distribution. We recommend reporting results across a small family of Π (e.g., weak vs strong interventions), and using adaptive CIF to probe likely failures without biasing headline metrics.

Boundedness and clipping. Our simplest guarantees assume bounded discrepancies/effects. For unbounded quantities (e.g., raw logit differences), one can either clip to a bounded range or use variance-adaptive CSs under additional assumptions [Howard et al., 2021]. In particular, sub-Gaussian and sub-exponential variants of time-uniform CSs [Howard et al., 2021], and asymptotic CSs [Waudby-Smith et al., 2024], extend the same anytime-valid workflow to unbounded estimands. Clipping is often acceptable in interpretability because extreme outliers can reflect out-of-distribution interventions rather than meaningful causal effects.

What CIF does not solve. CIF certifies estimates for a *given* metric and distribution. It does not guarantee that the chosen metric captures the desired notion of mechanistic truth. However, CIF makes it harder to overclaim by p-hacking sample sizes or adaptively searching without accounting for uncertainty. More fundamentally, CIF addresses the *estimation* problem of reliably estimating a causal estimand from finite data [Imbens and Rubin, 2015], but not the *identification* problem of whether the chosen abstraction captures the causal structure of interest.

6 CONCLUSION AND FUTURE DIRECTIONS

We have introduced Certified Interventional Fidelity (CIF), a statistical framework that provides anytime-valid confidence sequences for interventional interpretability metrics under both fixed and adaptive sampling regimes. CIF treats interventional metrics (interchange intervention accuracy, patching effects, circuit completeness scores) as formal causal estimands and wraps their evaluation with sequential inference guarantees that remain valid under arbitrary stopping, repeated monitoring, and adaptive counterexample hunting. CIF has a small interface: discrepancies must be bounded (or clipped to a bounded range), and the tracker can be applied as an evaluation layer around any existing interventional interpretability analysis.

A key practical finding is that the choice of CS construction can change certification cost by an order of magnitude. Hoeffding-style CSs are simple and transparent but conservative: their width depends only on n and the weight bound $W_{\max}$, not on the empirical variance of the observed scores. Betting CSs (Section 3.4) adapt to the data and reduce certification cost by 10–30$\times$ in our experiments, from thousands of forward passes to tens or low hundreds. This makes CIF-based certification practical even for expensive models like GPT-2, where additional forward passes are costly.

The immediate next step is replacing the Bonferroni correction in component scanning (Section 3.6) with e-value-based FDR control [Wang and Ramdas, 2022], which integrates natively with anytime-valid inference and matters most for hierarchical and per-component circuit scans, where Bonferroni is most lossy; extending CIF to multi-layer abstractions and richer attention-based interventions is the next challenge. CIF is deliberately a *verification* layer, the estimation side of the estimation/identification divide [Imbens and Rubin, 2015], but its anytime-valid bounds are natural controllers for *discovery*: integrated into automated circuit-discovery procedures [Conmy et al., 2023, Syed et al., 2024, Kramár et al., 2024], they can prune low-fidelity candidates early and reallocate evaluation budget toward contenders, turning certification from a post-hoc check into an active component of the search.

Acknowledgements

AA was partly supported by the Patient-Centered Outcomes Research Institute (PCORI) under award ME-2023C1-32148 and the National Institute of Mental Health under award R01MH139379.

Reproducibility. All experiments, tables, and figures in this paper can be reproduced with the code, data-processing scripts, and notebooks available at `https://github.com/AsiaeeLab/certified-interventional-fidelity`.

REFERENCES

Amir Asiaee. Causal mechanism reduction: Mechanism replacement for neural network pruning and abstraction. *arXiv preprint arXiv:2602.24266*, 2026.

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstractions. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 606–615. PMLR, 22–25 Jul 2020.

Sander L. Beckers and Joseph Y. Halpern. Abstracting causal models. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 2678–2685, 2019. doi: 10.1609/aaai.v33i01.33012678.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: a method for rigorously testing interpretability hypotheses. 2022. AI Alignment Forum / Redwood Research.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR, 17–23 Jul 2022.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR, 01–03 Apr 2024.

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

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

Jason Gross, Rajashree Agrawal, Thomas Kwa, Euan Ong, Chun Hei Yip, Alex Gibson, Soufiane Noubir, and Lawrence Chan. Compact proofs of model performance via mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011.

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 (COLM)*, 2024.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49 (2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 2021.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 68–87. PMLR, 11–14 Apr 2023.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic interpretability as statistical estimation: A variance analysis. *arXiv preprint arXiv:2510.00845*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *First Conference on Language Modeling*, 2024.

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

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2): 407–431, 2021. doi: 10.1111/rssa.12647.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, 2020.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

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 *International Conference on Learning Representations (ICLR)*, 2023.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022. doi: 10.1111/rssb.12489.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H. Kennedy, and Aaditya Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *The Annals of Statistics*, 52(6):2613–2640, 2024. doi: 10.1214/24-AOS2408.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Certified Interventional Fidelity: Anytime-Valid, Adaptive Evaluation of Causal Claims in Mechanistic Interpretability (Supplementary Material)

Amir Asiaee

Department of Biostatistics
Vanderbilt University Medical Center
Nashville, TN 37232, USA

A EXTENDED RELATED WORK

A.1 CAUSAL ABSTRACTION AND INTERVENTIONAL INTERPRETABILITY

Causal abstraction relates causal models at different granularities through state and intervention maps, with exact and approximate notions of commutativity [Rubenstein et al., 2017, Beckers and Halpern, 2019, Beckers et al., 2020, Massidda et al., 2023]. Interchange interventions operationalize these ideas in neural networks, yielding IIA as a graded faithfulness metric [Geiger et al., 2021, 2022, 2024, 2025]. Wu et al. [2023] show that distributed alignment search scales causal abstraction methods to large language models, making interchange interventions practical beyond small controlled settings.

A second strand uses causal mediation analysis inside neural networks. Vig et al. [2020] trace gender bias through transformer layers; activation patching, path patching, and causal tracing build on the same interventional logic [Meng et al., 2022, Goldowsky-Dill et al., 2023, Zhang and Nanda, 2024]. Causal scrubbing tests interpretability hypotheses through behavior-preserving resampling interventions [Chan et al., 2022].

A.2 CIRCUIT DISCOVERY AND SCALABLE APPROXIMATIONS

Automated circuit discovery methods attempt to localize important components with fewer interventions [Conmy et al., 2023, Syed et al., 2024, Kramár et al., 2024, Hanna et al., 2024]. Dunefsky et al. [2024] show that transcoders enable finer-grained circuit analysis than sparse autoencoders alone, increasing the number of hypotheses that require evaluation. These methods motivate CIF because they often scan many candidates and refine hypotheses adaptively, precisely the setting where fixed-sample point estimates can be misleading.

Recent work also studies the reliability of circuit evaluation itself. Makelov et al. [2024] identify an "interpretability illusion" in subspace activation patching, where apparently causal results arise from dormant pathways rather than the intended subspace. Miller et al. [2024] show that circuit faithfulness metrics are sensitive to the ablation method, complicating cross-study comparisons. Shi et al. [2024] develop formal hypothesis tests for circuit claims, and formal verification approaches provide complementary guarantees [Gross et al., 2024]. CIF complements these efforts by adding sequential, anytime-valid uncertainty quantification to adaptive interpretability workflows.

A.3 MECHANISM TRANSFORMATIONS AND COMPRESSION AS CAUSAL OPERATIONS

Recent work views a trained network as an SCM and interprets structured pruning as a hard intervention with an interventional-risk objective; soft interventions correspond to affine mechanism replacements and can be compiled exactly into a smaller dense network [Asiaee, 2026]. CIF serves a different role: it does not propose new abstractions, but certifies the interventional fidelity of a proposed abstraction, compression, or circuit.

A.4 SEQUENTIAL INFERENCE AND CONFIDENCE SEQUENCES

Ramdas et al. [2023] survey confidence sequences, e-values, and safe testing under the umbrella of game-theoretic statistics. CIF draws most directly on the nonparametric confidence sequences of Howard et al. [2021], which underlie our Hoeffding baseline, and the betting-based sequences of Waudby-Smith and Ramdas [2024], whose ONS strategy we use in Section 3.4. Safe testing [Grünwald et al., 2024, Shafer, 2021] and e-values [Vovk and Wang, 2021] provide related tools for optional stopping and evidence accumulation. The e-value false-discovery-rate methods of Wang and Ramdas [2022] are a natural replacement for Bonferroni correction when scanning many components.

In the bandits literature, Karampatziakis et al. [2021] develop off-policy confidence sequences with importance weighting, the closest statistical predecessor to adaptive CIF's mixture-importance-sampling layer. Concurrently, Méloux et al. [2025] diagnose variance and reliability issues in mechanistic interpretability evaluations; CIF provides one way to turn that diagnosis into certified evaluation practice.

A.5 IMPORTANCE SAMPLING AND ADAPTIVE MONTE CARLO

Importance sampling and Horvitz–Thompson estimators enable unbiased estimation under distribution shift [Horvitz and Thompson, 1952]. The potential-outcomes framework for causal inference [Imbens and Rubin, 2015] motivates treating interventional effects as estimands: each intervention defines a potential outcome, and the interventional risk R averages these outcomes under Π. CIF combines bounded-weight mixture proposals with martingale-based, anytime-valid bounds that remain valid when the proposal adapts over time.

B ADDITIONAL CIF METHOD MAPPINGS

Table 4 is organized by method family, with the claim class indicating what kind of mechanistic claim that family typically evaluates. Some circuit methods appear near the boundary depending on whether the reported score treats the circuit as a reduced mechanism to be faithful to the full model or as a component set whose causal effect is being measured.

Bernoulli IIA. When $Z = \mathbf{1}\{y_L^{\omega(I)}(x) \neq y_H^I(x)\}$ indicates disagreement, $R = \mathbb{P}(Z = 1)$ and $F = 1 - R = \mathbb{P}(Z = 0)$ is interchange intervention accuracy. Exact binomial confidence sequences can also be used; the betting CS in Section 3.4 already adapts to the low variance of such observations.

Patching recovery. For activation or path patching, the patched run follows the corrupted prompt except that selected activations are copied from the clean run. With behavior score g, the bounded recovery $\Delta \in [0, 1]$ measures the fraction of the clean-corrupted score gap recovered by the patch, and CIF certifies $\mu = \mathbb{E}[\Delta]$.

C PROOFS

We collect here the proofs of all formal results stated in the main text.

C.1 PROOF OF PROPOSITION 1 (HOEFFDING CONFIDENCE INTERVAL FOR A BOUNDED INTERVENTIONAL MEAN)

Proof. The random variables $Y_1, \ldots, Y_n$ are i.i.d. with $Y_t \in [0, 1]$ and $\mathbb{E}[Y_t] = \theta$. By Hoeffding's inequality [Hoeffding, 1963, Howard et al., 2021], for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{t=1}^{n} Y_t - \theta\right| > \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right).$$

Setting the right-hand side equal to δ gives $\epsilon = \sqrt{\log(2/\delta)/(2n)}$, and substituting $\widehat{\theta}_n = n^{-1}\sum_{t=1}^{n} Y_t$ completes the proof. □

C.2 PROOF OF THEOREM 1 (ANYTIME HOEFFDING CS VIA SPENDING)

Proof. The proof proceeds by a union bound over the fixed-sample Hoeffding inequality (Proposition 1) applied at each sample size n.

For each $n \geq 1$, Proposition 1 with confidence parameter δ_n gives

$$\mathbb{P}\left(|\widehat{\theta}_n - \theta| > \sqrt{\frac{\log(2/\delta_n)}{2n}}\right) \leq \delta_n.$$

Define $b_n = \sqrt{\log(2/\delta_n)/(2n)}$. By a union bound,

$$\mathbb{P}\Big(\exists\, n \geq 1 : |\widehat{\theta}_n - \theta| > b_n\Big) \leq \sum_{n=1}^{\infty} \mathbb{P}\Big(|\widehat{\theta}_n - \theta| > b_n\Big) \leq \sum_{n=1}^{\infty} \delta_n.$$

The spending schedule $\delta_n = 6\delta/(\pi^2 n^2)$ sums to δ, so

$$\mathbb{P}\Big(\forall\, n \geq 1 : \theta \in [\widehat{\theta}_n - b_n,\, \widehat{\theta}_n + b_n]\Big) \geq 1 - \delta.$$

Since the event $\{\forall n \geq 1 : \theta \in C_n\}$ holds uniformly over time, the CS remains valid under any possibly data-dependent stopping time T: if $\theta \in C_n$ for all n, then in particular $\theta \in C_T$. □

C.3 PROOF OF LEMMA 1 (UNBIASEDNESS UNDER ADAPTIVE PROPOSALS)

Proof. Fix time t and condition on the history $\mathcal{H}_{t-1} = \sigma(X_1, I_1, Y_1, \ldots, X_{t-1}, I_{t-1}, Y_{t-1})$. Given $\mathcal{H}_{t-1}$, the proposal q_t is deterministic, $X_t \sim \mathcal{D}$ is independent of the history, and $I_t \sim q_t(\cdot \mid \mathcal{H}_{t-1})$.

By the tower property of conditional expectation,

$$\begin{aligned}
\mathbb{E}[w_t Y_t \mid \mathcal{H}_{t-1}] &= \mathbb{E}\left[\frac{\Pi(I_t)}{q_t(I_t)} \cdot Y(X_t, I_t) \;\middle|\; \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}_{X_t \sim \mathcal{D}}\left[\mathbb{E}_{I_t \sim q_t}\left[\frac{\Pi(I_t)}{q_t(I_t)} \cdot Y(X_t, I_t) \;\middle|\; X_t, \mathcal{H}_{t-1}\right]\right] \\
&= \mathbb{E}_{X_t \sim \mathcal{D}}\left[\sum_{I \in \mathcal{I}} q_t(I) \cdot \frac{\Pi(I)}{q_t(I)} \cdot Y(X_t, I)\right] \\
&= \mathbb{E}_{X_t \sim \mathcal{D}}\left[\sum_{I \in \mathcal{I}} \Pi(I) \cdot Y(X_t, I)\right] \\
&= \mathbb{E}_{X \sim \mathcal{D},\, I \sim \Pi}[Y(X, I)] = \theta.
\end{aligned} \tag{3}$$

The full-support condition ensures the ratio is well-defined, and the cancellation $q_t(I)\Pi(I)/q_t(I) = \Pi(I)$ restores the target intervention distribution. In the continuous case, sums are replaced by integrals.

Taking unconditional expectations gives $\mathbb{E}[w_t Y_t] = \theta$ for all t, hence $\mathbb{E}[\widehat{\theta}_n^{\mathrm{IS}}] = \theta$.

For the martingale property, $M_0 = 0$ and $\mathbb{E}[M_n - M_{n-1} \mid \mathcal{H}_{n-1}] = \mathbb{E}[w_n Y_n - \theta \mid \mathcal{H}_{n-1}] = 0$, so $(M_n)_{n \geq 0}$ is a martingale. □

C.4 PROOF OF THEOREM 2 (ANYTIME-VALID CS UNDER ADAPTIVE SAMPLING)

Proof. By Lemma 1, $M_n = \sum_{t=1}^{n}(w_t Y_t - \theta)$ is a martingale. Since $0 \leq w_t Y_t \leq W_{\max}$ almost surely, each martingale difference has conditional range width at most $W_{\max}$. Hoeffding–Azuma for martingales with bounded conditional ranges gives, for any $\epsilon > 0$,

$$\mathbb{P}(|M_n| > \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2}{nW_{\max}^2}\right).$$

Setting $\epsilon' = \epsilon/n$ gives

$$\mathbb{P}\Big(|\widehat{\theta}_n^{\mathrm{IS}} - \theta| > \epsilon'\Big) \le 2\exp\left(-\frac{2n\,\epsilon'^2}{W_{\max}^2}\right).$$

Thus for each fixed n, $\mathbb{P}(|\widehat{\theta}_n^{\mathrm{IS}} - \theta| > b_n^{\mathrm{IS}}) \le \delta_n$, where $b_n^{\mathrm{IS}} = W_{\max}\sqrt{\log(2/\delta_n)/(2n)}$.

The same union bound over all $n \ge 1$ as in Theorem 1 yields

$$\mathbb{P}\Big(\forall\, n \ge 1 : \theta \in [\widehat{\theta}_n^{\mathrm{IS}} - b_n^{\mathrm{IS}},\ \widehat{\theta}_n^{\mathrm{IS}} + b_n^{\mathrm{IS}}]\Big) \ge 1 - \sum_{n=1}^{\infty}\delta_n = 1 - \delta.$$

□

D PRACTICAL REPORTING CHECKLIST

- State $\mathcal{D}$ (data/prompt distribution), Π (intervention distribution), and the bounded score being averaged, such as a discrepancy Z or recovery score Δ.
- Report a confidence sequence, not just a point estimate.
- If you sample adaptively, report α and confirm that you used mixture importance sampling.
- When scanning many components, either correct for multiple comparisons or clearly state that reported intervals are *post-selection* and optimistic.

E E1 CERTIFICATION AT ADDITIONAL PRUNING LEVELS

Table 5 extends Table 2 to the remaining pruning levels $k \in \{32, 384, 512\}$ ($p = 0.1$, i.i.d., $1 - \delta = 0.95$, n up to 10,000). At $k = 32$ no method reaches certifiable fidelity ($\hat{F} < 0.83$ everywhere). At $k = 512$ no units are removed, so all four methods coincide with the identity abstraction and certify at identical cost. The betting-vs-Hoeffding pattern of Table 2 persists at every level where certification is possible.

F ADAPTIVE VERSUS I.I.D. SAMPLING: MATCHED COMPARISON AND α-SWEEP

This appendix reports auxiliary adaptive-sampling studies that inform the efficiency discussion.

Matched three-way comparison. In the E1 setting, we compare three samplers under the betting CS ($\delta = 0.05$, $\alpha = 0.3$ for both adaptive arms): (i) i.i.d. sampling from Π; (ii) the *static* proposal $\tilde{q}(j) \propto \|W_{:,j}\|_2$ of Section 3.3; and (iii) a *history-updated* proposal that, after each batch of 100 samples, refits a smoothed per-coordinate failure rate $\tilde{q}_t(j) \propto (s_j+1)/(n_j+2)$, where s_j and n_j are the cumulative disagreement count and visit count of coordinate j. Two regimes are tested on five matched seeds each: *low-fidelity exclusion* (soft interventions, $k = 256$, $p = 0.5$, where $\hat{F}_{5000} \approx 0.55$; the task is to certify $F < 0.70$) and *borderline certification* (variance-based pruning, $k = 128$, $p = 0.1$, where $\hat{F} \approx 0.92$; the task is to certify $F \ge 0.90$).

Table 6 and Figure 6 report stopping times. The pattern is asymmetric. For failure-finding, the history-updated proposal (median 118, IQR 101–129) beats the static proposal on every seed (median 201, IQR 151–219) and is competitive with i.i.d. (median 88, IQR 86–150; the history proposal wins 3 of 5 matched seeds with a tighter IQR). For certification, every adaptive proposal loses to i.i.d. sampling: medians 3,800 (history, IQR 806–4,031) and 1,708 (static) versus 1,217 (i.i.d.). All samplers converge to the same $\hat{F}$ in both regimes, consistent with Lemma 1. We do not have a formal result for this direction-dependence; the observed pattern matches the one-sided importance-sampling intuition described in Section 3.3.

Sweep over the mixture rate α. On the borderline-certification configuration of Table 2 (variance-based pruning, $k = 256$, $p = 0.1$, betting CS, three seeds per point), we sweep $\alpha \in \{0, 0.1, 0.2, 0.3, 0.5, 0.7\}$ with the static proposal. Figure 7 shows median certification cost with IQR bars. For $F_0 = 0.90$ the medians are 64, 96, 101, 121, 156, and 255 samples; for $F_0 = 0.95$ they are 223, 190, 199, 240, 328, and 575. The non-monotonicity at small α is within three-seed noise; the slowdown above $\alpha = 0.3$ is systematic and matches the $W_{\max} = 1/(1-\alpha)$ weight-inflation bound (at $\alpha = 0.7$, $W_{\max} = 3.33$ and certification at $F_0 = 0.95$ is 2.6× slower than i.i.d.).

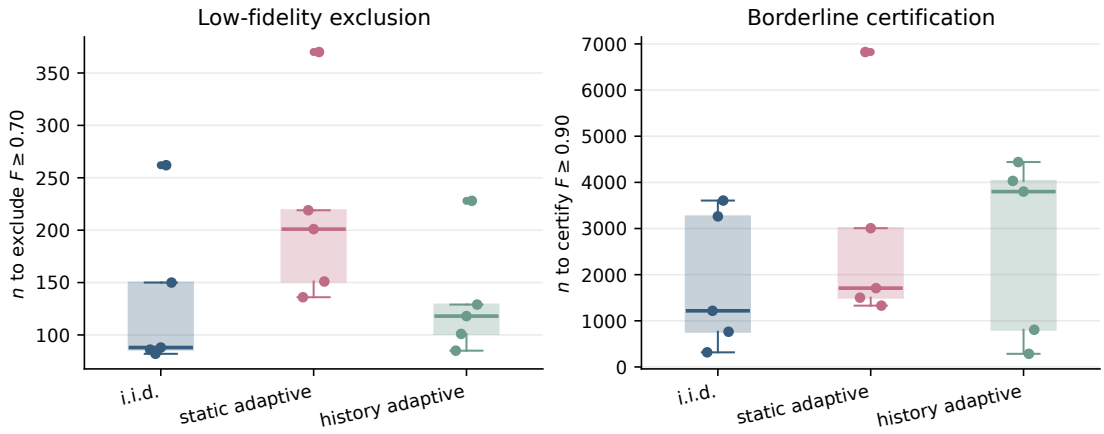

Figure 6: Stopping times across five matched seeds per sampler (boxes: IQR; dots: individual seeds). Left: low-fidelity exclusion (n to certify $F < 0.70$). Right: borderline certification (n to certify $F \geq 0.90$; note the static-adaptive outlier at $n = 6{,}824$). History-updated adaptive sampling helps failure-finding (left) and hurts certification (right).

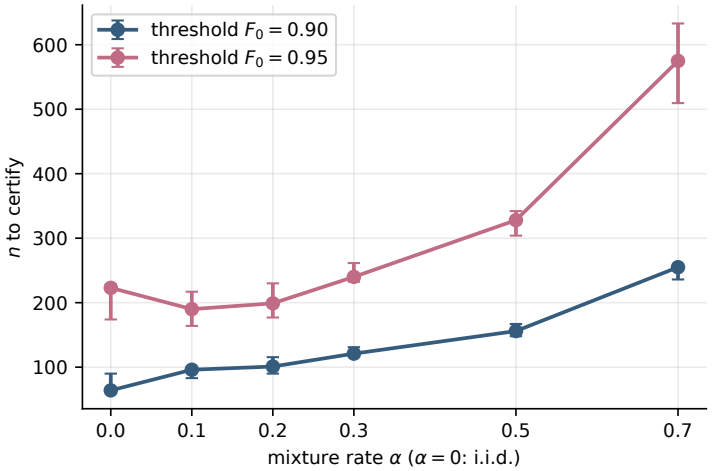

Figure 7: Median certification cost versus mixture rate α (variance-based pruning, $k = 256$, $p = 0.1$, betting CS; three seeds per point; bars: IQR). F_0 is the certification threshold: each run terminates when the lower confidence bound on F exceeds F_0.

G E4: COVERAGE VALIDATION DETAILS

Setup. In the MNIST MLP setting (E1), we fix one abstraction and estimate R across 500 independent runs under three protocols: i.i.d. sampling from Π, adaptive mixture sampling with $\alpha = 0.3$, and aggressive peeking, where the interval is checked after every 10 samples and the run stops at the first exclusion of a pre-specified value. For each protocol, we record whether the true R (estimated by Monte Carlo with 10^5 i.i.d. samples) falls inside the confidence sequence at the stopping time.

Results. Table 7 reports empirical coverage for Hoeffding and betting CSs across three confidence levels. Hoeffding coverage is 1.000 in every configuration, reflecting the conservatism of the range-based bound. Betting coverage ranges from 0.930 to 1.000, always at or above the nominal guarantee while leaving less unused error budget.

Claim class	Method family	Mechanistic claim	Intervention family / sampling rule	Certified quantity
Abstraction / reduction fidelity	Causal abstraction, interchange interventions, IIA [Geiger et al., 2021, 2025]	A high-level causal model preserves the intervention behavior of the original network.	Source input, donor input, and mask over abstract variables; Π samples masks and donors.	$F = 1 - \mathbb{E}[Z]$
Abstraction / reduction fidelity	Structured pruning, affine mechanism replacement, neural compression [Asiaee, 2026]	A reduced network or transformed mechanism preserves relevant causal behavior, not only ordinary predictions.	Coordinate, mechanism, or group interventions; Π samples masks, groups, donors, or perturbation strengths.	$F = 1 - \mathbb{E}[Z]$ or bounded risk
Abstraction / reduction fidelity	Causal scrubbing [Chan et al., 2022]	A hypothesis identifies which variables are causally relevant for behavior preservation.	Resampling interventions generated by the hypothesis; Π samples examples and resampling choices.	Behavior mismatch or degradation
Boundary case	Circuit completeness / circuit faithfulness [Wang et al., 2023, Hanna et al., 2024, Miller et al., 2024, Shi et al., 2024]	A selected circuit functions as a reduced mechanism for the behavior, or as a component set with high causal effect.	Restore, ablate, or compare selected nodes/edges under prompt distributions.	Fidelity F or recovery μ
Component effect / recovery	Activation patching, causal tracing [Vig et al., 2020, Meng et al., 2022, Zhang and Nanda, 2024]	A component causally supports a behavior because restoring it recovers the clean behavior.	Clean/corrupted prompt pair plus component set S; Π may be a point mass or a sampler over components.	$\mu = \mathbb{E}[\Delta]$
Component effect / recovery	Path patching [Goldowsky-Dill et al., 2023]	A directed path or edge set carries a causal effect between upstream and downstream components.	Clean/corrupted prompt pair plus path or edge set E; Π samples paths or fixes a discovered path set.	Recovery or clipped effect mean
Component effect / recovery	Ablation studies [Wang et al., 2023, Miller et al., 2024]	Removing or corrupting a component changes behavior, indicating causal relevance.	Node, head, edge, or feature set plus ablation baseline; Π samples components, prompts, or baselines.	Bounded degradation or effect
Discovery plus evaluation	ACDC, attribution patching, AtP*, related circuit-discovery methods [Conmy et al., 2023, Syed et al., 2024, Kramár et al., 2024, Hanna et al., 2024]	A search procedure proposes candidate components or circuits; a separate evaluation estimates their causal effect or fidelity.	Candidate-dependent proposals over nodes, edges, or paths; selection may require multiple-comparison correction.	Same CIF estimands after selection

Table 4: Extended mapping of common interventional interpretability and reduction evaluations into CIF. The common requirement is to declare the input distribution $\mathcal{D}$, intervention distribution Π, and bounded per-intervention score.

Method	k	$\hat{F}$	n ($F \geq 0.90$)		n ($F \geq 0.95$)	
			Hoeff.	Bet.	Hoeff.	Bet.
Soft interv.	32	0.824	—	—	—	—
Soft interv.	384	0.989	1,081	93	6,746	233
Hard pruning	32	0.102	—	—	—	—
Hard pruning	384	0.101	—	—	—	—
Variance-based	32	0.494	—	—	—	—
Variance-based	384	0.997	933	64	4,973	125
Random	32	0.196	—	—	—	—
Random	384	0.988	1,081	64	7,239	228
Identity ($k = 512$, all methods)	512	1.000	889	64	4,172	125

Table 5: Certification at additional pruning levels ($p = 0.1$, i.i.d., $1 - \delta = 0.95$). "—" = not certified within $n = 10{,}000$. At $k = 512$ nothing is pruned, so all methods reduce to the identity abstraction and share one row.

	i.i.d.	Static adaptive	History adaptive
Exclusion n (median, IQR)	**88** (86–150)	201 (151–219)	118 (101–129)
Certification n (median, IQR)	**1,217** (763–3,263)	1,708 (1,501–3,007)	3,800 (806–4,031)

Table 6: Matched five-seed comparison of samplers in the two E1 regimes (betting CS, $\alpha = 0.3$). Low-fidelity exclusion: soft interventions, $k = 256$, $p = 0.5$, certify $F < 0.70$. Borderline certification: variance-based pruning, $k = 128$, $p = 0.1$, certify $F \geq 0.90$.

Sampling regime	$1 - \delta$	Hoeffding		Betting	
		Coverage	Std	Coverage	Std
i.i.d. ($\alpha = 0$)	0.99	1.000	0.000	1.000	0.000
Adaptive ($\alpha = 0.3$)	0.99	1.000	0.000	1.000	0.000
Aggressive peeking	0.99	1.000	0.000	0.992	0.004
i.i.d. ($\alpha = 0$)	0.95	1.000	0.000	0.998	0.002
Adaptive ($\alpha = 0.3$)	0.95	1.000	0.000	1.000	0.000
Aggressive peeking	0.95	1.000	0.000	0.966	0.008
i.i.d. ($\alpha = 0$)	0.90	1.000	0.000	0.996	0.003
Adaptive ($\alpha = 0.3$)	0.90	1.000	0.000	0.996	0.003
Aggressive peeking	0.90	1.000	0.000	0.930	0.011

Table 7: Empirical coverage over 500 runs with $n = 2{,}000$ samples per run. Hoeffding is conservative in all configurations; betting remains at or above nominal coverage while producing tighter intervals.