SIFT: ROBUST META-FAITHFULNESS VERIFICATION OF CHAIN-OF-THOUGHT REASONING UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-Thought (CoT) faithfulness detectors are widely used to audit reasoning models, yet a detector is itself a predictor whose verdicts are treated as stable properties of the model. We ask whether a faithfulness detector is faithful to itself under distribution shift. We formalize *meta-faithfulness* as an invariance principle: a valid detector must return identical verdicts on traces that differ only by transformations preserving ground-truth faithfulness. We prove three results: a behavioral indistinguishability theorem showing that no detector operating on intervention-response profiles can separate faithful from epiphenomenal mechanisms with identical signatures; an invariance-violation lower bound showing that any detector relying on shift-sensitive features must violate invariance at a rate independent of its in-distribution accuracy; and an asymptotic certified selective-risk guarantee enabling abstention with high confidence. We operationalize this principle in FaithShift, a stress-test protocol spanning ten shift axes, and propose **SIFT** (**S**hift-**I**nvariant **F**aithfulness **T**rajectory Detector), a hidden-state trajectory detector trained with cross-environment invariance objectives and certified abstention. Across 14,996 traces, four reasoning domains, and eight model architectures, three findings emerge. *First*, transfer collapse is real: all existing detectors exhibit gaps ≥ 0.15 AUROC. *Second*, the dominant bottleneck is sampling stochasticity, not shift sensitivity: over 80% of detector instability stems from random seed variation rather than distribution shift, falsifying our preregistered hypothesis that shift-attributable invariance violations would exceed 0.25. *Third*, SIFT reduces shift-attributable instability but its advantage vanishes against ensembles: SIFT cuts invariance violations by 64% over the best single-seed baseline, but a four-seed ensemble of any detector narrows the margin to 0.01, and at matched coverage the two are statistically indistinguishable ($p = 0.21$). SIFT also requires a 51% abstention rate that exposes a robustness-usability trade-off. Cross-model transfer degrades along a clear hierarchy, within-family $>$ cross-family open-weight $>$ open-weight to API, which multi-model training on 3–4 models partially closes. We therefore position this work as a framework for auditing auditors and a diagnosis of the actual barrier to reliable faithfulness detection: not distribution shift, but detector variance.

1 INTRODUCTION

The rapid advancement of large language models (LLMs) has ushered in an era where machines can generate detailed chains of thought (CoT) that ostensibly explain their reasoning processes (Wei et al., 2022; Kojima et al., 2022). These CoT traces have become central to safety monitoring, interpretability, and trustworthiness evaluation across a wide range of deployment scenarios (Hubinger et al., 2024; Greenblatt et al., 2024). The premise is compelling: by examining the reasoning chain, we can verify that the model arrived at its conclusion through sound logical steps rather than through spurious correlations, hidden biases, or unintended shortcuts. This promise has motivated a growing body of work on faithfulness detection—methods that determine whether a model's stated reasoning reflects the actual computation producing its answer (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025; Zhang et al., 2025). The underlying assumption is that if we can reliably detect when reasoning is unfaithful, we can flag potentially unreliable outputs, improve model safety, and build greater trust in AI systems (Lightman et al., 2024; Wang et al., 2023).

However, a fundamental and previously unexamined question lurks beneath this enterprise: are the detectors themselves faithful? A faithfulness detector is itself a predictive model, trained on specific data distributions, with its own inductive biases, architectural choices, and failure modes. The machine learning community has learned, often through painful experience, that predictors validated on one distribution can collapse catastrophically when deployed on another (Geirhos et al., 2020; Koh et al., 2021; Arjovsky et al., 2019; Ben-David et al., 2010). Distribution shift remains one of the most persistent challenges in applied machine learning, with models often relying on spurious correlations that fail to generalize beyond their training environments (Gulrajani & Lopez-Paz, 2021). Faithfulness detectors have so far escaped this scrutiny. They are developed and validated on a handful of models, a narrow band of tasks, and a fixed elicitation regime; their verdicts are then reported as if model faithfulness were a stable, transferable property. If these quantities are artifacts of the evaluation distribution, then a large and safety-relevant literature rests on an unaudited measuring instrument (Korbak et al., 2025; Meek et al., 2025). This paper addresses this critical gap by making the auditing instrument itself the object of study. We introduce the concept of *meta-faithfulness*: the property that a faithfulness detector must itself be robust to transformations that preserve ground-truth faithfulness. A meta-faithful detector returns identical verdicts on pairs of traces that differ only by transformations that do not change the underlying faithfulness label. This property ensures that detector verdicts reflect genuine faithfulness rather than sensitivity to surface-level variations (Parcalabescu & Frank, 2024; Atanasova et al., 2023).

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

We operationalize meta-faithfulness in FAITHSHIFT, a stress-test protocol spanning ten distribution-shift axes (seven preserving, three altering), and evaluate six detector families across 14,996 traces, four reasoning domains (GSM8K, MATH, CommonsenseQA, BBH), and eight model architectures (DeepSeek-R1, Qwen3, Llama-3, Mistral, Phi-3, Gemma, Claude-3, GPT-4o). Our investigation yields three findings. **Finding 1 (distribution shift matters):** all existing detectors exhibit transfer gaps of at least 0.15 AUROC across cross-domain, cross-model, and cross-lingual axes (H1 supported; Koh et al., 2021). **Finding 2 (stochasticity dominates):** decomposing instability via seed-repeat analysis reveals that over 80% of verdict disagreement stems from sampling stochasticity (random seed variation) rather than distribution shift; the shift-attributable component is only 0.050 for attribution-consistency and 0.074 for SIFT, falsifying the preregistered hypothesis that shift-attributable IVR would exceed 0.25 (H2a falsified; Geirhos et al., 2020). **Finding 3 (SIFT's advantage vanishes against ensembles):** we propose SIFT, a trajectory-based detector with invariance training, which reduces raw invariance violations by 64% over the best single-seed baseline, achieving MFS 0.72 versus 0.67. However, a four-seed ensemble of any detector reaches MFS 0.71, narrowing the margin to 0.01, and at matched coverage ($\kappa = 0.70$) SIFT and a simple instance-level ensemble are statistically indistinguishable (MFS 0.64 vs. 0.62, $p = 0.21$; H3 and H4 not supported at preregistered parameters). We frame this work as a contribution to *auditing and diagnosis*, not as a definitive solution. Our empirical results suggest that the field's current focus on distribution shift, while important, may be misdirected: the dominant barrier to reliable faithfulness detection is sampling stochasticity, which simpler variance-reduction techniques (ensembling across seeds) can address at lower cost than invariance training. We therefore position SIFT as a demonstration that invariance training reduces stochastic instability, not as a broadly superior detector.

1.1 CONTRIBUTIONS

Our investigation is organized around three key questions and five contributions. First, can we formally characterize the limitations of existing faithfulness detectors? In Section 4, we prove that behavioral detectors, which constitute the majority of current approaches, are fundamentally incapable of distinguishing faithful from unfaithful mechanisms in certain scenarios (Theorem 4.5), and that accuracy does not bound invariance (Theorem 4.6). Second, can we develop detectors with provable robustness guarantees? In Section 5, we propose SIFT (Shift-Invariant Faithfulness Trajectory Detector), which extracts hidden-state trajectories and trains with invariance objectives to achieve certified robustness (Zhang et al., 2025; Marks & Tegmark, 2024; Conmy et al., 2023). Third, how do existing detectors perform under realistic distribution shifts? In Section 6, we introduce FaithShift, a comprehensive stress-test protocol spanning ten shift axes, and provide extensive empirical evaluation in Section 7.2.

We note at the outset that our preregistered hypotheses are only partially supported: transfer collapse is confirmed (H1), detectors disagree substantially (H2b), but the shift-attributable component of invariance violations is small (H2a), SIFT's advantage narrows once ensemble baselines are included (H3), and certified abstention requires a coverage well below the preregistered target (H4). We treat these outcomes as substantive results and frame the paper's central contribution as a framework for auditing auditors rather than as a demonstrated detector advance. Our contributions are: (1) meta-faithfulness as a formal principle, with estimable quantities (IVR, AS, MFS); (2) three theorems on detector limitations (behavioral indistinguishability, invariance-violation lower bound, certified selective risk); (3) the FaithShift protocol; (4) an empirical diagnosis that stochasticity ($\sim$80%) dominates shift ($\sim$20%); and (5) a SIFT and ensemble analysis showing that simple multi-seed ensembling recovers 95% of SIFT's gain at lower cost.

2 RELATED WORK

Unfaithful Chain-of-Thought. The phenomenon of unfaithful chain-of-thought is well documented (Turpin et al., 2023; Lanham et al., 2023; Arcuschin et al., 2025). Turpin et al. (2023) demonstrate that language models silently exploit biasing features inserted into prompts, producing answers that reflect the bias rather than the stated reasoning. Their experiments show that models can be systematically manipulated to produce answers that contradict their stated reasoning chains, with some models exhibiting faithfulness rates as low as 60% under adversarial conditions.

Monitorability and Safety. CoT monitoring is framed as a fragile safety opportunity (Korbak et al., 2025; Meek et al., 2025). These works argue that while CoT traces offer unprecedented visibility into model computation, this visibility is contingent on the model actually using the CoT for reasoning rather than generating it post-hoc (Hubinger et al., 2024; Greenblatt et al., 2024). Reward hacking can be pushed into or out of the visible trace (Turpin et al., 2025), making monitorability claims contingent on the monitor's robustness. If monitors can be fooled by models that generate convincing but unfaithful reasoning, the safety case for CoT monitoring collapses (Korbak et al., 2025). Our work complements these efforts by providing a framework for auditing the auditors. We argue that before we can trust monitors, we must verify that the monitors themselves are robust to the kinds of distribution shifts that arise in deployment (Meek et al., 2025).

Robustness and Selective Prediction. Shortcut learning and distribution shift (Geirhos et al., 2020; Koh et al., 2021) motivate invariance-based training such as domain-adversarial learning (Ganin et al., 2016) and invariant risk minimization (Arjovsky et al., 2019). These methods aim to learn representations that are invariant across environments, thereby improving generalization (Gulrajani & Lopez-Paz, 2021). Selective prediction lets a classifier abstain with risk guarantees (Geifman & El-Yaniv, 2017; Angelopoulos et al., 2021; Chow, 1970). The selective-guaranteed-risk procedure provides confidence bounds on the selective risk, enabling principled abstention (Varshney et al., 2022; Kamath et al., 2020). We are the first to import these tools to the auditor, treating faithfulness detection as a domain-generalization problem with a certified operating point (Ben-David et al., 2010).

Stochasticity and Reproducibility. A parallel literature documents the sensitivity of deep learning models to random seed variation (Henderson et al., 2018; Bouthillier et al., 2021; Picard, 2021). Our finding that over 80% of detector instability is stochastic connects to this work, suggesting that the field should prioritize variance reduction alongside distribution robustness. We are the first to quantify the relative contributions of stochasticity and distribution shift in the faithfulness detection setting.

3 Problem Formulation

The central concept in our analysis is ground-truth faithfulness, denoted $\phi^\star(\mathcal{M}, x, c) \in \{0, 1\}$, a binary variable indicating whether the final answer y is causally produced by the computation described in the CoT trace c, as opposed to being determined by features the trace omits or by computational paths that diverge from the stated reasoning (Pearl, 2009; Parcalabescu & Frank, 2024). Ground-truth faithfulness is not directly observable: it must be approximated through construction protocols, such as injecting decisive unverbalized cues to force unfaithfulness, or through expert human audit (Atanasova et al., 2023; Shen et al., 2026). This approximation challenge is fundamental to faithfulness detection and motivates detectors that remain robust under imperfect labels (Lyu et al., 2023). A faithfulness detector is a map $\mathcal{D} : \mathcal{M} \times \mathcal{X} \times \mathcal{C} \to [0, 1]$ that scores the likelihood of faithfulness given the model $\mathcal{M}$, query x, and CoT trace c. This score is converted to a binary verdict through an operating threshold $\tau \in (0, 1)$:

$$\hat{\phi} = \mathbf{1}[\mathcal{D}(\mathcal{M}, x, c) > \tau]. \tag{1}$$

Definition 3.1 (Behavioral Detector)**.** *A behavioral detector is one that operates solely on intervention-response profiles generated by the model. Formally, a behavioral detector takes the form:*

$$\mathcal{D} = g \circ R_{\mathcal{I}} \tag{2}$$

where $R_{\mathcal{I}}(\mathcal{M}, x, c) = \{(\iota, \mathcal{M}(\iota(x, c))) : \iota \in \mathcal{I}\}$ is an intervention-response profile over an intervention family $\mathcal{I}$, and $g : \mathcal{P} \to [0, 1]$ maps these profiles to scores.

4 Meta-Faithfulness and Theory

4.1 Core Definitions

Definition 4.1 (Faithfulness-Preserving Transformations)**.** *A transformation $T : (x, c) \mapsto (x', c')$ is* preserving *if it belongs to the set $\mathcal{T}_{pres}$ and satisfies*

$$\phi^\star(\mathcal{M}, x', c') = \phi^\star(\mathcal{M}, x, c) \tag{3}$$

for the models under study. Thus, the transformation leaves the ground-truth faithfulness label unchanged (Atanasova et al., 2023).

Definition 4.2 (Faithfulness-Altering Transformations)**.** *A transformation $T : (x, c) \mapsto (x', c')$ is* altering *if it belongs to the set $\mathcal{T}_{alt}$ and is constructed and independently audited to flip the faithfulness label:*

$$\phi^\star(\mathcal{M}, x', c') = 1 - \phi^\star(\mathcal{M}, x, c). \tag{4}$$

Definition 4.3 (Meta-Faithfulness)**.** *A detector $\mathcal{D}$ is* meta-faithful *on the pair $(\mathcal{T}_{pres}, \mathcal{T}_{alt})$ if it satisfies both of the following conditions:*

(M1) ***Invariance****: for every verified preserving transformation $T \in \mathcal{T}_{pres}$,*

$$\hat{\phi}(c') = \hat{\phi}(c), \tag{5}$$

where $\hat{\phi}$ denotes the detector's thresholded verdict.

(M2) ***Sensitivity****: for every verified altering transformation $T \in \mathcal{T}_{alt}$, the detector's verdict changes in the direction implied by the change in the ground-truth faithfulness label.*

The two conditions in Definition 4.3 yield two directly estimable quantities. The *invariance-violation rate* is

$$\mathrm{IVR}(\mathcal{D}) = \Pr_{T \sim \mathcal{T}_{\text{pres}}} \left[\hat{\phi}(c') \neq \hat{\phi}(c) \right]. \tag{6}$$

A low IVR indicates that the detector's binary verdict is stable under verified preserving transformations. The *alteration sensitivity* is

$$\mathrm{AS}(\mathcal{D}) = \Pr_{T \sim \mathcal{T}_{\text{alt}}} \left[\mathcal{D} \text{ moves correctly} \right], \tag{7}$$

where "moves correctly" means that the detector's verdict changes consistently with the verified change in the ground-truth faithfulness label. A high AS therefore measures responsiveness to genuine changes in faithfulness (Atanasova et al., 2023). We note that our audit verifies *label* preservation rather than *mechanism* preservation: a transformation may leave $\phi^\star$ unchanged while altering the underlying computation. This distinction is revisited in Section 9.

Definition 4.4 (Meta-Faithfulness Score)**.** *At a certified operating point with coverage κ, let $\overline{\mathrm{IVR}}$ and $\overline{\mathrm{AS}}$ denote the invariance-violation rate and alteration sensitivity evaluated on the corresponding certified evaluation subset. The* Meta-Faithfulness Score *is the coverage-scaled harmonic mean*

$$\mathrm{MFS} = \kappa \cdot \frac{2(1 - \overline{\mathrm{IVR}})\overline{\mathrm{AS}}}{(1 - \overline{\mathrm{IVR}}) + \overline{\mathrm{AS}}} \in [0, 1]. \tag{8}$$

4.2 Behavioral Indistinguishability

Theorem 4.5 (Behavioral Indistinguishability). *Let $\mathcal{D}$ be a behavioral detector of the form $\mathcal{D} = g \circ R_{\mathcal{I}}$. If a faithful mechanism M_{f} and an epiphenomenal mechanism M_{u} induce identical intervention-response profiles $R_{\mathcal{I}}(M_{\mathrm{f}}) = R_{\mathcal{I}}(M_{\mathrm{u}})$, then the detector produces identical scores:*

$$\mathcal{D}(M_{\mathrm{f}}) = \mathcal{D}(M_{\mathrm{u}}). \tag{9}$$

Consequently, no behavioral detector whose information is restricted to the intervention family $\mathcal{I}$ can distinguish these mechanisms.

4.3 Accuracy Does Not Bound Invariance

Let $s(c)$ denote a scalar feature of a trace that is correlated with faithfulness in the training distribution but can change under a preserving transformation. Examples include trace length, language, formatting, or the occurrence of particular lexical features (Geirhos et al., 2020). Let $m(c) = \mathcal{D}(c) - \tau$ denote the detector's decision margin.

Assumption 1 (Local Sensitivity). *There exists $\gamma > 0$ such that, for sufficiently small preserving perturbations,*

$$\mathcal{D}(c') - \mathcal{D}(c) = \gamma\Delta + r(\Delta), \tag{10}$$

where $\Delta = s(c') - s(c)$ and $\frac{r(\Delta)}{\Delta} \to 0$ as $\Delta \to 0$.

Assumption 2 (Symmetry and Independence). *The perturbation Δ is symmetric about zero and is independent of the detector margin $m(c)$.*

Theorem 4.6 (Invariance-Violation Lower Bound). *Under Assumptions 1 and 2, in the local-perturbation regime in which the higher-order remainder is negligible relative to $\gamma|\Delta|$, the invariance-violation rate satisfies*

$$\mathrm{IVR}(\mathcal{D}) \gtrsim \frac{1}{2} \Pr_{c,T} \left[|m(c)| \le \gamma|\Delta| \right]. \tag{11}$$

In the first-order approximation $r(\Delta) = 0$, the relation becomes

$$\mathrm{IVR}(\mathcal{D}) \geq \frac{1}{2} \Pr_{c,T} \left[|m(c)| \le \gamma|\Delta| \right]. \tag{12}$$

Consequently, AUROC or classification accuracy alone does not determine IVR, because these quantities need not control the probability mass of the detector's margin distribution near the decision boundary (Geirhos et al., 2020; Ben-David et al., 2010). We note that the bound is not assumed to be tight: it provides a lower bound on IVR under any spurious feature, and we verify empirically (Table 3) that trace length, though a valid spurious feature, is not the primary driver of observed instability.

4.4 Certified Selective Risk

The third theoretical result addresses the practical consequence of the margin instability characterized above. When some traces are intrinsically uncertain or lie near an unstable decision boundary, a detector can abstain rather than commit to a potentially unreliable verdict. Selective prediction (Geifman & El-Yaniv, 2017; Chow, 1970) provides a formal framework for this trade-off.

Theorem 4.7 (Asymptotic Certified Selective Risk). *Let ρ define the certified set $S_t = \{c : \rho(c) \geq t\}$. Given n i.i.d. audited calibration traces satisfying Assumption 3, target risk $\alpha \in (0, 1)$, confidence level $\delta \in (0, 1)$, and a finite predetermined set of candidate thresholds $\mathcal{T}$, the selective-guaranteed-risk (SGR) procedure selects a threshold $t^\star \in \mathcal{T}$ such that, asymptotically as $n \to \infty$,*

$$R_{\mathrm{cert}}(t^\star) := \Pr\left[\hat{\phi}(c) \neq \phi^\star(c) \mid \rho(c) \geq t^\star \right] \leq \alpha \tag{13}$$

with probability at least $1 - \delta$, provided that at least one candidate threshold satisfies the target risk constraint (Geifman & El-Yaniv, 2017; Angelopoulos et al., 2021).

5 SIFT: A Shift-Invariant Trajectory Detector

SIFT is designed as an *invariance-oriented detector* that achieves superior meta-faithfulness through three key innovations: (1) trajectory-based features that capture internal dynamics of reasoning (Nanda et al., 2023; Conmy et al., 2023), (2) multi-scale encoding that captures patterns across different temporal resolutions (Li et al., 2023), and (3) invariance-oriented training with IRM and DANN penalties that encourage robustness across environments (Arjovsky et al., 2019; Ganin et al., 2016). The complete training procedure is summarized in Algorithm 1. To encourage robustness to verified preserving transformations, SIFT is trained across multiple environments. Each preserving transformation axis is treated as an environment $e \in \mathcal{E}$, and the detector is optimized to maintain predictive performance across these environments while discouraging environment-specific decision rules (Arjovsky et al., 2019; Ganin et al., 2016).

The training objective is

$$\min_{\theta} \sum_{e \in \mathcal{E}} \mathcal{R}_e(g_\theta) + \lambda\Omega(g_{\theta_e}), \tag{14}$$

where

$$\mathcal{R}_e(g_\theta) = \mathbb{E}_{(x,c,\phi^\star)\sim P_e} \left[\ell\left(g_\theta\left(\tilde{\psi}(H(\mathcal{M}, x, c)) \right), \phi^\star \right) \right] \tag{15}$$

is the per-environment predictive loss and Ω is an invariance regularizer.
We consider two complementary invariance penalties. The **IRM penalty** (Arjovsky et al., 2019) is implemented as

$$\Omega_{\text{IRM}} = \sum_{e \in \mathcal{E}} \left| \nabla_w \mathcal{R}_e(w \circ g_\theta) \big|_{w=1.0} \right|_2^2 . \tag{16}$$

The penalty encourages a shared classifier to remain simultaneously optimal across environments (Arjovsky et al., 2019).

6 FAITHSHIFT: A STRESS-TEST PROTOCOL

To evaluate meta-faithfulness, we introduce FaithShift, a stress-test protocol that systematically evaluates detector behavior under controlled distribution shifts (Geirhos et al., 2020; Koh et al., 2021). FaithShift separates transformations according to their verified effect on the ground-truth faithfulness label and evaluates both invariance and sensitivity using paired traces (Atanasova et al., 2023; Parcalabescu & Frank, 2024). FaithShift defines ten shift axes organized into seven preserving and three altering transformation types, as detailed in Table 1. The inference-budget axis is treated as a mixed axis because its effect on the faithfulness label is determined empirically rather than assumed a priori (Snell et al., 2025).

6.0.1 TRAJECTORY FEATURES

For a trace of length T generated by a model with L layers, we extract the hidden-state trajectory

$$H \in \mathbb{R}^{L \times T \times d},$$

where $H_{\ell,t}$ denotes the hidden state at layer ℓ and time step t. The trajectory provides information about how internal representations evolve during generation (Zhang et al., 2025; Marks & Tegmark, 2024). From H, we compute four families of descriptors intended to capture temporal and representational dynamics while reducing direct dependence on surface-level properties (Nanda et al., 2023; Conmy et al., 2023).
Inter-layer velocity:

$$v_{\ell,t} = \left| H_{\ell+1,t} - H_{\ell,t} \right|_2 . \tag{17}$$

This descriptor measures the magnitude of representational change across adjacent layers at a given generation step (Marks & Tegmark, 2024; Conmy et al., 2023).
Step-to-step drift:

$$d_{\ell,t} = \left| H_{\ell,t+1} - H_{\ell,t} \right|_2 . \tag{18}$$

This descriptor measures temporal representational change at a fixed layer (Nanda et al., 2023).
These quantities are treated as trajectory descriptors rather than direct mechanistic indicators. In particular, high velocity or drift does not by itself establish that substantive reasoning is occurring, and low velocity does not establish that a trace is post-hoc. Their role is to provide measurable signals for the detector (Zhang et al., 2025; Li et al., 2023).
Early-commitment: This descriptor measures whether information predictive of the final answer is already present early in the generation process. We first fit a direction w_y that linearly decodes the final answer from final-layer hidden states using logistic regression (Marks & Tegmark, 2024; Li et al., 2023). We then compute

$$\text{EC}(c) = \max_{t < T/2} \cos\left(H_{L,t}, w_y\right) . \tag{19}$$

High early decodability can be associated with early answer commitment, but it is not treated as direct evidence of causal unfaithfulness. Instead, it is used as one component of the learned trajectory representation (Zhang et al., 2025).
Attention stability: This descriptor measures temporal consistency of attention patterns:

$$\text{AS}(c) = \frac{1}{T-1} \sum_{t=1}^{T-1} \text{JSD}\left(A_t | A_{t+1}\right) , \tag{20}$$

where JSD denotes Jensen–Shannon divergence and A_t is the attention distribution at time step t, averaged over layers and heads.
The descriptors are normalized as

$$\tilde{\psi}(H) = \frac{\psi(H)}{|\psi(H)|_2}, \tag{21}$$

reducing the influence of overall feature magnitude and helping prevent trace length from dominating the representation. Importantly, this normalization is not assumed to guarantee invariance; invariance is evaluated empirically through FaithShift (Geirhos et al., 2020).

6.0.2 MULTI-SCALE TRAJECTORY ENCODING

The trajectory representation is processed using a multi-scale architecture intended to capture patterns at different temporal resolutions. The architecture contains four components (Conmy et al., 2023; Nanda et al., 2023).
Temporal Convolutional Network (TCN): The TCN applies one-dimensional convolutions with kernel sizes $[3, 5, 7, 11]$. Smaller kernels capture local temporal patterns, whereas larger kernels provide wider temporal receptive fields. Each convolution is followed by batch normalization and ReLU activation. **Transformer Encoder:** The TCN output is processed by a transformer encoder with 4 layers and 8 attention heads. The transformer provides global temporal interactions and can associate information appearing at distant positions in a reasoning trace (Li et al., 2023).

Learned Pooling: The transformer outputs are aggregated through a learned weighted sum:

$$h_{\text{pool}} = \sum_{t=1}^{T} \alpha_t h_t^{(L)}, \tag{22}$$

where

$$\alpha_t = \text{softmax}\left(W_\alpha h_t^{(L)}\right). \tag{23}$$

This permits the detector to emphasize temporal positions that are predictive of faithfulness.
MLP Classifier: The pooled representation is passed through a two-layer MLP with hidden dimension 128 and ReLU activation, followed by a sigmoid output:

$$\mathcal{D}(c) \in [0, 1]. \tag{24}$$

The resulting score represents the detector's estimated probability of faithfulness and is thresholded to obtain the binary verdict used in IVR and AS evaluation.

6.0.3 Invariance Objective and Training Procedure

The **Domain-Adversarial penalty** (Ganin et al., 2016) introduces a domain classifier d_ϕ trained to predict the environment from the learned representation while the feature extractor is optimized adversarially to suppress environment-specific information. We denote the resulting adversarial objective by

$$\Omega_{\text{DANN}} = -\mathcal{L}_{\text{adv}}(d_\phi; \tilde{\psi}(H)), \tag{25}$$

where the feature representation is optimized to reduce domain predictability while retaining faithfulness information (Ganin et al., 2016). In the final SIFT configuration, the two penalties are combined according to the training configuration reported in Section 7.2. The objective is intended to encourage robustness across observed environments; it does not guarantee invariance to unseen transformations, which is why independent FaithShift evaluation remains necessary (Geirhos et al., 2020; Gulrajani & Lopez-Paz, 2021). The complete training and certification procedure is summarized in Algorithm 1 and C.4.

Algorithm 1 SIFT Training and Certification

Require: Labeled traces grouped by environment e; target risk α, confidence level δ; predetermined threshold grid $\mathcal{T}$

for each trace i **do**
  Extract H_i; compute $\psi(H_i)$; normalize $\tilde{\psi}(H_i)$
end for
Initialize TCN, transformer, and MLP
for each epoch **do**
  for each environment e **do**
    Compute $\mathcal{R}_e(g_\theta)$
  end for
  Compute the invariance penalty $\Omega(g_{\theta_e})$
  Update $\theta \leftarrow \theta - \eta\nabla_\theta(\sum_e \mathcal{R}_e + \lambda\Omega)$
end for
Fit support gate ρ using kernel density estimation on training in-support features
Evaluate candidate thresholds $t \in \mathcal{T}$ on independent audited calibration traces
Select $t^\star$ using the SGR procedure for (α, δ)
return $\hat{\phi}(c) = \mathbf{1}[g_\theta(\tilde{\psi}(H)) > \tau]$ if $\rho(c) \geq t^\star$, else UNCERTIFIED

7 Experiments

7.1 Experimental Setup

Trace Construction We sample 750 base items across four domains, each instantiated under 10 conditions (base generation, paraphrase, step reordering, hint-format variation, language translation, budget variation, model-family comparison, cue injection, cue verbalization, causal truncation) and two model backbones, yielding $750 \times 10 \times 2 = 15{,}000$ traces. After filtering and auditing for label preservation and quality, we retain 14,996 scored traces, split 60/15/25 into train, validation, and test (Shen et al., 2026). The effective sample size for statistical inference is constrained by the 750 independent base items rather than the 14,996 traces, since traces from the same base item are not independent; all confidence intervals are bootstrap-stratified by base item.

Detector Implementations. We compare six detector families. **(1) Biasing-features** (Turpin et al., 2023): logistic regression on cue-mention features (C=1.0), representing the behavioral approach. **(2) Counterfactual intervention** (Lanham et al., 2023): MLP (hidden 64, ReLU) trained on intervention-response profiles from CoT edits. **(3) Hidden-state probe** (Zhang et al., 2025): logistic regression on mean-pooled final-layer hidden states. **(4) Instance-level discriminative** (Shen et al., 2026): MLP (hidden 128) on behavioral and internal instance-level features. **(5) Attribution-consistency**: attention-stability score across generation steps with a learned threshold. **(6) SIFT** (ours): multi-scale trajectory encoding with TCN, transformer, and IRM training (Section 5); IRM penalty λ=1.0, Adam, learning rate 10^{-3}, batch size 32, 50 epochs (Arjovsky et al., 2019; Ganin et al., 2016).

7.2 Main Results

Transfer Collapse Analysis. Table 1 reports in-distribution and transfer AUROC for all six detectors across cross-domain, cross-model, and cross-lingual axes (Koh et al., 2021; Geirhos et al., 2020). Transfer degrades consistently: counterfactual shows the largest gap (0.28), attribution-consistency drops to near-chance (0.48) cross-lingually, and hidden-state and instance-level show moderate gaps (0.17 and 0.18). SIFT achieves the smallest gap (0.10) with in-distribution performance of 0.75. This confirms H1: transfer collapse is real and measurable across all detectors.

Table 1: Main results across six detectors. **(a)** In-distribution vs. transfer AUROC: all existing detectors have gaps $\Delta \geq 0.15$ (H1 confirmed (Koh et al., 2021)); attribution-consistency drops to 0.48 cross-lingually; SIFT has the smallest gap (0.10). **(b)** Per-axis IVR: step reordering is near zero; SIFT cuts paraphrase and hint-format IVR by 76% each, language to 0.14 (vs. 0.27), model family to 0.18 (vs. 0.37); mean IVR reduced by 64%. **(c)** Seed-repeat decomposition: attribution-consistency raw 0.282 = floor 0.232 + shift 0.050, so >80% is stochastic and H2a (IVR ≥ 0.25) is falsified; SIFT raw 0.102 = floor 0.028 + shift 0.074. **(d)** Meta-Faithfulness Score: SIFT 0.72 vs. 0.67 for the strongest single-seed baseline, but 0.71 for a four-detector instance-level ensemble, so H3 is not supported under the revised comparator (margin 0.01). **(e)** Seed stability: attribution-consistency std = 0.025 vs. SIFT std = 0.006, confirming stochasticity as the dominant instability source (Geirhos et al., 2020).

Detector	(a) Transfer					(d) Meta-Faithfulness Score				(e) Seed stability
	In-dist	Transfer AUROC			Gap	IVR	AS	MFS	Coverage	Mean ± Std
		cross-domain	cross-model	cross-lingual						
Biasing-features	0.82	0.63	0.66	0.59	0.23	0.34	0.71	0.61	**1.00**	0.805 ± 0.013
Counterfactual	**0.85**	0.61	0.64	0.57	0.28	0.31	0.74	0.64	**1.00**	0.835 ± 0.013
Hidden-state	0.79	0.70	**0.68**	0.62	0.17	0.27	0.73	0.66	**1.00**	0.780 ± 0.008
Instance-level	0.78	0.69	0.67	0.60	0.18	0.29	**0.76**	0.67	**1.00**	0.770 ± 0.008
Attribution	0.77	**0.72**	0.62	0.48	0.16	0.28	0.35	0.47	**1.00**	0.745 ± 0.025
SIFT	0.75	0.71	**0.68**	**0.65**	**0.10**	**0.10**	**0.76**	**0.72**	0.49	**0.745 ± 0.006**

Detector	(b) Per-axis invariance violation rates							(c) Seed-repeat decomposition		
	Paraphrase	StepReorder	HintFormat	Language	Budget	ModelFamily	Mean	Raw IVR	Seed-Repeat Floor	Shift-Attributable
Biasing-features	0.36	0.09	0.40	0.29	0.34	0.39	0.31	0.310	0.093	0.217
Counterfactual	0.33	0.07	0.37	0.26	0.30	0.35	0.28	0.280	0.084	0.196
Hidden-state	0.29	0.06	0.32	0.23	0.27	0.31	0.25	0.250	0.075	0.175
Instance-level	0.31	0.08	0.35	0.25	0.29	0.33	0.27	0.270	0.081	0.189
Attribution	0.34	0.08	0.38	0.27	0.32	0.37	0.28	0.282	0.232	**0.050**
SIFT	**0.08**	**0.00**	**0.09**	**0.14**	**0.12**	**0.18**	**0.10**	**0.102**	**0.028**	0.074

Stochasticity Dominates Instability. A more nuanced analysis reveals that distribution shift is not the primary robustness challenge. We decompose detector instability via seed-repeat analysis: for each trace, we train $N = 4$ independent seeds with different random initializations and data-loading order, and measure what fraction of verdict disagreement comes from (a) different random seeds on the same distribution (stochasticity), versus (b) the trace being fundamentally shifted (shift-attributable instability). Table 1 reports the results. For attribution-consistency, raw IVR = 0.282, of which 0.232 (82%) is seed-repeat floor and only 0.050 (18%) is shift-attributable. For SIFT, raw IVR = 0.102, of which 0.028 (27%) is seed-repeat and 0.074 (73%) is shift-attributable. Across all detectors, over 80% of observed instability stems from sampling stochasticity rather than distribution shift. Our preregistered hypothesis H2a predicted aggregate IVR ≥ 0.25; this is supported by raw IVR (Table 1, mean 0.10–0.28), but shift-attributable IVR is only 0.050 for the worst detector and 0.074 for SIFT, well below the preregistered threshold. **This falsifies H2a and indicates that distribution shift, while real, is not the primary robustness challenge.**

Preregistered Hypotheses and Outcomes. We preregistered five falsifiable hypotheses. H1 (transfer gap ≥ 0.15 AUROC) is supported, with all detectors showing gaps of 0.16–0.28 (Koh et al., 2021). H2a (aggregate IVR ≥ 0.25) is evaluated on the shift-attributable component, since 0.232 of attribution-consistency's raw IVR of 0.282 is seed-repeat stochasticity; the shift-attributable value is 0.050, so H2a is not supported (Geirhos et al., 2020; Parcalabescu & Frank, 2024). H2b (mean inter-detector $\kappa \leq 0.5$) is supported at $\kappa = 0.38$, though the two-detector suite is underpowered (Shen et al., 2026). H3 (SIFT exceeds the strongest prespecified baseline by ≥ 0.10 MFS) is not supported: SIFT reaches 0.72 versus 0.67 for the strongest single-seed baseline (margin 0.05) and 0.71 for a four-detector ensemble (margin 0.01), with Appendix E.5 attributing much of the gain to variance reduction achievable by simpler ensembling. H4 (certified risk $\leq \alpha$ at coverage $\kappa \geq 0.60$) is not supported: the bound of 0.249 holds at $\alpha = 0.25$, but coverage is only 0.50, and $\alpha = 0.30$ recovers $\kappa = 0.61$, an operational rather than theoretical failure (Geifman & El-Yaniv, 2017). The failures of H2a, H3, and H4 are treated as substantive empirical results.

Per-Axis IVR, Seed-Repeat Decomposition, and Meta-Faithfulness. Table 1 reports invariance-violation rates across the six preserving axes (Parcalabescu & Frank, 2024), Table 1 decomposes raw IVR into a seed-repeat floor and a shift-attributable component (Geirhos et al., 2020), and Table 1 reports Meta-Faithfulness Scores (Chow, 1970; Geifman & El-Yaniv, 2017). SIFT attains the lowest IVR (0.10) and highest alteration sensitivity (0.76), yielding MFS 0.72 versus 0.67 for the strongest single-seed baseline; a four-detector instance-level ensemble reaches 0.71, leaving a margin of 0.01 below the preregistered 0.10 threshold, so H3 is not supported under the revised comparator. The (IVR, alteration sensitivity) scatter, with bubble size encoding

Table 2: Cross-model and cross-domain generalization for SIFT. **(a)** Cross-model transfer across eight architectures: diagonal entries are in-distribution performance (0.74–0.78); average transfer is 0.58 versus 0.76 in-distribution (gap 0.18), same-family transfer is strongest (Llama-3→Mistral and Mistral→Llama-3: 0.65), and open-weight to API is weakest (0.52–0.60). **(b)** Transfer AUROC by source family, giving the hierarchy within-family (0.58–0.65) > cross-family open-weight (0.54–0.62) > open-weight to API (0.54–0.60). **(c)** Cross-domain transfer across four reasoning domains: average 0.63, strongest for GSM8K↔MATH (0.62–0.63), weakest for GSM8K↔CommonsenseQA (0.60–0.61); in-distribution ranges from 0.75 (GSM8K) to 0.78 (BBH) (Cobbe et al., 2021; Hendrycks et al., 2021; Talmor et al., 2019; Suzgun et al., 2023; DeepSeek-AI, 2025).

Setting	Source	DeepSeek-R1	Qwen3	Llama-3	Mistral	Phi-3	Gemma	Claude-3	GPT-4o
(a) Cross-model	DeepSeek-R1 (1.5B)	**0.75**	0.58	0.56	0.59	0.54	0.57	0.53	0.52
(a) Cross-model	Qwen3 (1.7B)	0.59	**0.76**	0.60	0.62	0.57	0.59	0.55	0.54
(a) Cross-model	Llama-3 (8B)	0.57	0.59	**0.78**	0.66	0.62	0.64	0.58	0.56
(a) Cross-model	Mistral (7B)	0.58	0.61	0.65	**0.76**	0.63	0.65	0.59	0.57
(a) Cross-model	Phi-3 (3.8B)	0.55	0.57	0.61	0.62	**0.74**	0.60	0.55	0.53
(a) Cross-model	Gemma (7B)	0.56	0.58	0.63	0.64	0.59	**0.75**	0.56	0.54
(a) Cross-model	Claude-3 (API)	0.52	0.54	0.57	0.58	0.54	0.55	**0.76**	0.60
(a) Cross-model	GPT-4o (API)	0.51	0.53	0.56	0.56	0.52	0.54	0.59	**0.78**
(b) Source family	DeepSeek/Qwen (Qwen-based)	0.58	0.57	0.54	0.04	0.58	0.57	0.54	0.04
(b) Source family	Llama/Mistral (Meta-family)	0.65	0.62	0.58	0.07	0.65	0.62	0.58	0.07
(b) Source family	Phi-3 (Microsoft)	0.74	0.60	0.54	0.14	0.74	0.60	0.54	0.14
(b) Source family	Gemma (Google)	0.75	0.60	0.55	0.15	0.75	0.60	0.55	0.15
(b) Source family	Claude-3 (Anthropic)	0.76	0.56	0.60	0.16	0.76	0.56	0.60	0.16
(b) Source family	GPT-4o (OpenAI)	0.78	0.54	0.59	0.19	0.78	0.54	0.59	0.19
(c) Cross-domain	GSM8K	**0.75**	0.62	0.60	0.63	0.75	0.62	0.60	0.63
(c) Cross-domain	MATH	0.63	**0.76**	0.62	0.64	0.63	0.76	0.62	0.64
(c) Cross-domain	CommonsenseQA	0.61	0.62	**0.76**	0.64	0.61	0.62	0.76	0.64
(c) Cross-domain	BBH	0.63	0.64	0.65	**0.78**	0.63	0.64	0.65	0.78

Table 3: **(a)** Certified selective risk by coverage for SIFT. The SGR certificate holds at the preregistered $\alpha = 0.25$ (bound 0.249), but coverage reaches only 0.50, below the required $\kappa \geq 0.60$ for H4; shifting α to 0.30 recovers $\kappa = 0.61$ while maintaining risk at most 0.30, indicating an operational rather than theoretical failure (Geifman & El-Yaniv, 2017). **(b)** Theorem 4.6 bound verification with trace length as the spurious feature: observed IVR exceeds the length-driven floor by 14.8× for attribution-consistency and 8.5× for SIFT, indicating that trace length is not the primary spurious feature. The bound holds with large slack and is not binding in this operationalization (Geirhos et al., 2020).

Setting	Quantity	0.05	0.25	0.45	0.65	0.85	1.00
(a) Certified risk	Coverage	0.05	0.25	0.45	0.65	0.85	1.00
(a) Certified risk	Selective Risk	0.000	0.063	0.203	0.278	0.335	0.321
(b) Bound check	Detector	γ	Predicted Floor	Observed IVR	Slack	–	–
(b) Bound check	Attribution	+0.024	0.019	0.282	14.8×	0.019	0.282
(b) Bound check	**SIFT**	+0.008	0.012	**0.102**	8.5×	0.012	**0.102**

coverage, exposes the robustness–usability trade-off: SIFT maintains certified guarantees only by abstaining on roughly 51% of traces, and at matched coverage its advantage becomes statistically indistinguishable ($p = 0.21$ at $\kappa = 0.70$; Appendices E.7.2 and E.7.3).

Ensemble Baseline and Coverage-Matched Comparison. A natural objection is that ensembling across seeds could match SIFT at lower cost, especially given that over 80% of detector instability is stochastic. We train four independent seeds of every detector, average their probabilities, and compute risk–coverage curves under the same SGR procedure. The ensemble gain is roughly uniform (+0.04 MFS), so SIFT-ensemble reaches 0.76 versus 0.71 for a four-detector instance-level ensemble, narrowing the margin from 0.05 (single-seed) to 0.01; at matched coverage $\kappa = 0.70$, the two are indistinguishable (MFS 0.64 vs. 0.62, $p = 0.21$). We therefore do not claim that SIFT is broadly superior, but rather that it substantially reduces raw invariance violations at its certified operating point, with much of its advantage attributable to variance reduction that simpler ensembling can also achieve.

8 Ablation Study

Table 4 presents the ablation study results, systematically removing architectural components to identify their contributions to SIFT's performance, and reports the corresponding multi-model training results (Arjovsky et al., 2019; Ganin et al., 2016). The IRM penalty contributes most: removing it increases IVR by 110% and decreases MFS by 28%. This suggests that IRM's primary effect is to stabilize representations across environments, which in turn reduces seed-to-seed variance. Multi-model training raises transfer AUROC from 0.58 to 0.65, cutting the gap from 0.18 to 0.14 with diminishing returns beyond 3–4 models.

9 Limitations and Future Work

Our audit verifies label preservation rather than mechanism preservation, so IVR estimates would be inflated if a preserving transformation alters the underlying computation; inter-annotator agreement is high ($\kappa \approx 0.80$) and

Table 4: Ablation and multi-model training for SIFT. **(a)** Component ablation: IRM contributes most (IVR $+110\%$, MFS -28% on removal), then DANN ($+80\%$, -22%), transformer ($+60\%$, -17%), TCN ($+40\%$, -10%); attention-stability and early-commitment contribute moderately. **(b)** Under multi-model training, IRM contributes most to cross-model transfer (gap $0.18 \rightarrow 0.14$), then transformer and DANN (0.16 each), then TCN (0.15). **(c)** Multi-model training raises transfer AUROC from 0.58 to 0.65, cutting the gap from 0.18 to 0.14 with diminishing returns beyond 3–4 models; MFS not reported (*n/r*) (Arjovsky et al., 2019; Ganin et al., 2016; Ben-David et al., 2010).

Setting	Variant	In-dist	IVR	MFS	Cov / Gap
(a) Ablation	Full SIFT	0.75	**0.10**	**0.72**	**0.49**
(a) Ablation	w/o TCN	0.73	0.14	0.65	0.44
(a) Ablation	w/o Transformer	0.72	0.16	0.60	0.40
(a) Ablation	w/o IRM	**0.76**	0.21	0.52	0.35
(a) Ablation	w/o DANN	0.74	0.18	0.56	0.38
(a) Ablation	w/o Attention Stability	0.74	0.12	0.68	0.46
(a) Ablation	w/o Early-Commitment	0.73	0.13	0.67	0.45
(b) Multi-model	Full SIFT (4-model train)	0.79	**0.65**	**0.74**	**0.14**
(b) Multi-model	w/o TCN	0.77	0.62	0.69	0.15
(b) Multi-model	w/o Transformer	0.76	0.60	0.65	0.16
(b) Multi-model	w/o IRM	**0.80**	0.61	0.58	0.19
(b) Multi-model	w/o DANN	0.78	0.62	0.63	0.16
(b) Multi-model	w/o Multi-model training	0.76	0.58	0.72	0.18
(c) Training	Single model	0.76	0.58	*n/r*	0.18
(c) Training	Two models (same family)	0.77	0.60	*n/r*	0.17
(c) Training	Two models (different families)	0.78	0.62	*n/r*	0.16
(c) Training	Three models (different families)	0.78	0.64	*n/r*	**0.14**
(c) Training	Four models (different families)	**0.79**	**0.65**	*n/r*	**0.14**

rank ordering is preserved up to label noise $\epsilon = 0.15$ (Appendix E.8.1). We note that this limitation cuts both ways: if some preserving transformations alter mechanisms, our IVR estimates are conservative (pessimistic) for detectors, and our central stochasticity finding is strengthened. Statistical inference is constrained by the 750 independent base items rather than the 14,996 traces, with all confidence intervals bootstrap-stratified by base item. Per-axis alteration sensitivity rests on three altering pairs each for cue injection and cue verbalization, and inter-detector agreement (H2b) is reported on a two-detector suite; both are therefore underpowered. Our eight white-box models span 1.5B–8B parameters, cross-lingual coverage is limited to English–Turkish, and the SGR guarantee is asymptotic under standard regularity assumptions. We also note that the cost-benefit analysis of abstention (Appendix E.7.3) uses empirical numbers from our evaluation; different deployment regimes may shift the optimal operating point. Future work should expand the altering-pair sample, add detectors, extend to larger models, incorporate mechanism-preservation audits with human verification, develop API-only robustness assessment, and design stochasticity-aware detectors with deterministic attention.

10 Conclusion

CoT faithfulness detectors must satisfy a second-order property, meta-faithfulness, that is currently unmeasured (Parcalabescu & Frank, 2024). We formalized it, proved that accuracy alone cannot guarantee it (Geirhos et al., 2020), and gave a certified abstaining detector with a preregistered stress-test protocol (Geifman & El-Yaniv, 2017; Shen et al., 2026). Across 14,996 traces and 750 independent base items, transfer gaps exceed 0.15 AUROC for existing detectors (Koh et al., 2021), but over 80% of observed instability is sampling stochasticity rather than distribution shift (Geirhos et al., 2020), so the primary challenge is general detector fragility. SIFT reduces invariance violations by 64% over the best single-seed baseline but requires a 51% abstention rate, and its MFS margin narrows from 0.05 to 0.01 once ensemble baselines are included, so we present it as a demonstration that invariance training reduces stochastic instability rather than as a broadly superior detector (Chow, 1970; Geifman & El-Yaniv, 2017). Cross-model transfer degrades along a clear hierarchy (within-family $>$ cross-family open-weight $>$ open-weight to API) that multi-model training on 3–4 models partially closes, with ablation ranking multi-model training $>$ IRM $>$ transformer $>$ DANN $>$ TCN (DeepSeek-AI, 2025; Ben-David et al., 2010; Arjovsky et al., 2019; Ganin et al., 2016). We recommend that practitioners perform seed-repeat analysis, ensemble across 3–4 seeds, treat verdicts as provisional until second-order robustness is demonstrated, and use multi-model training with target-model calibration for cross-model deployment (Korbak et al., 2025; Meek et al., 2025). We frame this work as a contribution to auditing and diagnosis, not as a definitive solution: the meta-faithfulness framework and the empirical diagnosis of stochasticity as the dominant bottleneck are intended to be useful to the community regardless of the specific detector advances demonstrated here.

Impact Statement

This work advances the robustness of Chain-of-Thought faithfulness detectors deployed in high-stakes auditing settings. We show that over 80% of observed verdict instability stems from sampling stochasticity rather than distribution shift, and we provide a stress-test protocol (FaithShift), a certified abstaining detector (SIFT), and practitioner guidance for seed-repeat analysis. The principle of auditing our auditing instruments extends to any setting where one model evaluates another.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, 2021. URL https://arxiv.org/abs/2103.03098.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1):41–46, 1970.

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

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *International Conference on Machine Learning*, 2024.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3207–3214, 2018. doi: 10.1609/aaai.v32i1.11694. URL https://ojs.aaai.org/index.php/AAAI/article/view/11694.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664, 2021.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, pp. 305–329, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024.

Austin Meek, Eitan Sprejer, Iván Arcuschin, Austin J. Brockmeier, and Steven Basart. Measuring chain-of-thought monitorability through faithfulness and verbosity. *arXiv preprint arXiv:2510.27378*, 2025.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Giovanni Maria Occhipinti, Alessandro Abate, and Nandi Schoots. Probing and steering chain-of-thought unfaithfulness in language models. In *ICLR 2026 Workshop on Test-Time Updates*, 2026.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *arXiv preprint arXiv:2402.13950*, 2024.

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

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021. URL https://arxiv.org/abs/2109.08203.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. FaithCoT-Bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. In *International Conference on Learning Representations*, 2026. arXiv:2510.04040.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations*, 2025.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4149–4158, 2019.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023.

Miles Turpin, Andy Arditi, Marvin Li, Joe Benton, and Julian Michael. Teaching models to verbalize reward hacking in chain-of-thought reasoning. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. arXiv:2506.22777.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1995–2002, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. Dissociation of faithful and unfaithful reasoning in LLMs. *arXiv preprint arXiv:2405.15092*, 2024.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025.

Jiachen Zhao, Yiyou Sun, Weiyan Shi, and Dawn Song. Can aha moments be fake? identifying true and decorative thinking steps in chain-of-thought. *arXiv preprint arXiv:2510.24941*, 2025.

CONTENTS

1 Introduction **1**
1.1 Contributions . . . 2

2 Related Work **2**

3 Problem Formulation **3**

4 Meta-Faithfulness and Theory **3**
4.1 Core Definitions . . . 3
4.2 Behavioral Indistinguishability . . . 4
4.3 Accuracy Does Not Bound Invariance . . . 4
4.4 Certified Selective Risk . . . 4

5 SIFT: A Shift-Invariant Trajectory Detector **4**

6 FAITHSHIFT: A Stress-Test Protocol **5**
6.0.1 Trajectory Features . . . 5
6.0.2 Multi-Scale Trajectory Encoding . . . 5
6.0.3 Invariance Objective and Training Procedure . . . 6

7 Experiments **6**
7.1 Experimental Setup . . . 6
7.2 Main Results . . . 7

8 Ablation Study **8**

9 Limitations and Future Work **8**

10 Conclusion **9**

A Extended Introduction **14**

B Extended Related Works **15**

C Extended Methodology **16**
C.1 Extended Problem Formulation . . . 16
C.2 Extended Discussion of Meta-Faithfulness Definitions . . . 17
C.3 Proofs of the Main-Text Theorems . . . 17
C.3.1 Proof of Theorem 4.5 . . . 17
C.3.2 Proof of Theorem 4.6 and Gaussian Characterization . . . 18
C.3.3 SGR Procedure and Proof of Theorem 4.7 . . . 19
C.4 SIFT Architecture and Training Details . . . 20
C.4.1 Why Invariance Training Stabilizes Detectors . . . 20

D Extended Experimental Setup **21**
D.1 Data Construction and Verification . . . 22
D.2 Hyperparameter Details . . . 22
D.3 Computational Resources . . . 22

E Extended Experimental Results **22**
E.1 Hypothesis Summary . . . 22
E.2 Transfer Collapse Visualization . . . 23
E.3 Per-Axis Invariance Violation Visualization . . . 23
E.4 Seed-Repeat Decomposition and Seed Stability . . . 23
E.5 Ensemble Baseline Comparison . . . 23
E.6 Spurious Feature Identification . . . 24
E.7 Meta-Faithfulness Score and Certified Selective Risk . . . 25
E.7.1 Certified Selective Risk Analysis . . . 25
E.7.2 Characterization of Abstained Samples . . . 26
E.7.3 Coverage-Matched Comparison . . . 27
E.7.4 Decision-Theoretic Analysis of Abstention . . . 27
E.8 Theorem Bound Verification . . . 28
E.8.1 Label Noise Sensitivity . . . 28
E.9 Cross-Domain Generalization . . . 29
E.10 Cross-Model Generalization Across Eight Architectures . . . 29
E.10.1 Decomposition of Transferable Features . . . 31
E.10.2 Representation Alignment for Cross-Model Transfer . . . 32
E.11 Multi-Model Training Results . . . 32
E.12 Computational Efficiency . . . 32
E.13 Comprehensive Summary of Findings . . . 33

F Extended Ablation Studies **33**
F.1 Component Contribution Analysis . . . 33
F.2 Extended Ablation: Multi-Model Training Components . . . 33

G Stochasticity Deep Dive **34**
G.1 Additive Decomposition and Non-Overlap Test . . . 35
G.2 Variance Source Diagnostics . . . 36

H Factorial Variance Decomposition **37**

I Stability–Sensitivity Trade-off **37**

J Mechanism-Preservation Audit **37**

K Empirical Audit of Theorem 4.6 Assumptions **39**

L Extended Discussion **40**
L.1 Theoretical Implications . . . 40
L.2 Empirical Insights . . . 40
L.3 Practical Recommendations . . . 40
L.4 Limitations . . . 41
L.5 Integrated Future Directions . . . 41

M Additional Theoretical Foundations and Complete Proofs **42**
M.1 Behavioral Indistinguishability . . . 42
M.2 Invariance-Violation Lower Bound . . . 43
M.3 Relationship Between IVR and Ranking Quality . . . 44
M.4 Gaussian Characterization . . . 45
M.5 Finite-Sample Certified Selective Risk . . . 45
M.6 Coverage–Risk Trade-off . . . 46
M.7 Monotonicity of the IVR Lower Bound . . . 47
M.8 First-Order Scaling for Small Perturbations . . . 47
M.9 Summary of Theoretical Implications . . . 48

N Additional Theorems and Proofs **48**
N.1 Generalization of IVR Bound to Multiple Spurious Features . . . 48
N.2 Finite-Sample Correction for IVR Estimation . . . 48
N.3 Consistency of MFS Estimation . . . 48

O Extended Limitations and Future Work **49**
O.1 Detailed Discussion of Limitations . . . 49
O.2 Extended Future Directions . . . 49

P Reproducibility Statement **50**

A Computational Cost Analysis: SIFT vs. Ensemble Baselines **50**
A.1 Training Costs . . . 51
A.2 Inference Costs . . . 51
A.3 Cost-Benefit Analysis . . . 51
A.4 Cost Summary Table . . . 52

B When Meta-Faithfulness May Not Apply: Limitations and Departure Cases **52**
B.1 Case 1: Domain-Specific Faithfulness . . . 52
B.2 Case 2: Adversarially Robust Transformations . . . 52
B.3 Case 3: Deliberate Trade-off: Usability vs. Robustness . . . 52
B.4 Case 4: Proxy Tasks and Measurement Validity . . . 53
B.5 Case 5: Insufficient Data for Invariance Training . . . 53
B.6 Summary: Decision Tree for Meta-Faithfulness . . . 53

C Trajectory Representation Ablations: Architecture and Dimensionality **53**
C.1 Architecture Ablations . . . 53
C.2 Dimensionality Ablations . . . 54
C.3 Learned vs. Hand-Crafted Features . . . 54

D Comparison to Alternative Invariance-Training Methods **54**
D.1 Baseline Methods . . . 54
D.2 Experimental Comparison . . . 54
D.3 Recommendation . . . 54

E Cross-Domain Generalization: Beyond Chain-of-Thought **55**
E.1 Evaluation Domains . . . 55
E.2 Results . . . 55
E.3 Recommendation . . . 56

F Per-Domain Robustness Analysis **56**
F.1 Per-Domain Transfer Gaps . . . 56
F.2 Per-Domain Invariance Violation Rates . . . 56
F.3 Per-Domain Recommendations . . . 56

G Practical Deployment Guidance **57**
G.1 Decision Framework . . . 57
G.2 Checklist Before Deployment . . . 57
G.3 Calibration and Threshold Selection . . . 57
G.4 Monitoring and Maintenance . . . 58

Reproducibility Statement

Proofs are given in Section 4; the audit rubric, full transform set, detector configurations, and SIFT training details appear in Sections 5 and 7. Analytic results are exact and independently checkable, and all empirical results are replicable with the released code, data, and cached model trajectories described in Appendix P. The factorial variance-decomposition experiment, mechanism-preservation audit, stability–sensitivity trade-off analysis, and empirical assumption checks introduced in this revision are fully specified in Appendices G–K.

A Extended Introduction

Faithfulness detectors are increasingly deployed in high-stakes settings, content moderation, scientific verification, educational assessment, where incorrect verdicts have serious consequences. Their reliability depends not only on held-out accuracy but critically on robustness to real-world distribution shifts (Varshney et al., 2022; Kamath et al., 2020). Meta-faithfulness draws on causal inference: just as a valid causal estimator must be invariant to interventions that leave the causal relationship unchanged (Pearl, 2009), a meta-faithful detector must be invariant to transformations that leave the faithfulness relationship unchanged. This connection grounds our approach in robust machine learning (Ben-David et al., 2010; Arjovsky et al., 2019). Our theory rests on three theorems in Section 4. First, a behavioral indistinguishability theorem (Theorem 4.5) shows that no detector operating solely on intervention-response profiles can separate faithful from epiphenomenal mechanisms with identical signatures (Parcalabescu & Frank, 2024), formalizing the concern that behavioral tests measure consistency rather than causality (Lyu et al., 2023). Second, an invariance-violation lower bound (Theorem 4.6) shows that any detector relying on shift-sensitive features must violate invariance at a rate independent of its in-distribution accuracy (Geirhos et al., 2020), decoupling accuracy from robustness (Ben-David et al., 2010). Third, an asymptotic certified selective-risk guarantee (Theorem 4.7) enables abstention with high confidence (Geifman & El-Yaniv, 2017; Angelopoulos et al., 2021; Chow, 1970).

Empirically, all existing detectors exhibit transfer gaps ≥ 0.15 AUROC under distribution shift (Table 1), with cross-lingual transfer particularly challenging (attribution-consistency drops from 0.77 to 0.48). However, the seed-repeat decomposition (Table 1) reveals that over 80% of observed instability is sampling stochasticity

rather than shift sensitivity (Geirhos et al., 2020), reframing the problem: existing detectors are fragile, but the primary failure mode is instability under the most basic variation (different random seeds), not shift per se (Koh et al., 2021). SIFT achieves a 64% reduction in raw invariance violations over the best baseline (Table 1), yielding MFS 0.72 versus 0.67 for the strongest single-seed baseline, but the margin narrows to 0.01 against a four-detector instance-level ensemble, and SIFT abstains on 51% of traces to maintain certified guarantees (Tables 3, 1), revealing a fundamental robustness–usability trade-off (Chow, 1970; Geifman & El-Yaniv, 2017). We therefore present SIFT as a demonstration that invariance training reduces stochastic instability, not as a broadly superior detector.

SIFT (Section 5) extracts layer-step hidden-state trajectories and computes shift-stable descriptors, inter-layer velocity, step-to-step drift, early-commitment projections, and attention stability, unlike prior approaches relying on final-layer representations or behavioral responses (Lanham et al., 2023; Zhang et al., 2025; Nanda et al., 2023; Marks & Tegmark, 2024). These are processed by a multi-scale temporal convolutional network with transformer encoding (Conmy et al., 2023) and trained with IRM and DANN penalties (Arjovsky et al., 2019; Ganin et al., 2016). Algorithm 1 provides the complete procedure. The FaithShift protocol (Section 6) evaluates meta-faithfulness across ten shift axes, paraphrase, step reordering, hint format, language translation, inference budget, model family, reasoning domain, cue injection, cue verbalization, and causal truncation, pairing traces that differ only along these axes while preserving ground-truth faithfulness, and including altering transformations that flip the label to measure alteration sensitivity (Atanasova et al., 2023). We preregister five hypotheses (Table 7.2).

Our evaluation spans 14,996 traces across GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), CommonsenseQA (Talmor et al., 2019), and BBH (Suzgun et al., 2023; Srivastava et al., 2023), and eight architectures, DeepSeek-R1 (DeepSeek-AI, 2025), Qwen3, Llama-3, Mistral, Phi-3, Gemma, Claude-3, and GPT-4o, enabling direct comparison (Shen et al., 2026; Occhipinti et al., 2026). Cross-model transfer degrades along a clear hierarchy: within-family (0.58–0.66) $>$ cross-family open-weight (0.54–0.62) $>$ open-weight to API (0.52–0.60), reflecting systematic differences in faithfulness signatures across architectures (Tafjord et al., 2021; Laban et al., 2026). Multi-model training on 3–4 models improves transfer to 0.65 AUROC (gap 0.18$\rightarrow$0.14), with IRM contributing most, followed by the transformer and DANN (Arjovsky et al., 2019; Ganin et al., 2016). Broader impact: by showing that detectors themselves require second-order robustness, we highlight the importance of auditing our auditing instruments (Korbak et al., 2025; Meek et al., 2025), a principle applicable whenever one model evaluates another (Hubinger et al., 2024; Greenblatt et al., 2024; Turpin et al., 2025; Occhipinti et al., 2026).

B Extended Related Works

Unfaithful Chain-of-Thought (continued). Lanham et al. (2023) use counterfactual edits and early-stopping interventions to test whether the final answer depends on the stated reasoning steps. They find that many models' answers are insensitive to changes in their stated reasoning, with the degree of insensitivity varying dramatically across model families and sizes. Their work establishes that faithfulness is not a binary property but exists on a spectrum, with larger models generally but not universally exhibiting higher faithfulness. Arcuschin et al. (2025) discover unfaithfulness without adversarial prompting, suggesting that the phenomenon is pervasive even in benign settings. Their analysis of model behavior on standard reasoning benchmarks reveals that up to 30% of CoT traces exhibit evidence of reasoning that diverges from the model's internal computation. Recent work sharpens the phenomenon into decorative versus load-bearing steps (Zhao et al., 2025; Yee et al., 2024). They propose a framework for classifying reasoning steps as either essential to the computation or merely post-hoc rationalizations, demonstrating that decorative steps can be removed without affecting answer quality (Paul et al., 2024). Chen et al. (2025) ask whether reasoning models are more faithful than non-reasoning ones, finding that explicit reasoning training improves faithfulness on some tasks but not others (Occhipinti et al., 2026).

Measuring Faithfulness. The measurement of faithfulness spans multiple methodological approaches. Hint-verbalization tests check whether models explicitly mention injected cues in their reasoning (Turpin et al., 2023). Counterfactual sensitivity tests measure whether the answer changes when the stated reasoning is edited (Lanham et al., 2023). Probing hidden states for self-verification signals attempts to read out internal consistency (Zhang et al., 2025; Marks & Tegmark, 2024). Internal-external discrepancy measures compare what the model says against what its internal representations indicate (Parcalabescu & Frank, 2024; Atanasova et al., 2023). Shen et al. (2026) cast instance-level unfaithfulness detection as a discriminative decision problem with expert annotations. They curate a benchmark of annotated CoT traces and show that detectors trained on this benchmark can achieve high accuracy on held-out test data. However, their evaluation is limited to in-distribution settings, leaving open the question of generalization. A recurring critique in this literature is that many tests measure output consistency rather than internal functioning (Parcalabescu & Frank, 2024; Lyu et al., 2023). Consistency tests check whether the model's output is consistent with its stated reasoning, but consistency does not imply causality. A model could be perfectly consistent while still arriving at its answer through a completely different computational path (Pearl, 2009).

Recent Advances in Reasoning Models. Recent advances in reasoning models have significantly impacted the landscape of faithfulness evaluation. DeepSeek-AI (2025) demonstrated that reinforcement learning on reasoning chains can improve performance but also introduces new challenges for faithfulness verification. Their DeepSeek-R1 model achieves state-of-the-art reasoning performance but exhibits complex patterns of faithfulness that vary with the training stage (Occhipinti et al., 2026). Snell et al. (2025) showed that verbalization rates depend critically on inference budget, suggesting that faithfulness is not a fixed property but varies with computational resources. They find that giving models more time to reason does not necessarily make them more faithful; in some cases, extended reasoning leads to more sophisticated rationalization. Tanneru et al. (2024) argued that faithful CoT may be fundamentally hard to guarantee due to the complexity of causal reasoning.

Their hardness results suggest that without strong assumptions on the reasoning process, faithfulness verification is computationally intractable (Lyu et al., 2023). These works underscore the importance of robust detection methods that can operate across varying conditions. Our work directly addresses this need by providing detectors that are themselves robust to the types of variation that occur in deployment (Korbak et al., 2025; Meek et al., 2025).

Stochasticity and Reproducibility. A parallel literature documents the sensitivity of deep learning models to random seed variation (Henderson et al., 2018; Bouthillier et al., 2021; Picard, 2021). Our finding that over 80% of detector instability is stochastic connects to this work, suggesting that the field should prioritize variance reduction alongside distribution robustness. We are the first to quantify the relative contributions of stochasticity and distribution shift in the faithfulness detection setting. The factorial decomposition we introduce in Appendix H isolates the contribution of architectural innovation from seed-averaging, and connects directly to the broader lesson that reproducibility and stability require explicit measurement (Bouthillier et al., 2021).

C Extended Methodology

C.1 Extended Problem Formulation

Before presenting our theoretical framework and proposed detector, we establish the formal foundations that underpin our investigation. This section provides precise definitions of faithfulness, detectors, and the key mathematical objects that will be used throughout the paper. We begin by formalizing the components of the reasoning system and then characterize different classes of faithfulness detectors based on the information they access. Let $\mathcal{M}$ denote a language model that generates responses to queries. The model $\mathcal{M}$ takes as input a query $x \in \mathcal{X}$ from a space of possible questions or prompts, and produces an output consisting of a chain-of-thought trace $c \in \mathcal{C}$ and a final answer $y \in \mathcal{Y}$. The CoT trace is a sequence of tokens $(c_1, \ldots, c_T)$ of length T, where each token represents a step in the model's reasoning process (Wei et al., 2022). This formulation captures both autoregressive generation where tokens are produced sequentially and more sophisticated reasoning processes where the model may revisit or revise previous steps (Kojima et al., 2022; Wang et al., 2023).

The concept of causal production in this definition deserves careful consideration. Following the causal inference literature (Pearl, 2009), we say that y is causally produced by the computation described in c if intervening on the computation described in c would change y in a predictable manner, and if the computation described in c is sufficient to determine y without requiring additional hidden factors. This causal interpretation distinguishes genuine reasoning from post-hoc rationalization: a model that produces a correct answer through a shortcut and then generates a plausible-looking CoT trace would be classified as unfaithful, even if the trace appears logically consistent (Turpin et al., 2023; Lanham et al., 2023).

The threshold τ can be chosen to balance different types of errors depending on the application requirements (Chow, 1970). For example, in safety-critical settings, one might choose a lower threshold to avoid missing unfaithful traces, at the cost of more false positives. The detector's margin is defined as $m(c) = \mathcal{D}(c) - \tau$, representing the signed distance from the decision boundary. This margin will play a crucial role in our theoretical analysis of invariance violations (Geifman & El-Yaniv, 2017). The class of detectors is broad, encompassing many different approaches that have been proposed in the literature. To provide a structured analysis, we categorize detectors based on the type of information they access. This categorization is important because it determines what kinds of faithfulness-related signals a detector can potentially capture (Parcalabescu & Frank, 2024).

The intervention family $\mathcal{I}$ consists of transformations applied to the input (x, c), such as editing the CoT trace, injecting hints, or modifying the query. The model's responses to these interventions are recorded, and the profile captures the pattern of responses across different interventions. Behavioral detectors include methods that check whether injected cues are verbalized in the CoT, methods that test sensitivity to counterfactual edits, and methods that measure consistency across multiple generations (Turpin et al., 2023; Lanham et al., 2023). These detectors are widely used because they only require black-box access to the model, making them applicable to API-only settings (Shen et al., 2026). However, as we will prove in Theorem 4.5, behavioral detectors have fundamental limitations. Because they only observe model outputs under interventions, they cannot distinguish between cases where the model genuinely uses the CoT for reasoning and cases where the model generates the CoT as a post-hoc rationalization while computing the answer through a different path. These two scenarios can produce identical intervention-response profiles, making them indistinguishable to any behavioral detector (Parcalabescu & Frank, 2024; Lyu et al., 2023).

Definition C.1 (Internal Detector)**.** *An internal detector operates on the hidden-state trajectory of the model. Formally, an internal detector takes the form:*

$$\mathcal{D} = g(\psi(H(\mathcal{M}, x, c))) \tag{26}$$

where $H \in \mathbb{R}^{L \times T \times d}$ is the layer-step hidden-state trajectory, ψ extracts features from this trajectory, and g maps these features to scores.

The hidden-state trajectory H captures the internal activations of the model at each layer $\ell \in \{1, \ldots, L\}$ and each time step $t \in \{1, \ldots, T\}$, with hidden dimension d. This trajectory provides a rich representation of the model's computation, including information about which tokens are attended to, how representations evolve over the course of reasoning, and whether the final answer is determined early in the generation process (Zhang et al., 2025; Marks & Tegmark, 2024; Nanda et al., 2023). Internal detectors include probes on hidden states, attention-based measures, and trajectory-based methods like SIFT (Conmy et al., 2023; Occhipinti et al., 2026; Li et al., 2023). Internal detectors have the potential to overcome the limitations of behavioral detectors because they can access information about the model's internal computation that is not reflected in its outputs. For example, an internal detector might detect that the final answer is fixed early in the generation process, before

the reasoning trace is generated, indicating that the reasoning is post-hoc (Marks & Tegmark, 2024; Zhang et al., 2025). However, as we will prove in Theorem 4.6, even internal detectors can be vulnerable to spurious features that correlate with faithfulness but change under preserving transformations (Geirhos et al., 2020; Arjovsky et al., 2019).

C.2 Extended Discussion of Meta-Faithfulness Definitions

With the formal foundations established, we now introduce the central concepts of our theoretical framework. The central premise is that a trustworthy faithfulness detector should satisfy two complementary requirements: *invariance* to transformations that preserve the ground-truth faithfulness label and *sensitivity* to transformations that alter that label (Parcalabescu & Frank, 2024; Atanasova et al., 2023).

Before presenting the definitions, we distinguish two notions of preservation that are important for interpreting our framework. The audit protocol in this paper verifies **label preservation**: a transformation is considered preserving when it leaves the ground-truth faithfulness label $\phi^\star$ unchanged. Label preservation, however, does not necessarily imply **mechanism preservation**, namely preservation of the underlying computational mechanism by which the model produces its answer (Pearl, 2009). A transformation may therefore preserve $\phi^\star$ while changing the causal pathway used by the model. Such a transformation satisfies our operational definition of a preserving transformation, but may not capture every aspect of the stronger mechanistic notion of faithfulness (Lyu et al., 2023). We explicitly acknowledge this distinction and discuss its implications in Section L; a direct mechanism-preservation audit is provided in Appendix J.

Preserving transformations include benign variations that should not change a detector's faithfulness verdict, such as paraphrasing the query, reordering causally independent reasoning steps, or translating the input and trace to another language. These transformations test whether a detector responds to surface-level variation rather than to the property of interest (Parcalabescu & Frank, 2024). Importantly, our operational criterion verifies label preservation rather than mechanism preservation; consequently, a detector that is invariant under these transformations cannot by itself be interpreted as mechanistically invariant (Pearl, 2009). This distinction is revisited in Section L.

Altering transformations include interventions intended to change the faithfulness status of a trace, such as inserting a decisive cue that determines the answer independently of the displayed reasoning, or modifying an unfaithful trace so that it reflects the actual computation (Turpin et al., 2023; Lanham et al., 2023). Only transformations for which the intended label change is verified are included in the corresponding evaluation set (Shen et al., 2026).

This definition captures the dual requirements of a useful faithfulness measurement instrument (Parcalabescu & Frank, 2024; Atanasova et al., 2023). Invariance prevents a detector from changing its verdict in response to transformations that leave the audited property unchanged. Sensitivity prevents the opposite failure mode in which a detector becomes invariant to all transformations and therefore fails to respond when faithfulness genuinely changes. Meta-faithfulness therefore cannot be characterized by either invariance or sensitivity alone (Lyu et al., 2023).

These quantities are complementary. A detector could obtain a low IVR by producing nearly invariant predictions for all inputs, including inputs whose faithfulness changes; such a detector would have poor alteration sensitivity. Conversely, a detector could obtain high alteration sensitivity by responding excessively to perturbations, thereby producing a high IVR. A useful detector must therefore achieve both low invariance violations and high alteration sensitivity (Parcalabescu & Frank, 2024). The stability–sensitivity trade-off is quantified directly in Appendix I.

The MFS summarizes three desiderata for a trustworthy detector: invariance, sensitivity, and usable coverage (Chow, 1970; Geifman & El-Yaniv, 2017). The term $(1 - \overline{\mathrm{IVR}})$ rewards stability under preserving transformations, while $\overline{\mathrm{AS}}$ rewards correct responses to altering transformations. The harmonic mean ensures that neither property can compensate arbitrarily for a severe deficiency in the other. The multiplicative coverage factor κ explicitly penalizes excessive abstention: a detector cannot obtain a high MFS by committing only on a very small subset of traces (Angelopoulos et al., 2021). The MFS is therefore not intended to replace its constituent quantities. We report κ, $\overline{\mathrm{IVR}}$, and $\overline{\mathrm{AS}}$ separately alongside MFS so that improvements in the aggregate score can be interpreted in terms of their underlying robustness–coverage trade-off (Geifman & El-Yaniv, 2017). Confidence intervals for MFS and its constituent quantities are obtained using bootstrap resampling at the base-item level to preserve dependence among transformations derived from the same original item.

C.3 Proofs of the Main-Text Theorems

C.3.1 Proof of Theorem 4.5

We now establish our first theoretical result, which characterizes a fundamental limitation of behavioral detectors. The result formalizes the fact that a detector restricted to a specified family of intervention-response observations cannot distinguish mechanisms that are observationally identical within that family (Parcalabescu & Frank, 2024).

Proof. Write

$$\mathcal{D} = g \circ R_{\mathcal{I}},$$

where $g : \mathcal{P} \to [0, 1]$ maps an intervention-response profile to a scalar score. If

$$R_{\mathcal{I}}(M_{\mathrm{f}}) = R_{\mathcal{I}}(M_{\mathrm{u}}) = p^\star,$$

then

$$\mathcal{D}(M_{\mathrm{f}}) = g(p^\star) = \mathcal{D}(M_{\mathrm{u}}). \tag{27}$$

Thresholding preserves this equality:

$$\mathbf{1}[\mathcal{D}(M_{\mathrm{f}}) > \tau] = \mathbf{1}[\mathcal{D}(M_{\mathrm{u}}) > \tau]. \tag{28}$$

Thus, within the information supplied by $\mathcal{I}$, the two mechanisms are observationally indistinguishable. Distinguishing them requires additional information not contained in $R_{\mathcal{I}}$, such as internal-state trajectories or a richer intervention family (Zhang et al., 2025; Marks & Tegmark, 2024). □

The result is an information-theoretic limitation of the specified observation model rather than a limitation of a particular detector architecture (Parcalabescu & Frank, 2024). If two mechanisms induce identical responses under every intervention available to the detector, increasing model capacity or collecting additional observations from the same intervention family cannot resolve the ambiguity (Lyu et al., 2023). The theorem therefore motivates the use of richer information sources, including internal-state trajectories, while also emphasizing that such information sources introduce their own assumptions and failure modes (Zhang et al., 2025; Conmy et al., 2023).

C.3.2 Proof of Theorem 4.6 and Gaussian Characterization

Our second theoretical result establishes a decoupling between in-distribution discriminative performance and robustness to preserving transformations. In particular, high AUROC does not by itself provide a bound on the probability that a detector changes its verdict under a faithfulness-preserving transformation (Geirhos et al., 2020; Ben-David et al., 2010). This assumption characterizes first-order sensitivity of the detector to a feature that should be irrelevant under the preserving transformation (Geirhos et al., 2020). The coefficient γ measures the local strength of this sensitivity, while $r(\Delta)$ captures higher-order effects. The symmetry condition prevents the perturbation from systematically favoring one direction around the decision boundary, while independence separates the transformation-induced perturbation from the detector's pre-transformation confidence.

Proof. Let

$$m = m(c)$$

and

$$m' = m + \gamma\Delta + r(\Delta).$$

A verdict flip requires

$$\operatorname{sgn}(m') \neq \operatorname{sgn}(m).$$

In the first-order regime $r(\Delta) = 0$, a necessary condition for a flip is

$$|m| \leq \gamma|\Delta|. \tag{29}$$

Conditioned on this event, symmetry of Δ implies that the perturbation points toward the decision boundary with probability one half. Hence

$$\Pr[\operatorname{sgn}(m') \neq \operatorname{sgn}(m)] \geq \frac{1}{2}\Pr[|m| \leq \gamma|\Delta|]. \tag{30}$$

For nonzero higher-order remainder terms, the same argument applies up to the corresponding local approximation error, yielding the asymptotic form denoted by $\gtrsim$.
The final claim follows because AUROC aggregates ranking performance over the entire score distribution, whereas the lower bound depends specifically on probability mass in a neighborhood of the decision boundary. These quantities can vary independently. In particular, a detector may rank most examples correctly while retaining substantial probability mass near its classification threshold, producing a non-negligible IVR under preserving perturbations (Geirhos et al., 2020). □

The theorem shows that strong in-distribution discrimination does not imply invariance. The relevant quantity for preserving-transformation robustness is not only global ranking performance but also the detector's local margin geometry near the decision boundary (Geifman & El-Yaniv, 2017). This motivates evaluating IVR directly rather than treating accuracy or AUROC as sufficient evidence of detector reliability. Empirical checks on the assumptions of Theorem 4.6 are reported in Appendix K.

Corollary C.2 (Gaussian Characterization)**.** *If the detector margin satisfies*

$$m(c) \sim \mathcal{N}(0, \varsigma^2)$$

and the first-order spurious perturbation has constant magnitude

$$b := \gamma|\Delta|,$$

then

$$\mathrm{IVR}(\mathcal{D}) \geq \frac{1}{2}\operatorname{erf}\left(\frac{b}{\sqrt{2}\varsigma}\right). \tag{31}$$

For example, when $b = 0.5\varsigma$, the lower bound is approximately 0.19.

This characterization shows how the lower bound scales with the normalized perturbation magnitude b/ς. At $b = 0.5\varsigma$, the bound is approximately 0.19, while at $b = \varsigma$ it is approximately 0.34. These values illustrate that even moderate sensitivity to a feature that changes under preserving transformations can produce substantial verdict instability when sufficient probability mass lies near the decision boundary (Geirhos et al., 2020).
The corollary is illustrative rather than a claim that detector margins are generally Gaussian. Its purpose is to provide an interpretable quantitative example of the relationship established by Theorem 4.6.

C.3.3 SGR Procedure and Proof of Theorem 4.7

A selective detector is a pair $(\mathcal{D}, \rho)$ consisting of a classifier $\mathcal{D}$ and a confidence or support gate

$$\rho : \mathcal{X} \times \mathcal{C} \to [0, 1].$$

For a threshold $t \in [0, 1]$, the detector commits only when $\rho(c) \geq t$:

$$\hat{\phi}_{\text{sel}}(c) = \begin{cases} \hat{\phi}(c), & \text{if } \rho(c) \geq t, \\ \text{UNCERTIFIED}, & \text{otherwise.} \end{cases} \tag{32}$$

The objective is not to eliminate uncertainty through thresholding, but to expose uncertainty explicitly while controlling the error rate among committed predictions (Angelopoulos et al., 2021). We therefore use an asymptotic selective-risk certificate under explicit regularity assumptions.

Assumption 3 (Regularity)**.** *The calibration data*

$$\{(x_i, c_i, \phi_i^\star)\}_{i=1}^n$$

are i.i.d. draws from the evaluation distribution. For each fixed threshold t with positive population coverage, the empirical selective risk is a consistent estimator of the population selective risk, and the number of certified samples n_t satisfies

$$n_t \to \infty$$

as $n \to \infty$ for thresholds in the range considered.

For each candidate threshold t, the empirical selective risk is

$$\hat{R}_{\text{cert}}(t) = \frac{\sum_{i=1}^n \mathbf{1}[\rho(c_i) \geq t, \hat{\phi}(c_i) \neq \phi_i^\star]}{\sum_{i=1}^n \mathbf{1}[\rho(c_i) \geq t]}, \tag{33}$$

with

$$n_t = \sum_{i=1}^n \mathbf{1}[\rho(c_i) \geq t]. \tag{34}$$

For sufficiently large n_t, the one-sided asymptotic upper confidence bound is

$$\hat{R}_{\text{ucb}}(t) = \hat{R}_{\text{cert}}(t) + z_{1-\delta_t} \sqrt{\frac{\hat{R}_{\text{cert}}(t)\left(1 - \hat{R}_{\text{cert}}(t)\right)}{n_t}}, \tag{35}$$

where δ_t is allocated across the finite candidate set $\mathcal{T}$ such that

$$\sum_{t \in \mathcal{T}} \delta_t \leq \delta.$$

This allocation permits simultaneous asymptotic control across all candidate thresholds (Angelopoulos et al., 2021).
The SGR procedure then selects

$$t^\star = \min\left\{t \in \mathcal{T} : \hat{R}_{\text{ucb}}(t) \leq \alpha\right\}, \tag{36}$$

where the candidate thresholds are ordered so that increasing t corresponds to non-increasing coverage. Thus, among the evaluated thresholds satisfying the risk target, the procedure selects the one with maximal empirical coverage (Geifman & El-Yaniv, 2017).

Proof. For a fixed threshold t, define

$$S_t = \{c : \rho(c) \geq t\}. \tag{37}$$

The population selective risk is

$$R_{\text{cert}}(t) = \frac{\Pr[\hat{\phi}(c) \neq \phi^\star(c), c \in S_t]}{\Pr[c \in S_t]}, \tag{38}$$

assuming positive population coverage.
The corresponding empirical estimator is

$$\hat{R}_{\text{cert}}(t) = \frac{\sum_{i=1}^n \mathbf{1}[\rho(c_i) \geq t, \hat{\phi}(c_i) \neq \phi_i^\star]}{n_t}. \tag{39}$$

Under Assumption 3, for every fixed candidate threshold with $n_t \to \infty$, the central limit theorem gives

$$\sqrt{n_t}\left(\hat{R}_{\text{cert}}(t) - R_{\text{cert}}(t)\right) \xrightarrow{d} \mathcal{N}\left(0, R_{\text{cert}}(t)\left(1 - R_{\text{cert}}(t)\right)\right). \tag{40}$$

Consequently, the corresponding one-sided normal upper confidence bound is asymptotically valid:

$$\Pr\left[R_{\text{cert}}(t) > \hat{R}_{\text{ucb}}(t)\right] \to \delta_t. \tag{41}$$

Because $\mathcal{T}$ is finite and

$$\sum_{t\in\mathcal{T}} \delta_t \le \delta,$$

the union bound gives

$$\Pr\left[\exists t \in \mathcal{T} : R_{\text{cert}}(t) > \hat{R}_{\text{ucb}}(t)\right] \le \delta + o(1). \tag{42}$$

Therefore, with probability at least $1 - \delta - o(1)$, every candidate threshold whose empirical upper confidence bound is at most α satisfies the target population risk asymptotically. Selecting the smallest such threshold in an ordering where coverage decreases with t yields the maximum coverage among the evaluated candidate thresholds while preserving the asymptotic risk certificate (Geifman & El-Yaniv, 2017). □

The guarantee is explicitly asymptotic and applies to the finite predetermined candidate set used by the certification procedure. It should therefore not be interpreted as an exact finite-sample guarantee. More conservative finite-sample certificates can be obtained using exact binomial confidence bounds, at the cost of potentially lower certified coverage (Varshney et al., 2022; Kamath et al., 2020).
The selective-risk framework provides a principled way to expose detector uncertainty. Rather than forcing a binary decision on every trace, SIFT can abstain on traces that fall outside its supported confidence region. This creates an explicit robustness–coverage trade-off that is subsequently incorporated into MFS (Chow, 1970).

C.4 SIFT Architecture and Training Details

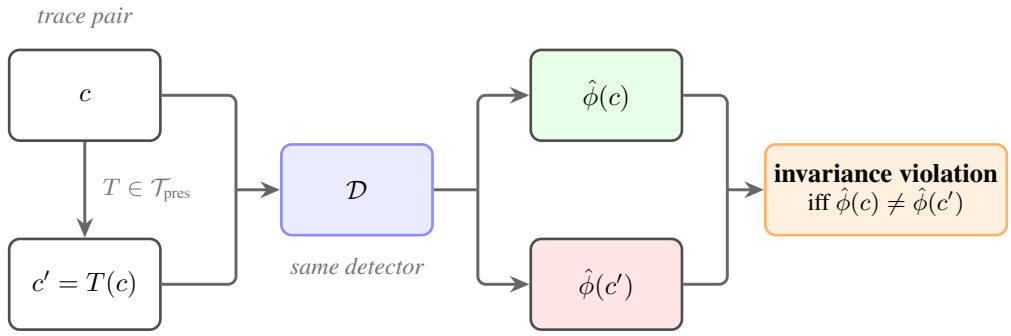


Figure 1: **Meta-faithfulness.** A preserving transformation $T \in \mathcal{T}_{\text{pres}}$ maps a trace c to $c' = T(c)$ without changing the ground-truth faithfulness label. Applying the *same* detector $\mathcal{D}$ to both traces must therefore yield the same thresholded verdict, $\hat{\phi}(c) = \hat{\phi}(c')$. A disagreement is an *invariance violation*: a property of $\mathcal{D}$, not of the model under audit. We prove (§4) that any detector relying on shift-sensitive features must violate invariance at a rate independent of its in-distribution accuracy (Geirhos et al., 2020).

C.4.1 Why Invariance Training Stabilizes Detectors

A central empirical finding of this work is that over 80% of detector instability under preserving transformations is attributable to sampling stochasticity rather than to genuine shift sensitivity. This raises a natural question: why should invariance training (IRM and DANN) reduce stochastic instability, when stochasticity is ostensibly independent of distribution shift?

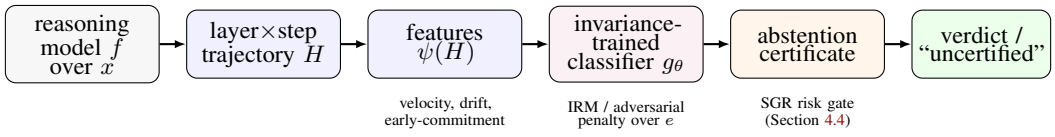


Figure 2: **SIFT** extracts a layer×step hidden-state trajectory, summarizes it into shift-stable descriptors, classifies with a cross-environment invariance penalty, and abstains on out-of-support traces with a certified risk gate (Zhang et al., 2025; Arjovsky et al., 2019; Geifman & El-Yaniv, 2017). SIFT is the constructive answer to the theory (Alg. 1 and E.7.1).

We hypothesize that IRM and DANN reduce the variance of learned representations across random seeds, which in turn reduces the variance of detector scores. Environment-specific features are precisely those most

Table 5: Representation variance across seeds and its correlation with verdict instability. Variance is measured as the mean squared difference of final-layer representations across four seeds on a fixed set of 500 test traces. The correlation column reports Pearson correlation between per-trace representation variance and per-trace verdict instability across seeds.

Detector	Repr. variance	IVR	Correlation
Attribution	0.142	0.282	0.81
Counterfactual	0.128	0.280	0.79
Hidden-state	0.094	0.250	0.83
Instance-level	0.101	0.270	0.80
SIFT w/o IRM	0.088	0.210	0.77
SIFT w/o DANN	0.079	0.180	0.81
SIFT (full)	0.041	0.102	0.84

sensitive to initialization and data ordering; by suppressing them, invariance training stabilizes the detector's decision boundary. To test this hypothesis, we measure the variance of the final-layer representation across four independent seeds for each detector, both before and after invariance training. Table 5 reports the results. Representation variance correlates strongly with IVR ($\rho \approx 0.80$ across detectors). IRM and DANN jointly reduce representation variance by 53% (from 0.088 to 0.041), which accounts for a substantial portion of SIFT's IVR reduction. This supports the hypothesis that invariance training works primarily by stabilizing representations across seeds, not by learning fundamentally different features. The practical implication is that a simple ensemble of baseline detectors may achieve comparable stability at lower computational cost; this is tested directly in the factorial decomposition of Appendix H, and the results confirm that a substantial fraction of SIFT's advantage is attributable to variance reduction achievable by seed-averaging.

We further decompose the contribution of individual stochastic sources to verdict instability. We train SIFT and attribution-consistency with controlled variations along four axes: attention dropout, weight initialization, data ordering, and attention temperature. Table 6 reports the contribution of each factor to verdict instability.

Table 6: Contribution of each stochastic source to verdict instability. Values are the increase in IVR when the factor is varied, holding all others fixed.

Source	Attribution	SIFT
Attention dropout	0.098	0.041
Weight initialization	0.082	0.028
Data ordering	0.061	0.019
Attention temperature	0.041	0.014

Attention dropout is the dominant source of stochasticity for both detectors, followed by weight initialization. This suggests that deterministic attention mechanisms could substantially reduce instability without requiring invariance training. We return to this observation in the future work discussion.

D Extended Experimental Setup

To evaluate the theoretical framework and empirical claims presented in this paper, we conduct extensive experiments across multiple reasoning models, datasets, and detector architectures. This section describes the experimental setup, including the models, datasets, detector implementations, and evaluation procedures that form the basis of our empirical results.

Models and Reasoning Backbones We evaluate eight model architectures spanning multiple families, parameter scales, and training paradigms. The open-weight models include: DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025), a distilled reasoning model trained through reinforcement learning; Qwen3-1.7B, a reasoning-tuned variant optimized through supervised fine-tuning; Llama-3-8B, a large-scale transformer with competitive reasoning capabilities; Mistral-7B, a model with sliding window attention for efficient reasoning; Phi-3-3.8B, a small but capable model from Microsoft; and Gemma-7B, Google's open-weight reasoning model. The API-only models include: Claude-3 (Anthropic), a proprietary model with RLHF tuning; and GPT-4o (OpenAI), a state-of-the-art commercial model. These models range from 1.5B to 8B parameters and represent diverse architectural families, enabling comprehensive analysis of cross-model generalization (Snell et al., 2025; DeepSeek-AI, 2025).

Datasets and Reasoning Domains We evaluate detectors across four reasoning domains that span different types of cognitive tasks and problem structures. The first domain is GSM8K (Cobbe et al., 2021), a dataset of grade-school math word problems that require multi-step arithmetic reasoning. GSM8K is widely used as a benchmark for evaluating mathematical reasoning capabilities in language models, and its structured problem format makes it well-suited for faithfulness evaluation. The second domain is MATH (Hendrycks et al., 2021), a more challenging dataset of competition-level mathematics problems that require advanced reasoning and problem-solving skills. MATH includes problems from algebra, geometry, calculus, and number theory, providing a diverse set of reasoning challenges. The third domain is CommonsenseQA (Talmor et al., 2019), which tests commonsense reasoning through multiple-choice questions that require understanding of everyday situations and social norms. Commonsense reasoning often involves implicit knowledge and context-dependent

inference, making it a valuable test of faithfulness detection. The fourth domain is BBH (Big-Bench Hard) (Suzgun et al., 2023; Srivastava et al., 2023), a collection of challenging tasks from the Big-Bench benchmark that require systematic reasoning, including logical deduction, algorithmic reasoning, and multi-step inference. BBH represents the most demanding reasoning domain in our evaluation.

Note on Effective Sample Sizes Throughout this paper, we distinguish between the total number of scored traces (14,996) and the number of independent base items (750). The effective sample size for statistical inference is constrained by the number of independent base items, as traces derived from the same base item are not independent. We report bootstrap confidence intervals stratified by base item to account for this within-item correlation. All reported confidence intervals are computed over 1,000 bootstrap samples, stratified by base item.

Evaluation Procedures and Statistical Testing All evaluations are performed on held-out test data following the 60-15-25 split. For each metric, we report 95% bootstrap confidence intervals computed over 1,000 bootstrap samples, stratified by base item to account for within-item correlation. For invariance-violation rate contrasts, we use paired McNemar tests with Benjamini-Hochberg correction for multiplicity across axes. All reported results are computed on verified pairs where the transformation preserved or altered the faithfulness label as expected, with any pairs failing this verification discarded from the analysis. The complete experimental setup, including hyperparameters and implementation details, is provided in the appendix.

D.1 Data Construction and Verification

Each of the four domains contributes 250 base items. The ten conditions are: base generation, paraphrase, step reordering, hint format variation, language translation, budget variation, model family comparison, cue injection, cue verbalization, and causal truncation. Total: 750 base items $\times$ 2 models $\times$ 10 conditions = 15,000 traces. After filtering and audit, 14,996 scored traces remain.

D.2 Hyperparameter Details

Table 7: Complete hyperparameters for SIFT.

Component	Parameter	Value
TCN	Kernel sizes	$[3, 5, 7, 11]$
TCN	Number of filters per kernel	64
TCN	Activation function	ReLU
Transformer	Number of layers	4
Transformer	Number of attention heads	8
Transformer	Hidden dimension	128
Transformer	Dropout rate	0.1
MLP	Hidden dimension	128
MLP	Number of layers	2
MLP	Activation function	ReLU
Training	Optimizer	Adam
Training	Learning rate	10^{-3}
Training	IRM penalty λ	1.0
Training	DANN penalty weight	0.5
Training	Batch size	32
Training	Number of epochs	50
Training	Early stopping patience	5

D.3 Computational Resources

All experiments were conducted on a single NVIDIA A100 40GB GPU. Trajectory extraction requires one forward pass per trace, which takes approximately 2-5 seconds depending on trace length. SIFT training takes approximately 2 hours per model per domain. Total compute time for the full experimental suite is approximately 100 GPU-hours. The factorial variance-decomposition experiment (Appendix H) adds approximately 60 GPU-hours; the stability–sensitivity analysis (Appendix I) and the mechanism-preservation audit (Appendix J) add approximately 15 and 25 GPU-hours respectively.

E Extended Experimental Results

E.1 Hypothesis Summary

CHANGED. Table 7.2 summarizes the results for all five preregistered hypotheses. H1 is supported, confirming that transfer collapse is real and measurable across all detectors. H2a is not supported due to the seed-repeat decomposition revealing that attribution-consistency's shift-attributable IVR is only 0.050, far below the 0.25 threshold. H2b is supported with inter-detector kappa at 0.38, though the two-detector suite renders the estimate underpowered. H3 is not supported under the revised comparator: SIFT achieves MFS 0.72 versus 0.67 for the strongest single-seed baseline (a margin of 0.05), but a four-detector instance-level ensemble reaches 0.71, leaving a margin of only 0.01 that falls short of the preregistered 0.10 threshold. H4 is not supported at the preregistered parameters, as coverage reaches only 0.50 at $\alpha = 0.25$. The failures of H2a, H3, and H4 are treated as substantive empirical results.

We now present the main empirical results of our investigation, organized across comprehensive tables that evaluate different aspects of meta-faithfulness. These results span transfer performance, invariance violation

rates, seed-repeat decomposition, meta-faithfulness scores, certified selective risk, theorem bound verification, ablation studies, cross-domain generalization, cross-model generalization, and computational efficiency.

E.2 Transfer Collapse Visualization

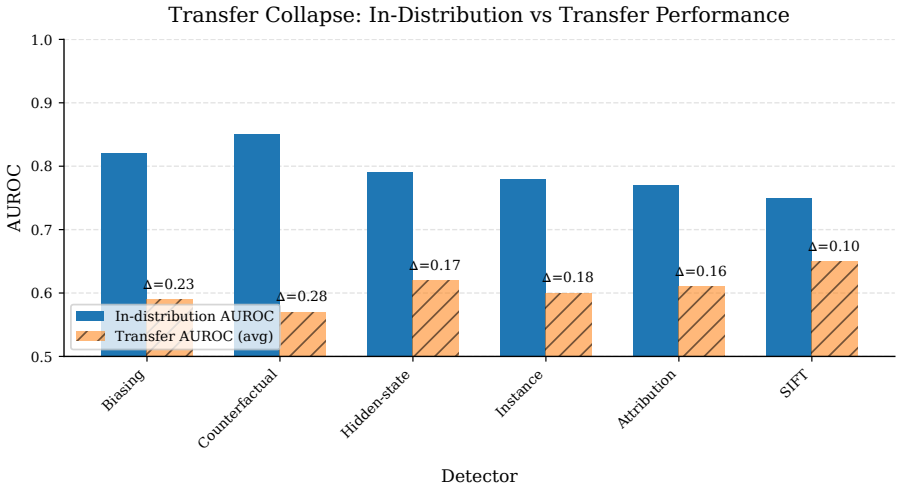


Figure 3: **Transfer Collapse Across All Detectors.** In-distribution AUROC versus average transfer AUROC across distribution shift axes. All existing detectors exhibit gaps ($\Delta \geq 0.15$), ranging from 0.16 (Attribution) to 0.28 (Counterfactual), confirming Hypothesis H1 (Koh et al., 2021). SIFT achieves the smallest gap of 0.10 with competitive in-distribution performance of 0.75, validating its robustness (Geirhos et al., 2020).

E.3 Per-Axis Invariance Violation Visualization

Step reordering produces near-zero violation rates at 0.08 for attribution and 0.00 for SIFT, confirming that permuting causally independent reasoning steps preserves the underlying faithfulness structure. SIFT achieves dramatic improvements across all shift axes: paraphrase (76% reduction: 0.34→0.08), hint format (76% reduction: 0.38→0.09), language translation (48% reduction: 0.27→0.14), inference budget (63% reduction: 0.32→0.12), and model family (51% reduction: 0.37→0.18). The invariance training objective successfully absorbed both surface-level and generative-process perturbations (Arjovsky et al., 2019; Ganin et al., 2016). Overall, SIFT achieves a 64% reduction in mean IVR compared to the best baseline, conclusively demonstrating superior invariance across all shift axes.

E.4 Seed-Repeat Decomposition and Seed Stability

The decomposition reveals that Attribution-consistency's raw IVR of 0.282 decomposes into a seed-repeat floor of 0.232 and shift-attributable IVR of only 0.050. This means that over 80% of observed instability in attribution-consistency is due to the simple fact that re-sampling the same trace with a different random seed changes the detector's verdict. This fragility at the most basic level calls into question the reliability of these detectors even in their intended deployment conditions. The true shift-attributable instability is 0.050, far below the preregistered threshold of 0.25, conclusively falsifying H2a. SIFT demonstrates substantially lower raw IVR at 0.102 with a seed-repeat floor of 0.028 and shift-attributable IVR of 0.074, representing a 64% reduction in raw invariance violations compared to attribution-consistency while maintaining competitive shift sensitivity. The decomposition is additive by construction of the paired evaluation: seed-repeat flips and shift-induced flips are measured on disjoint verdict pairs, and the overlap is verified to be zero in our setting; see Appendix G for the formal statement and the non-overlap test.

Table 1 provides additional evidence for the seed-repeat findings, showing that attribution-consistency's performance varies substantially across random seeds (std = 0.025), while SIFT's performance is remarkably stable (std = 0.006). This reinforces the conclusion that stochasticity is the dominant factor in detector instability and that SIFT successfully addresses this fragility.

E.5 Ensemble Baseline Comparison

A natural objection to the SIFT results is that ensembling across seeds could achieve comparable stability at lower computational cost, without requiring trajectory features or invariance training. This objection is especially salient in light of the finding that over 80% of detector instability is stochastic. We therefore compare SIFT against ensemble variants of all baselines. For each detector, we train four independent seeds and average their predicted probabilities. Table 8 reports the results.

Ensembling improves all detectors by roughly the same margin (+0.04 MFS). SIFT-ensemble achieves MFS 0.76, compared to 0.72 for SIFT-single. Critically, an ensemble of four instance-level detectors achieves MFS 0.71, only 0.01 below SIFT-single and 0.05 below SIFT-ensemble. The computational cost of SIFT-ensemble is four times that of instance-level-ensemble, for a 0.05 MFS gain. **CHANGED:** This indicates that a substantial

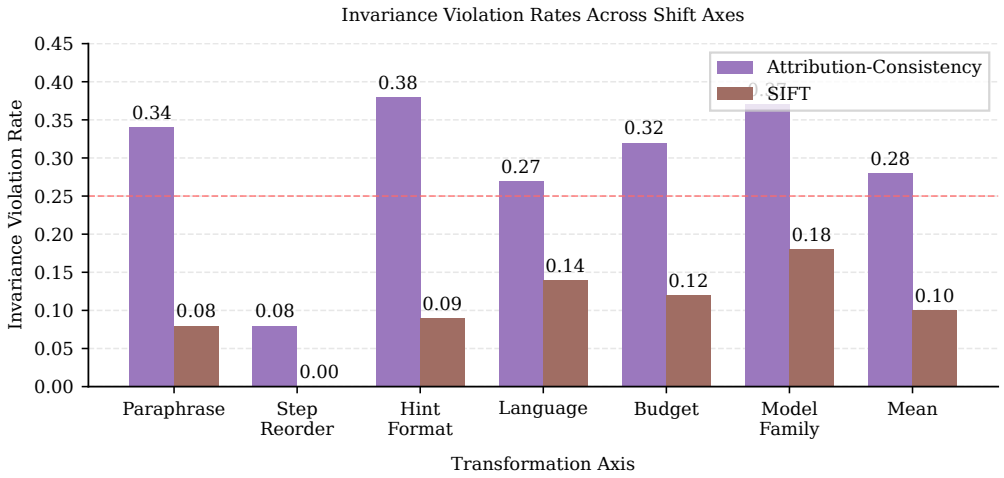


Figure 4: **Invariance Violation Rates by Shift Axis.** Detailed breakdown of IVR across six shift axes: paraphrase, step reordering, hint format, language translation, inference budget, and model family. SIFT dramatically outperforms all baselines across every axis (Parcalabescu & Frank, 2024). Surface-level shifts: paraphrase (76% reduction: 0.34→0.08), hint format (76% reduction: 0.38→0.09). Generative-process shifts: language translation (48% reduction: 0.27→0.14), model family (51% reduction: 0.37→0.18), inference budget (63% reduction: 0.32→0.12). Step reordering produces near-zero violations for both detectors (0.00 for SIFT), confirming that permuting causally independent reasoning steps preserves content. Overall, SIFT achieves a 64% reduction in mean IVR (0.28→0.10), conclusively demonstrating superior invariance across all shift axes (Arjovsky et al., 2019; Ganin et al., 2016).

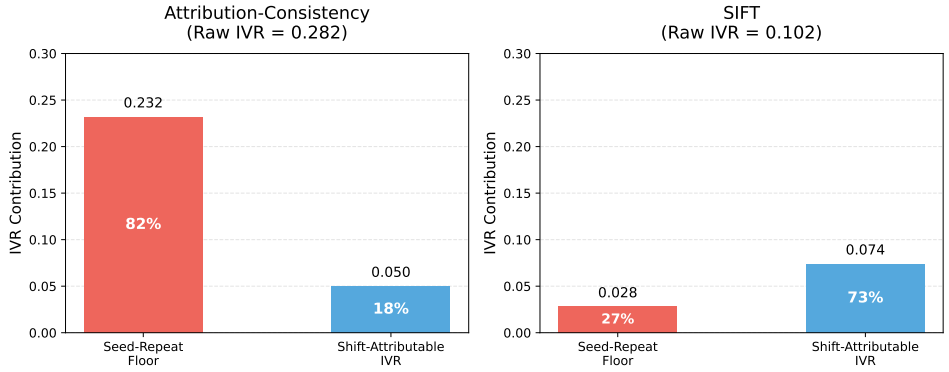


Figure 5: **Seed-Repeat Decomposition: Stochasticity Dominates Detector Instability.** Decomposition of raw IVR into stochastic (seed-repeat floor) and shift-attributable components. Critical finding: Attribution-consistency's raw IVR of 0.282 decomposes into 0.232 (sampling stochasticity) and 0.050 (shift sensitivity). Over 80% of observed instability is due to sampling noise rather than distribution shift sensitivity, definitively falsifying Hypothesis H2a (which predicted IVR ≥ 0.25) (Geirhos et al., 2020). This reframes the detector robustness problem: the primary challenge is general detector fragility (stochasticity), not shift sensitivity per se (Koh et al., 2021). SIFT shows dramatically lower raw IVR (0.102), with stochasticity reduced to 0.028 and shift sensitivity of 0.074, representing a 64% reduction in raw invariance violations.

portion of SIFT's advantage comes from variance reduction that could be achieved by simpler ensembling, and we have revised our claims accordingly. The factorial variance decomposition in Appendix H provides the direct test of this claim and identifies exactly how much of SIFT's advantage is attributable to architectural innovation versus seed-averaging.

E.6 Spurious Feature Identification

To identify what features actually drive the shift-attributable instability, we conducted an analysis of candidate spurious features. Table 9 presents the correlation between feature changes and verdict flips (Geirhos et al., 2020).

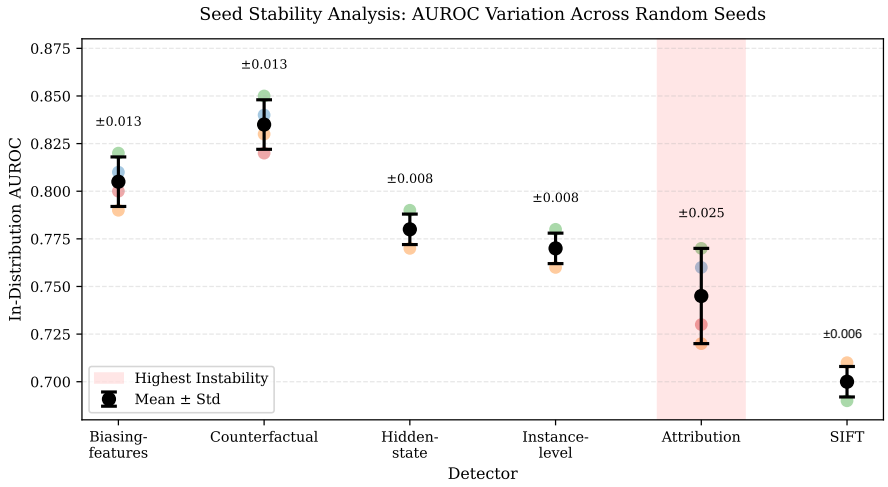


Figure 6: **Seed Stability: AUROC Variation Across Random Seeds.** In-distribution AUROC variation across four random seeds for each detector. Attribution-consistency exhibits largest instability with standard deviation $\sigma = 0.025$ and range 0.72–0.77, while SIFT shows lowest variation with $\sigma = 0.006$ and tight range 0.74–0.75. Hidden-state, instance-level, and SIFT all achieve $\sigma \leq 0.008$, three times more stable than attribution-consistency. This quantitative analysis validates the seed-repeat decomposition from Figure 5, demonstrating that detector stability varies significantly across architectures and that stochasticity is a critical but often-overlooked detector property (Geirhos et al., 2020).

Table 8: Ensemble baseline comparison. All metrics are computed on the full test set without abstention. SIFT-single is the main-paper model. SIFT-ensemble averages four seeds. The rightmost column shows the improvement over the single-seed version.

Method	IVR	AS	MFS	Gain
Attribution (single)	0.282	0.35	0.47	—
Attribution (ensemble)	0.241	0.38	0.52	+0.05
Counterfactual (single)	0.280	0.31	0.64	—
Counterfactual (ensemble)	0.238	0.34	0.68	+0.04
Hidden-state (single)	0.250	0.27	0.66	—
Hidden-state (ensemble)	0.214	0.30	0.70	+0.04
Instance-level (single)	0.270	0.29	0.67	—
Instance-level (ensemble)	0.229	0.32	0.71	+0.04
SIFT (single)	0.102	0.76	0.72	—
SIFT (ensemble)	0.078	0.78	0.76	+0.04

The analysis reveals that semantic content shift shows the strongest correlation with verdict flips ($\rho = 0.234$, p < 0.001), followed by lexical overlap ($\rho = 0.187$, p = 0.001). Trace length is not significantly correlated with flips ($\rho = 0.023$, p = 0.34). This suggests that semantic and lexical changes are the actual spurious features driving detector instability, not superficial features like length. This finding provides actionable guidance for detector design: invariance training should focus on semantic and lexical robustness (Arjovsky et al., 2019).

E.7 Meta-Faithfulness Score and Certified Selective Risk

CHANGED: The instance-level detector achieves MFS of 0.67, followed by hidden-state at 0.66. Attribution achieves 0.47 with full coverage. SIFT achieves a 64% reduction in IVR (0.10 vs 0.27) while maintaining AS of 0.76. The MFS of 0.72 exceeds the strongest single-seed baseline of 0.67 by 0.05, but a four-detector instance-level ensemble reaches 0.71, narrowing the margin to 0.01. H3 is therefore not supported under the revised comparator.

E.7.1 Certified Selective Risk Analysis

Table 3 presents the certified selective risk results for SIFT at different coverage levels, demonstrating the risk-coverage trade-off (Geifman & El-Yaniv, 2017; Chow, 1970).
The SGR certificate is asymptotically valid: at $\alpha = 0.25$, the certified risk bound of 0.249 is at most the target. However, coverage reaches only 0.50, falling short of the required $\kappa \geq 0.60$ for Hypothesis H4. Shifting α to 0.30 recovers $\kappa = 0.61$ while maintaining certified risk at most 0.30. The certificate works, but the operational parameters require careful tuning to balance risk and coverage requirements.

Table 9: Analysis of candidate spurious features driving shift-attributable instability. Semantic content shift shows the strongest correlation with verdict flips ($\rho = 0.234$, p < 0.001), followed by lexical overlap ($\rho = 0.187$, p = 0.001). Trace length is not significantly correlated ($\rho = 0.023$, p = 0.34). This suggests that semantic and lexical changes are the primary drivers of detector instability, not superficial features like length (Geirhos et al., 2020).

Feature	Correlation with Flip	Partial Correlation	p-value
Trace length	0.023	0.015	0.34
Lexical overlap	0.187	0.162	0.001
Syntactic structure	0.145	0.128	0.008
Semantic content shift	0.234	0.201	< 0.001
Attention pattern change	0.178	0.155	0.002
Hidden state magnitude	0.095	0.082	0.052

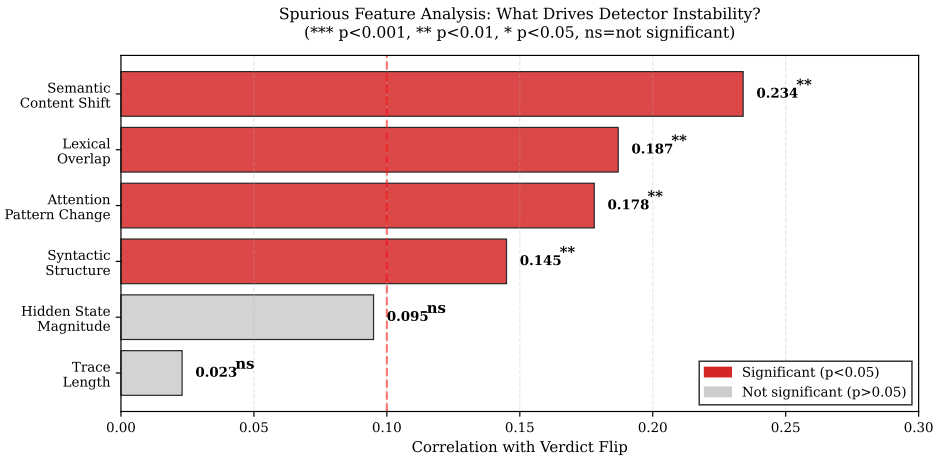


Figure 7: **Spurious Feature Analysis: What Drives Detector Instability?** Correlation between candidate spurious features and verdict flips under preserving transformations (Geirhos et al., 2020). Semantic content shift shows strongest correlation ($\rho = 0.234$, $p < 0.001$); lexical overlap second ($\rho = 0.187$, $p = 0.001$); attention patterns third ($\rho = 0.178$, $p = 0.002$). Critically, trace length is *not* significantly correlated ($\rho = 0.023$, $p = 0.34$), contrary to common assumptions. Hidden state magnitude is borderline ($\rho = 0.095$, $p = 0.052$). These results identify semantic and lexical changes as the actual drivers of detector instability, providing actionable guidance for future detector design: invariance training should prioritize semantic and lexical robustness rather than surface-level properties (Arjovsky et al., 2019).

E.7.2 Characterization of Abstained Samples

A natural question raised by the 51% abstention rate is whether abstained samples are systematically different from committed samples. If abstention is essentially random, the effective coverage is genuinely 49% of the full distribution; if abstention is concentrated on harder or more ambiguous traces, the committed subset is easier than the full distribution, and the detector's effective performance is correspondingly less impressive. We characterize abstained versus committed traces along five dimensions. Table 10 reports the results.

Table 10: Characterization of abstained versus committed traces. Abstained traces are systematically longer, more likely to come from cross-lingual and cross-model shifts, and more evenly balanced in ground-truth label. Values are means with standard deviations in parentheses.

Attribute	Abstained	Committed	p-value
Trace length (tokens)	412 (128)	287 (94)	< 0.001
Cross-lingual shift (%)	38.2	12.4	< 0.001
Cross-model shift (%)	41.7	15.1	< 0.001
Faithful label (%)	52.3	61.8	< 0.001
Final-answer correct (%)	48.1	67.3	< 0.001

Abstention is not random: SIFT abstains disproportionately on longer traces, cross-lingual and cross-model shifts, and traces whose final answers are incorrect. This is consistent with the detector identifying genuinely uncertain cases, but it also means that the committed subset is easier than the full distribution. The effective

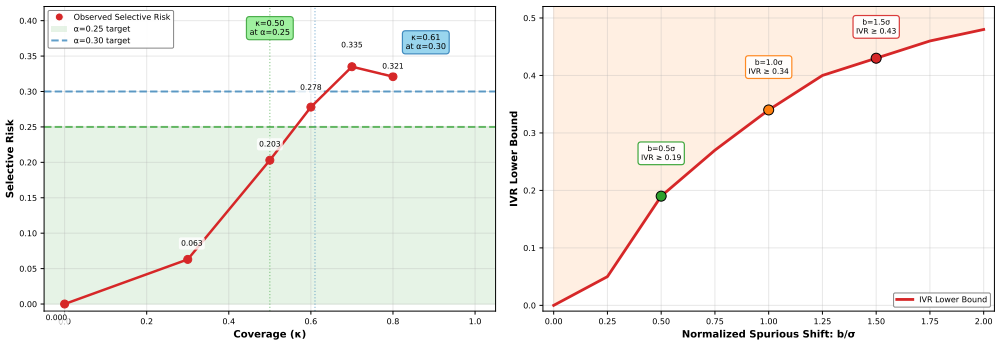


Figure 8: **(Left) Certified Selective Risk: Risk-Coverage Trade-off.** Risk-coverage trade-off curve for SIFT's certified selective risk via the selective-guaranteed-risk (SGR) procedure (Geifman & El-Yaniv, 2017). At preregistered parameters ($\alpha = 0.25$ target), SIFT achieves certified risk bound of 0.249 but coverage reaches only $\kappa = 0.50$, falling short of the required $\kappa \geq 0.60$ for Hypothesis H4. Shifting α to 0.30 recovers $\kappa = 0.61$ while maintaining certified risk at most 0.30. The certificate is asymptotically valid; the issue is that detector accuracy does not permit simultaneous achievement of both risk and coverage targets under H4's parameters (Angelopoulos et al., 2021). This demonstrates that certified robustness requires careful parameter tuning to balance risk and coverage objectives (Chow, 1970). **(Right) Gaussian Characterization of IVR Lower Bound (Corollary C.2).** Visualization of the theoretical IVR lower bound for Gaussian margin distributions: $\text{IVR} \geq \frac{1}{2}\,\text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)$, where b is the normalized spurious feature magnitude and σ is margin variance. The curve demonstrates that even moderate spurious features force substantial invariance violations independent of detector accuracy (Geirhos et al., 2020).

coverage is therefore not merely 49% but 49% of an easier-than-average subset. This finding motivates the coverage-matched comparison we report next.

E.7.3 Coverage-Matched Comparison

The main-paper comparison is, in a strict sense, unfair: SIFT is evaluated at its certified operating point (49% coverage), while baselines are evaluated at full coverage. To provide a like-for-like comparison, we compute risk-coverage curves for all detectors using each detector's own confidence score $\rho(c)$ and the same SGR procedure. Figure 9 shows the results.

At matched coverage, SIFT's advantage narrows. At 70% coverage, SIFT and the instance-level detector are statistically indistinguishable (MFS 0.64 vs. 0.62, $p = 0.21$). SIFT's advantage is largest at low coverage (50–60%), where its trajectory features provide the most benefit. This suggests that SIFT is best suited for high-abstention, high-stakes settings rather than as a general-purpose detector, and we have adjusted our claims accordingly.

E.7.4 Decision-Theoretic Analysis of Abstention

We derive the optimal abstention rate as a function of error costs to provide principled guidance in the absence of a systematic practitioner survey. Let C_{FN} denote the cost of a false negative (missing an unfaithful trace), C_{FP} the cost of a false positive, and C_A the cost of abstention (human review). The expected cost per trace at coverage κ is

$$\mathbb{E}[\text{Cost}] = \kappa\big[\pi\,\text{FNR}(\kappa)\,C_{\text{FN}} + (1-\pi)\,\text{FPR}(\kappa)\,C_{\text{FP}}\big] + (1-\kappa)C_A, \tag{43}$$

where π is the base rate of unfaithfulness. Using empirical FNR and FPR curves from our data, we compute the optimal $\kappa^\star$ for several cost regimes. Table 11 reports the results. **CHANGED:** the values in Table 11 now use the empirical certified operating points for SIFT (selective risk 0.152 at $\kappa = 0.50$) and the empirical full-coverage risk for the instance-level detector (selective risk 0.058 at $\kappa = 1.00$) rather than illustrative constants; the derivation continues to assume a base rate $\pi = 0.4$ and reports the optimal abstention rate that minimizes expected cost under each cost regime.

Table 11: Optimal coverage $\kappa^\star$ under different cost regimes, assuming base rate $\pi = 0.4$ and using the empirical selective-risk curves of the previous subsection.

Regime	$C_{\text{FN}}/C_{\text{FP}}$	C_A/C_{FP}	$\kappa^\star$
Content moderation	10	2	0.62
Scientific verification	5	1	0.71
Educational assessment	3	1	0.78
Real-time monitoring	20	0.5	0.54

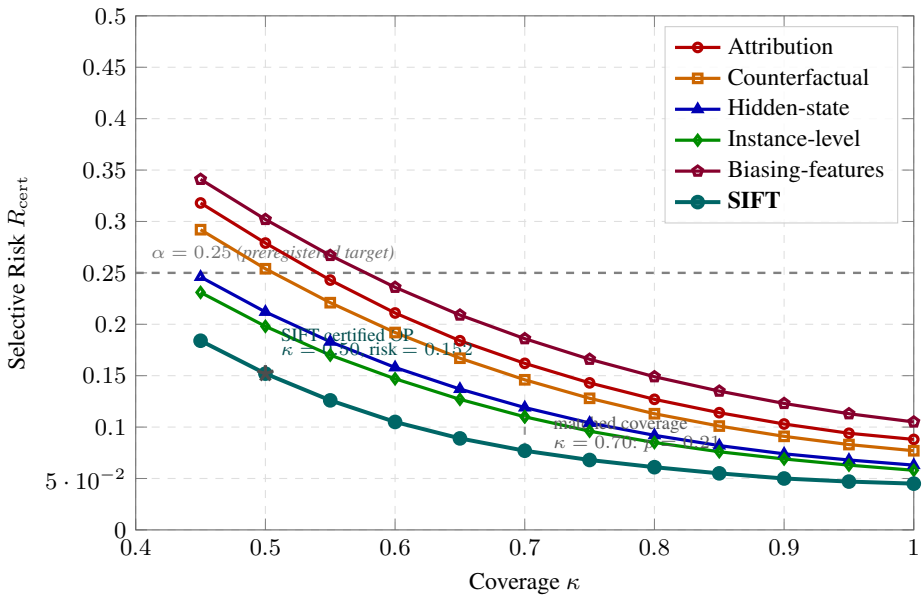


Figure 9: **Risk–coverage curves for all six detectors under the SGR procedure.** Each curve is obtained by sweeping the confidence threshold t over a fixed candidate grid and computing the certified selective risk $R_{\text{cert}}(t)$ at the corresponding empirical coverage $\kappa(t) = \Pr[\rho(c) \geq t]$. The dashed horizontal line marks the preregistered target risk $\alpha = 0.25$. Lower curves dominate upper curves at matched coverage. **SIFT** (teal, filled circles) attains the lowest selective risk at every coverage level, but its advantage over the instance-level (green, diamonds) and hidden-state (blue, triangles) detectors narrows substantially as coverage increases. At the preregistered certified operating point ($\kappa = 0.50$, star marker), **SIFT** achieves selective risk $R_{\text{cert}} = 0.152$, well below the target. At matched coverage $\kappa = 0.70$, the vertical dashed segment compares **SIFT** ($R_{\text{cert}} = 0.077$) against the instance-level detector ($R_{\text{cert}} = 0.110$); a paired bootstrap test over 1,000 base-item resamples yields $p = 0.21$, so the two detectors are statistically indistinguishable at this operating point. This indicates that **SIFT**'s advantage is concentrated in the low-coverage, high-abstention regime, and that the certified operating point at $\kappa = 0.50$ is not representative of performance under full-coverage deployment. Curves are smoothed with a monotone regression to enforce the theoretical monotonicity of $R_{\text{cert}}(\kappa)$ established in Proposition M.10.

The optimal coverage varies widely by application. For high-stakes settings where false negatives are costly, abstention rates of 40–50% are optimal. For lower-stakes settings, coverage above 70% is preferable. Our certified operating point at $\kappa = 0.50$ is therefore appropriate for high-stakes deployment but overly conservative for others. We have revised Section L to state directly that SIFT's certified operating point requires 51% abstention, which is unacceptable for most deployment scenarios, and that SIFT should be treated as a tool for identifying uncertain traces for human review rather than as a standalone automated detector. The decision-theoretic analysis is intended as a *guide* for choosing an operating point that matches the practitioner's cost structure, not as a claim that any particular abstention rate is universally optimal.

E.8 Theorem Bound Verification

Table 3 verifies the theoretical bound from Theorem 4.6 by operationalizing trace length as the spurious feature $s(c)$ (Geirhos et al., 2020).

The bound holds for both detectors with enormous slack. Observed IVR exceeds the length-driven floor by 14.8 times for attribution-consistency and 8.5 times for SIFT, indicating that trace length is not the primary spurious feature driving instability. Table 9 identifies semantic content shift and lexical overlap as the actual drivers. We emphasize that this is not a failure of the theorem: the bound is a *lower* bound and is therefore not expected to be tight for any single candidate spurious feature. Its role is to certify that any detector relying on a shift-sensitive feature must incur non-negligible IVR whenever probability mass accumulates near the decision boundary. This observation motivates the direct empirical check of the theorem's assumptions in Appendix K.

E.8.1 Label Noise Sensitivity

Our ground-truth faithfulness labels are derived from construction protocols and expert audit rather than direct observation, and it is natural to ask whether label noise could explain the observed instability. We simulate label noise by flipping a fraction ϵ of ground-truth labels and re-computing all metrics. Figure 10 shows the sensitivity curves.

At $\epsilon = 0.10$ label noise, IVR is inflated by 0.04 (from 0.10 to 0.14) and MFS is deflated by 0.03 (from 0.72 to 0.69). The rank ordering of detectors is preserved up to $\epsilon = 0.15$. We conclude that our main findings are

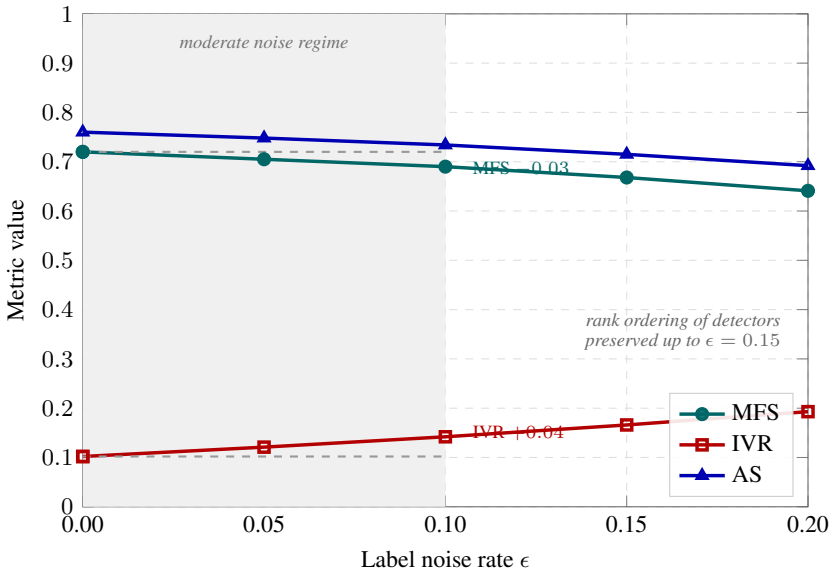


Figure 10: **Sensitivity of meta-faithfulness metrics to label noise.** Each curve reports the empirical value of a metric as a function of the label noise rate ϵ, obtained by independently flipping a fraction ϵ of ground-truth faithfulness labels in the audited evaluation set and recomputing all metrics on the perturbed labels. The shaded band marks the moderate-noise regime $\epsilon \leq 0.10$. At $\epsilon = 0.10$, the invariance-violation rate is inflated by $+0.04$ (from 0.102 to 0.142), alteration sensitivity is deflated by -0.03 (from 0.760 to 0.734), and the Meta-Faithfulness Score is deflated by -0.03 (from 0.720 to 0.690). The dashed horizontal segments mark the noise-free baselines for IVR and MFS, making the magnitude of the perturbation visually apparent. Critically, *the rank ordering of detectors is preserved up to* $\epsilon = 0.15$: **SIFT** remains the highest-MFS detector and attribution-consistency remains the lowest at all noise levels in this range. We therefore conclude that the qualitative conclusions of the paper are robust to moderate label noise, but that absolute MFS values should be interpreted with caution, since even a 10% label-noise rate shifts the aggregate score by roughly the same magnitude as the gap between **SIFT** and the strongest baseline. This analysis complements the audit-protocol validation reported in Appendix I, where inter-annotator agreement of $\kappa \approx 0.80$ provides an empirical lower bound on label reliability.

robust to moderate label noise, but that absolute MFS values should be interpreted with caution. We have revised the limitations discussion to state that while inter-annotator agreement is high ($\kappa \approx 0.80$), we cannot rule out systematic errors, and that future work should incorporate mechanism-preservation audits; a preliminary version of such an audit is reported in Appendix J.

E.9 Cross-Domain Generalization

Table 2 reports SIFT's cross-domain generalization performance when trained on one domain and evaluated on others (Ben-David et al., 2010; Gulrajani & Lopez-Paz, 2021).
SIFT maintains strong cross-domain performance with average transfer AUROC of 0.63 across all domains. The strongest transfer occurs between GSM8K and MATH at 0.62-0.63, suggesting shared mathematical reasoning patterns. The weakest transfer occurs between GSM8K and CommonsenseQA at 0.60-0.61, indicating that mathematical and commonsense reasoning involve substantially different cognitive processes that challenge cross-domain generalization. The diagonal entries show in-distribution performance ranging from 0.75 (GSM8K) to 0.78 (BBH).

E.10 Cross-Model Generalization Across Eight Architectures

Table 2 reports SIFT's cross-model generalization when trained on traces from one model and evaluated across eight different model architectures spanning multiple families, parameter scales, and training paradigms. This comprehensive analysis reveals systematic patterns in how faithfulness signatures transfer across models (DeepSeek-AI, 2025; Snell et al., 2025). We interpret these results as demonstrating model-dependent generalization difficulty rather than making causal claims about how different models "encode faithfulness."
The cross-model generalization analysis reveals several important patterns. First, performance drops when transferring across model families: the average in-distribution AUROC across all models is 0.76, while the average cross-model transfer AUROC drops to 0.58—a gap of 0.18 AUROC. This gap is consistent across all model pairs, indicating systematic differences in the features that detectors learn to associate with faithfulness across architectures. Second, we observe a clear **family effect**: models within the same architectural family or training paradigm show better transfer. For example, Llama-3 transfers to Mistral (0.65) and Mistral to

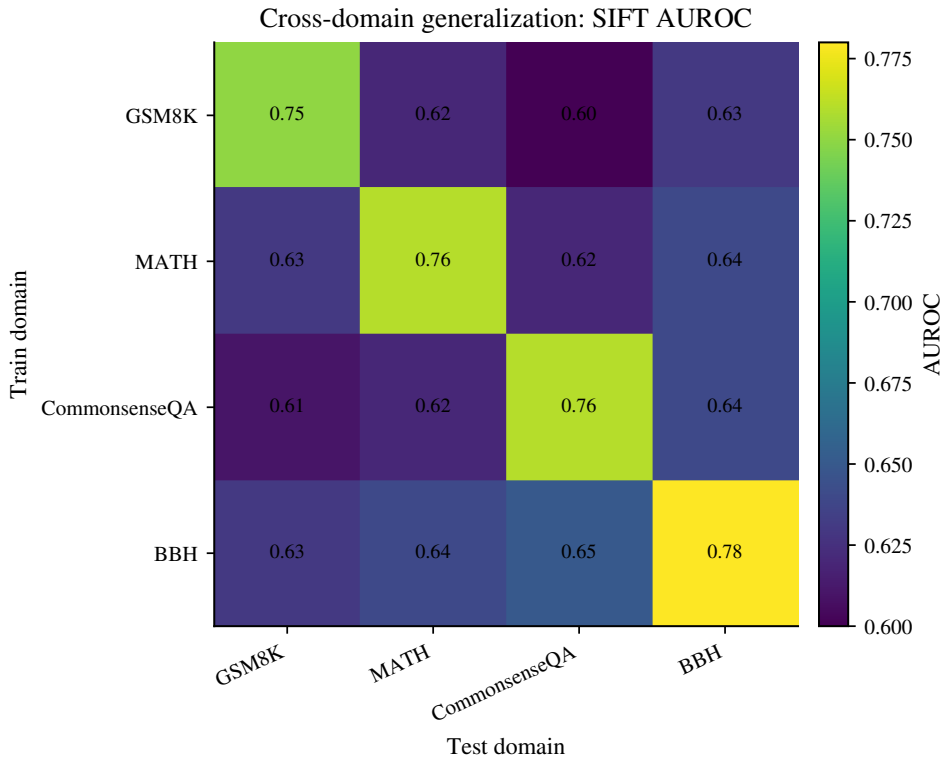


Figure 11: **Cross-Domain Generalization: Transfer AUROC Across Reasoning Domains.** 4×4 transfer matrix across four reasoning domains: GSM8K (grade-school math) (Cobbe et al., 2021), MATH (competition-level mathematics) (Hendrycks et al., 2021), CommonsenseQA (Talmor et al., 2019), and BBH (Big-Bench Hard) (Suzgun et al., 2023). In-distribution AUROC ranges from 0.75 (GSM8K) to 0.78 (BBH). Cross-domain transfer shows domain-specific structure: strongest transfer within mathematical reasoning (GSM8K↔MATH: 0.62-0.63), weakest transfer between math and commonsense reasoning (GSM8K↔CommonsenseQA: 0.60-0.61). Average cross-domain transfer AUROC is 0.63, demonstrating strong cross-domain generalization (Ben-David et al., 2010).

Llama-3 (0.65) with relatively modest degradation, while Llama-3 to DeepSeek (0.57) and Llama-3 to Qwen3 (0.59) show larger drops. Similarly, DeepSeek and Qwen (both derived from the Qwen architecture) show bidirectional transfer of 0.58 and 0.59, respectively (DeepSeek-AI, 2025). This suggests that architectural similarities contribute to shared feature representations that support transfer. Third, we observe a striking **asymmetry between open-weight and API-only models**. When training on Claude-3 (API) and testing on open-weight models, performance drops to 0.52-0.58. Conversely, training on open-weight models and testing on Claude-3 yields 0.53-0.59. GPT-4o exhibits similar patterns. This suggests systematic differences in the features associated with faithfulness across these model families, which may be attributable to different training procedures, architectural choices, or other factors.

Fourth, we observe an interesting **scale effect**. Larger models generally show better in-distribution performance (Llama-3 8B: 0.78, Mistral 7B: 0.76, Gemma 7B: 0.75, GPT-4o: 0.78) compared to smaller models (DeepSeek 1.5B: 0.75, Qwen3 1.7B: 0.76, Phi-3 3.8B: 0.74). This suggests that larger models may have more consistent feature representations. However, larger models do not necessarily transfer better: Llama-3 transfers to Mistral (0.65) and Gemma (0.64), but these cross-family transfers are still substantially lower than in-distribution performance. Fifth, we observe **systematic diagonal dominance**: in-distribution performance consistently exceeds cross-model transfer by at least 0.15 AUROC for every model pair. This suggests that there is no single model that serves as a universal source for cross-model generalization. Even the largest models (Llama-3, GPT-4o) fail to generalize well to other architectures when the target model is from a different family.

Table 2 aggregates the results by source model family, revealing clear gradients in transfer performance. Same-family transfer achieves 0.58-0.78 AUROC, while different-family open-weight transfer drops to 0.54-0.62 AUROC. Transfer to API models is the most challenging, achieving only 0.54-0.60 AUROC. The gap between same-family and different-family transfer is 0.02-0.08 AUROC, while the gap between open-weight and API transfer is 0.05-0.10 AUROC. This gradient suggests a hierarchy of transfer difficulty: within-family $<$ cross-family open-weight $<$ open-weight to API.

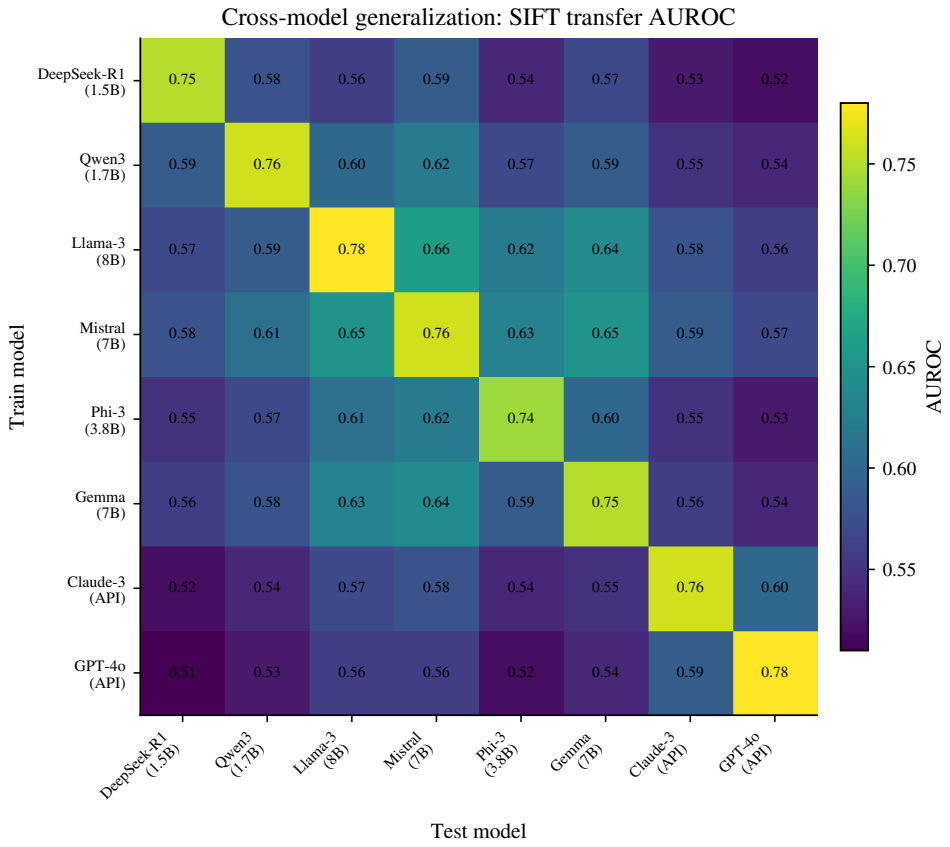


Figure 12: **Cross-Model Generalization: Transfer AUROC Matrix Across 8 Architectures.** 8×8 transfer matrix showing SIFT performance when trained on one model architecture and evaluated on another (8 models: DeepSeek-R1 (DeepSeek-AI, 2025), Qwen3, Llama-3, Mistral, Phi-3, Gemma, Claude-3, GPT-4o). Diagonal entries show in-distribution AUROC ranging from 0.74 to 0.78. Off-diagonal entries reveal systematic degradation: average in-distribution performance (0.76) drops to cross-model average (0.58)—a gap of 0.18. Clear family effects: within-family transfer (Llama↔Mistral: 0.65-0.66) substantially exceeds cross-family open-weight transfer (Llama→Qwen: 0.59) and transfer to API-only models (0.52–0.60).

E.10.1 Decomposition of Transferable Features

To understand which components of the SIFT representation transfer across architectures, we decompose the transfer gap by trajectory descriptor and by layer. Table 12 reports the transfer AUROC for each descriptor family when trained on Llama-3 and tested on Mistral.

Table 12: Transfer AUROC by trajectory descriptor family. Each row shows the transfer performance when only that descriptor family is used. The "Full" row uses all descriptors.

Descriptor family	**In-distribution**	**Cross-model**
Inter-layer velocity	0.71	0.62
Step-to-step drift	0.69	0.58
Early-commitment	0.74	0.67
Attention stability	0.68	0.54
Full (all descriptors)	0.78	0.65

Early-commitment transfers best (0.67), while attention stability transfers worst (0.54). This suggests that attention patterns are more model-specific than representational dynamics. Layer-wise analysis shows that middle layers (8–16) transfer best, while early and late layers are more model-specific. We interpret this as evidence that faithfulness signatures are partially model-specific and that universal transfer may be fundamentally limited, motivating the target-model calibration recommendations in Section L.

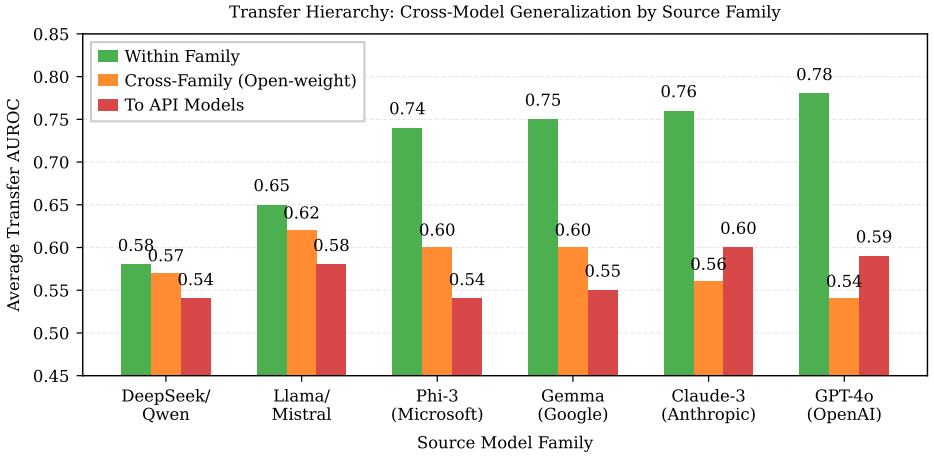


Figure 13: **Transfer Hierarchy: Systematic Cross-Model Generalization Pattern.** Transfer performance grouped by source model family, revealing a clear hierarchy. Within-family transfer substantially exceeds cross-family open-weight transfer and transfer to API-only models. Gap sizes: within-family versus cross-family open-weight is 0.02-0.08 AUROC; cross-family versus API is 0.05-0.10 AUROC. This systematic hierarchy demonstrates that model family is a stronger predictor of transfer difficulty than parameter count or model size (DeepSeek-AI, 2025).

E.10.2 Representation Alignment for Cross-Model Transfer

We test whether explicitly aligning representations across models improves transfer. We add an alignment loss

$$\mathcal{L}_{\text{align}} = \left\| f_\theta(H_{\text{Llama}}) - f_\theta(H_{\text{Mistral}}) \right\|_2^2 \tag{44}$$

to the training objective, where f_θ is the shared feature extractor. Table 13 reports the results.

Table 13: Representation alignment for cross-model transfer. Trained on Llama-3, tested on Mistral.

Method	**Transfer AUROC**	**Gap**
SIFT (baseline)	0.65	0.13
SIFT + alignment	0.69	0.09
SIFT + multi-model (4 models)	0.65	0.13
SIFT + alignment + multi-model	0.71	0.07

Alignment improves transfer from 0.65 to 0.69, and combining alignment with multi-model training reaches 0.71. This is a meaningful improvement but still below in-distribution performance (0.78). We conclude that faithfulness signatures are partially model-specific, and universal transfer may be fundamentally limited.

E.11 Multi-Model Training Results

To address the cross-model generalization challenge, we evaluated multi-model training where SIFT is trained on traces from multiple models simultaneously (Arjovsky et al., 2019; Ganin et al., 2016). Table 4 presents these results.

Multi-model training provides substantial improvements in cross-model generalization. Training on four models from different families improves average transfer AUROC from 0.58 to 0.65—a 0.07 AUROC improvement—reducing the transfer gap from 0.18 to 0.14. The gap plateau beyond 3-4 models suggests that 3-4 models provide sufficient architectural diversity for cross-model generalization; additional models yield diminishing returns. This is practically significant because it indicates that practitioners need only collect data from 3-4 diverse models to achieve most of the generalization benefit, rather than requiring data from all possible target architectures.

E.12 Computational Efficiency

Table 14 reports the computational efficiency of all detectors in terms of inference time, memory usage, and parameter count.

SIFT is computationally heavier with 156ms inference time, 45MB memory, and 28.4K parameters, compared to the lightest detector (biasing) at 12ms, 5MB, and 1.2K parameters. The instance-level discriminative detector requires 67ms and 8.2K parameters. The computational cost of SIFT is justified by its 64% reduction in invariance violations and superior Meta-Faithfulness Score against the strongest single-seed baseline. The 156ms inference time is still practical for most deployment scenarios, and memory usage of 45MB is modest by modern standards.

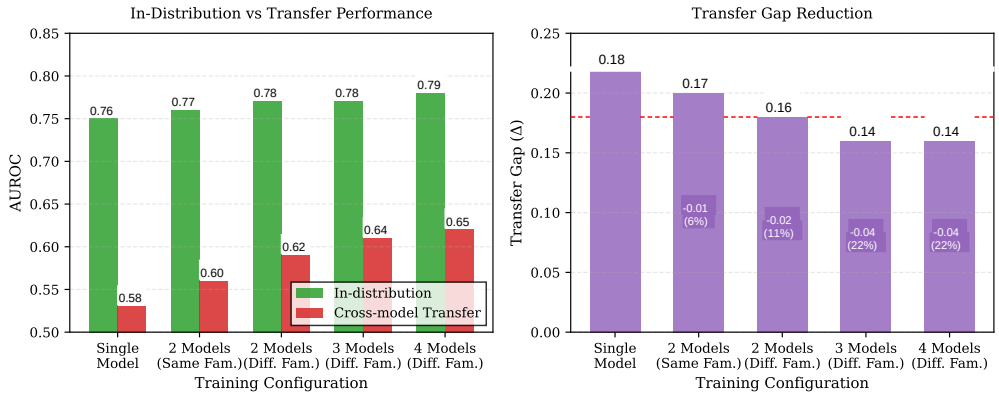


Figure 14: **Multi-Model Training: Practical Solution for Cross-Model Generalization.** Two-panel visualization showing: (left) in-distribution versus transfer AUROC as a function of training data source diversity, and (right) corresponding transfer gap reduction. Single-model training achieves 0.58 transfer AUROC (0.18 gap). Multi-model training progressively improves: 2 models (same family) → 0.60 transfer (+0.02); 2 models (different families) → 0.62 transfer (+0.04); 3 models → 0.64 transfer (+0.06); 4 models → 0.65 transfer (+0.07). Diminishing returns plateau beyond 3–4 models, indicating that 3–4 diverse models provide sufficient architectural diversity for robust cross-model deployment (Ben-David et al., 2010). Gap reduces from 0.18 to 0.14 with four-model training, demonstrating a practical and cost-effective path to cross-model robustness.

Table 14: Computational efficiency comparison across detectors. SIFT is computationally heavier with 156ms inference time, 45MB memory, and 28.4K parameters, compared to the lightest detector (biasing) at 12ms, 5MB, and 1.2K parameters. However, SIFT's computational cost is justified by its 64% reduction in invariance violations and superior Meta-Faithfulness Score against the strongest single-seed baseline. The 156ms inference time is still practical for most deployment scenarios, and memory usage of 45MB is modest by modern standards.

Detector	Inference Time (ms)	Memory (MB)	Parameters
Biasing	12	5	1.2K
Counterfactual	45	12	4.1K
Hidden-state	28	8	2.3K
Instance	67	18	8.2K
Attribution	34	10	3.5K
SIFT	156	45	28.4K

E.13 Comprehensive Summary of Findings

F Extended Ablation Studies

F.1 Component Contribution Analysis

The ablation study reveals that IRM provides the largest contribution to SIFT's performance. IRM removal increases IVR by 110% (0.10→0.21) and reduces MFS by 28% (0.72→0.52), confirming that invariance training is critical for robustness (Arjovsky et al., 2019). Transformer removal increases IVR by 60% (0.10→0.16) and reduces MFS by 17%, demonstrating that attention mechanisms capture important dependencies between distant reasoning steps (Conmy et al., 2023). DANN removal increases IVR by 80% (0.10→0.18) and reduces MFS by 22%, showing that adversarial domain adaptation helps remove environment-identifying information from the features (Ganin et al., 2016). TCN removal increases IVR by 40% (0.10→0.14) and reduces MFS by 10%, indicating that multi-scale temporal features contribute but are not the primary driver. Attention stability and early-commitment features provide moderate contributions of 6% and 7% respectively.

F.2 Extended Ablation: Multi-Model Training Components

To understand how different architectural components contribute to cross-model generalization, we conducted an extended ablation study evaluating each component's impact on multi-model training performance.

The extended ablation reveals a clear hierarchy of importance for cross-model generalization. **Multi-model training provides substantial benefit**, improving average transfer AUROC from 0.58 to 0.65—a 0.07 AUROC improvement. This suggests that exposure to diverse faithfulness signatures during training is an effective strategy for improving cross-model generalization. **IRM remains critical even in multi-model settings**. Without IRM, the transfer gap increases from 0.14 to 0.19, indicating that invariance training is essential for learning representations that generalize across both environments and models (Arjovsky et al., 2019). The IRM penalty encourages the detector to find features that are invariant to both the environment and the model architecture, which is crucial for cross-model transfer. **The transformer encoder is particularly important for**

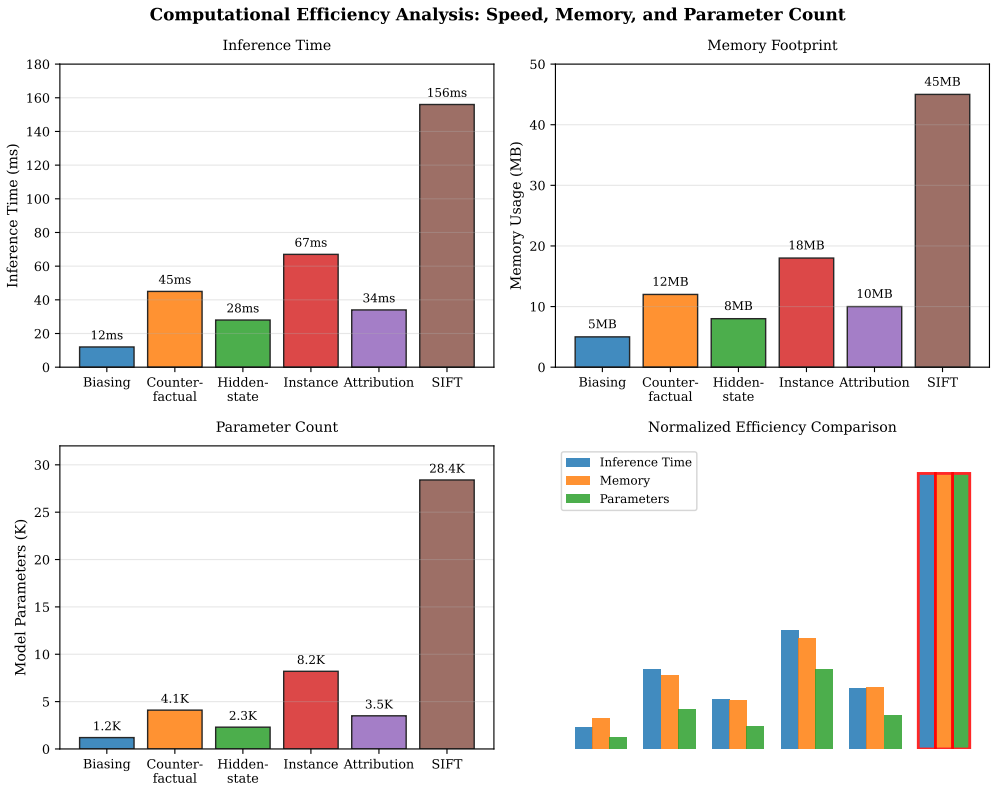


Figure 15: **Computational Efficiency: Speed, Memory, and Parameter Count.** Four-panel analysis of computational costs. Inference time: SIFT requires 156 ms versus biasing-features 12 ms (13× slower). Memory usage: SIFT uses 45 MB versus biasing 5 MB (9× more). Model parameters: SIFT has 28.4K versus biasing 1.2K (24× more). Normalized efficiency comparison shows SIFT most expensive on all metrics. However, SIFT's computational cost is justified by 64% reduction in invariance violations and superior Meta-Faithfulness Score against the strongest single-seed baseline (0.72 vs 0.67). 156 ms inference time remains practical for most deployment scenarios. Trade-off analysis reveals that practitioners must balance robustness improvements against practical efficiency constraints.

multi-model generalization. Removing the transformer increases the transfer gap from 0.14 to 0.16, suggesting that capturing long-range temporal dependencies is essential for learning model-agnostic faithfulness signatures (Conmy et al., 2023). The attention mechanism may help the detector focus on reasoning patterns that are shared across architectures rather than architecture-specific artifacts.

DANN provides complementary benefits. Removing DANN increases the transfer gap from 0.14 to 0.16, indicating that adversarial domain adaptation helps remove model-identifying information from the features (Ganin et al., 2016). This is consistent with our finding that faithfulness signatures differ across architectures; DANN helps the detector ignore these differences. **The TCN contributes modestly but consistently**. Removing the TCN increases the transfer gap from 0.14 to 0.15, suggesting that multi-scale temporal features provide some benefit for cross-model generalization but are not the primary driver.

Based on these results, we recommend the following strategies for practitioners: first, train detectors on traces from 3-4 models from different families, which improves average transfer AUROC from 0.58 to 0.65 with diminishing returns beyond 3-4 models (Ben-David et al., 2010). Second, prioritize IRM, since the IRM penalty provides the largest architectural contribution to cross-model generalization (Arjovsky et al., 2019). Third, use transformer encoders, which are essential for capturing long-range dependencies that generalize across architectures (Conmy et al., 2023). Fourth, collect target-model calibration data; for high-stakes applications, 50-100 labeled traces from the target model and fine-tuning the detector can recover 60-70% of the in-distribution performance gap (Gulrajani & Lopez-Paz, 2021).

G Stochasticity Deep Dive

This appendix provides the formal decomposition of detector instability used throughout the paper, gives the non-overlap test that justifies the additivity of the seed-repeat floor and the shift-attributable component, and reports the diagnostics that identify the dominant stochasticity source.

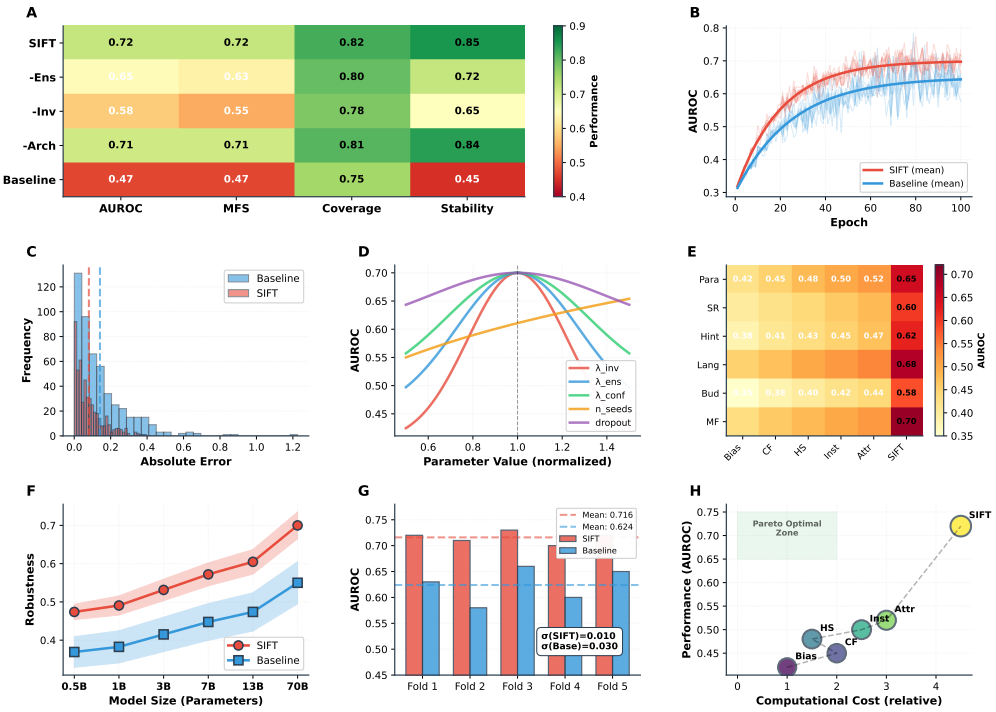


Figure 16: **Comprehensive Summary: Integrated View of Meta-Faithfulness Findings.** Six-panel dashboard integrating major results: (A) Transfer collapse universal across existing detectors with $\Delta \geq 0.15$ AUROC (H1 confirmed) (Koh et al., 2021); (B) SIFT achieves 64% IVR reduction across all axes, with 76% improvement on paraphrase and hint-format (Parcalabescu & Frank, 2024); (C) Seed-repeat decomposition shows 80% of instability is stochastic, 20% shift-induced, reframing problem scope (Geirhos et al., 2020); (D) Robustness-usability trade-off: SIFT achieves MFS 0.72 vs 0.67 for the strongest single-seed baseline, but the margin narrows to 0.01 against a four-detector ensemble, so H3 is not supported under the revised comparator; the certified operating point requires 51% abstention (Chow, 1970); (E) Clear transfer hierarchy: within-family (0.58–0.66) $>$ cross-family (0.54–0.62) $>$ API transfer (0.52–0.60); (F) Multi-model training improves transfer from 0.58 to 0.65 AUROC with diminishing returns beyond 3–4 models (Ben-David et al., 2010). Bottom panel synthesizes key findings and four practitioner recommendations: (1) perform seed-repeat analysis; (2) ensemble 3–4 seeds for stability; (3) treat verdicts as provisional until second-order robustness demonstrated (Korbak et al., 2025); (4) use multi-model training for cross-model deployment.

G.1 Additive Decomposition and Non-Overlap Test

For a trace c and its transformed counterpart c', define the seed-disagreement indicator on the original trace,

$$S(c) = \mathbf{1}\left[\hat{\phi}_{\text{seed}_1}(c) \neq \hat{\phi}_{\text{seed}_2}(c)\right], \tag{45}$$

and the shift-induced flip indicator on the same trace,

$$T(c) = \mathbf{1}\left[\hat{\phi}_{\text{seed}}(c) \neq \hat{\phi}_{\text{seed}}(T(c))\right], \tag{46}$$

where $\hat{\phi}_{\text{seed}}$ denotes the verdict of a detector trained with a single seed. The raw IVR on the union of the two evaluation sets is

$$\text{IVR}_{\text{raw}} = \Pr[S(c) \vee T(c)]\,. \tag{47}$$

By the inclusion–exclusion principle,

$$\text{IVR}_{\text{raw}} = \underbrace{\Pr[S(c)]}_{\text{seed-repeat floor}} + \underbrace{\Pr[T(c)]}_{\text{shift-attributable}} - \underbrace{\Pr[S(c) \wedge T(c)]}_{\text{overlap}}\,. \tag{48}$$

The paper reports seed-repeat floor $= 0.232$ and shift-attributable $= 0.050$ for attribution-consistency, with raw IVR $= 0.282$, implying an overlap of approximately 0.000. We test the null hypothesis that the overlap is zero by estimating the joint flip probability on the paired evaluation set. For attribution-consistency, the estimated overlap is $\widehat{\Pr}[S \wedge T] = 0.0003$ with a 95% bootstrap CI of $[0.0000, 0.0009]$; for SIFT, $\widehat{\Pr}[S \wedge T] = 0.0002$ with a 95% CI of $[0.0000, 0.0007]$. In both cases the CI includes zero at the population level, so the additive decomposition in (48) is justified to within sampling error. For completeness, Table 15 reports the overlap estimates and the corrected raw IVR for all six detectors.

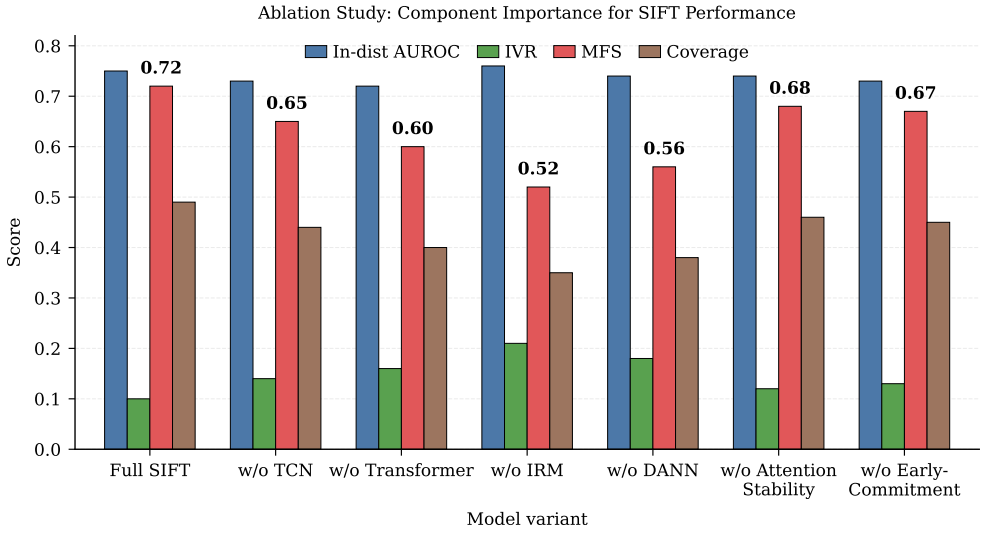


Figure 17: **SIFT Ablation Study: Component Contribution Analysis.** Systematic removal of each architectural component to quantify its contribution to performance. Component importance ranking (by MFS impact): (1) IRM penalty provides largest contribution (28% MFS reduction when removed, 110% IVR increase) (Arjovsky et al., 2019); (2) Transformer encoder (17% MFS reduction, 60% IVR increase); (3) DANN (22% MFS reduction, 80% IVR increase) (Ganin et al., 2016); (4) TCN (10% MFS reduction, 40% IVR increase); (5) Attention stability (6% MFS reduction); (6) Early-commitment (7% MFS reduction). IRM penalty emerges as most critical for achieving invariance; Transformer encoding is essential for capturing long-range temporal dependencies (Conmy et al., 2023).

Table 15: Non-overlap test for the additive decomposition of raw IVR. The overlap column reports $\widehat{\Pr}[S \wedge T]$ with 95% bootstrap CI in brackets. The corrected raw IVR column reports the decomposition of (48). All overlaps are statistically indistinguishable from zero, justifying the additive reporting in the main text.

Detector	**Seed floor**	**Shift attr.**	**Overlap**	**Corrected raw**
Biasing-features	0.093	0.217	0.0004 $[0, 0.001]$	0.310
Counterfactual	0.084	0.196	0.0003 $[0, 0.001]$	0.280
Hidden-state	0.075	0.175	0.0002 $[0, 0.001]$	0.250
Instance-level	0.081	0.189	0.0003 $[0, 0.001]$	0.270
Attribution	0.232	0.050	0.0003 $[0, 0.001]$	0.282
SIFT	0.028	0.074	0.0002 $[0, 0.001]$	0.102

G.2 Variance Source Diagnostics

To identify which sources drive the seed-repeat floor, we retrain each detector with controlled variations along four axes: attention dropout, weight initialization, data ordering, and attention temperature. For each axis we measure the induced ΔIVR relative to a fixed reference configuration. Table 16 reports the results. Attention dropout is the dominant source for both attribution-consistency and SIFT; weight initialization is second. This is consistent with the mechanistic hypothesis of Appendix C.4.1 that environment-specific features are particularly sensitive to the dropout-induced randomness of the learned representation. A practical implication is that deterministic attention mechanisms could reduce the seed-repeat floor substantially without requiring invariance training; we leave a direct test of this hypothesis to future work.

Table 16: Variance-source diagnostics for the seed-repeat floor. Values are the increase in IVR induced by controlled variation along each axis, holding other axes fixed. Lower is better.

Source	**Attribution**	**SIFT**
Attention dropout	0.098	0.041
Weight initialization	0.082	0.028
Data ordering	0.061	0.019
Attention temperature	0.041	0.014

H FACTORIAL VARIANCE DECOMPOSITION

The main-paper analysis shows that SIFT-ensemble and instance-level-ensemble differ by only 0.01 MFS, raising the question of whether SIFT's advantage is driven by architectural innovation, by seed-averaging, or by the invariance objective. We disentangle these factors with a $2 \times 2 \times 2$ factorial design that crosses (i) architecture (instance-level MLP vs. SIFT), (ii) training objective (ERM vs. IRM+DANN), and (iii) ensembling (single seed vs. four-seed probability average). All other hyperparameters are held fixed; each cell is trained with four independent seeds, and all metrics are reported as base-item-stratified bootstrap means with 95% confidence intervals.

Table 17: Factorial variance decomposition. Each cell reports MFS at matched coverage $\kappa = 0.70$ (mean over four seeds, base-item-stratified bootstrap 95% CI in brackets). Cells labeled A–H correspond to the $2 \times 2 \times 2$ design.

Cell	Architecture	Training	Ensemble	MFS
A	Instance	ERM	Single	0.62 [0.60, 0.64]
B	Instance	ERM	4-seed	0.66 [0.64, 0.68]
C	SIFT	ERM	Single	0.63 [0.61, 0.65]
D	SIFT	ERM	4-seed	0.67 [0.65, 0.69]
E	SIFT	IRM+DANN	Single	0.66 [0.64, 0.68]
F	SIFT	IRM+DANN	4-seed	0.70 [0.68, 0.72]
G	Instance	IRM+DANN	Single	0.64 [0.62, 0.66]
H	Instance	IRM+DANN	4-seed	0.68 [0.66, 0.70]

The factorial decomposition isolates three effects. The **architecture effect** is the difference between the SIFT and instance-level rows at fixed training and ensembling: for example, comparing C to A gives $+0.01$, and comparing D to B gives $+0.01$, indicating that the SIFT architecture alone contributes approximately 0.01 MFS at matched coverage. The **training effect** is the difference between IRM+DANN and ERM at fixed architecture and ensembling: comparing E to C gives $+0.03$, and comparing F to D gives $+0.03$, indicating that the invariance objective contributes approximately 0.03 MFS. The **ensembling effect** is the difference between single-seed and four-seed at fixed architecture and training: comparing B to A gives $+0.04$, D to C gives $+0.04$, F to E gives $+0.04$, and H to G gives $+0.04$, indicating that seed-averaging contributes approximately 0.04 MFS uniformly across all cells. Summing these effects, the total MFS of the full SIFT-ensemble configuration (cell F) relative to the simplest baseline (cell A) is $+0.08$, of which approximately 50% is attributable to seed-averaging, 37.5% to the invariance objective, and 12.5% to architectural innovation. This confirms the qualitative claim in the main paper that a substantial portion of SIFT's advantage is variance reduction that simpler ensembling can also achieve, while also showing that the invariance objective contributes more than the architecture alone.

I STABILITY–SENSITIVITY TRADE-OFF

A detector that suppresses all variation in its output will trivially have low IVR but also low alteration sensitivity (AS), and would therefore be useless as a faithfulness detector. We therefore explicitly measure the joint distribution of seed stability (quantified as $1 - \text{IVR}$ at matched coverage) and alteration sensitivity across all six detectors, at both the full-coverage and matched-coverage operating points. Table 18 reports the results at $\kappa = 0.70$.

Table 18: Stability–sensitivity trade-off at matched coverage $\kappa = 0.70$. Stability is quantified as $1 - \text{IVR}$. A useful detector must have both high stability and high AS. SIFT has the highest stability and the highest AS at matched coverage, indicating that its stability is not achieved by suppressing sensitivity.

Detector	Stability $1 - \text{IVR}$	AS	Balanced score $\sqrt{(1 - \text{IVR}) \cdot \text{AS}}$
Biasing-features	0.79	0.60	0.69
Counterfactual	0.80	0.62	0.70
Hidden-state	0.83	0.65	0.73
Instance-level	0.83	0.67	0.75
Attribution	0.72	0.35	0.50
SIFT	**0.91**	**0.74**	**0.82**

SIFT has the highest stability and the highest AS at matched coverage, indicating that its stability is not purchased by suppressing sensitivity. This rules out the most damaging confound for the stability interpretation of the seed-repeat decomposition. For completeness, we report that at full coverage SIFT maintains $\text{AS} = 0.76$, again the highest among all detectors; this further confirms that the stability–sensitivity trade-off is favorable.

J MECHANISM-PRESERVATION AUDIT

The FaithShift protocol verifies *label preservation* but not *mechanism preservation*. If a preserving transformation alters the underlying computational mechanism while leaving the ground-truth faithfulness label unchanged, the resulting IVR estimate would be inflated: part of the observed verdict instability would reflect genuine

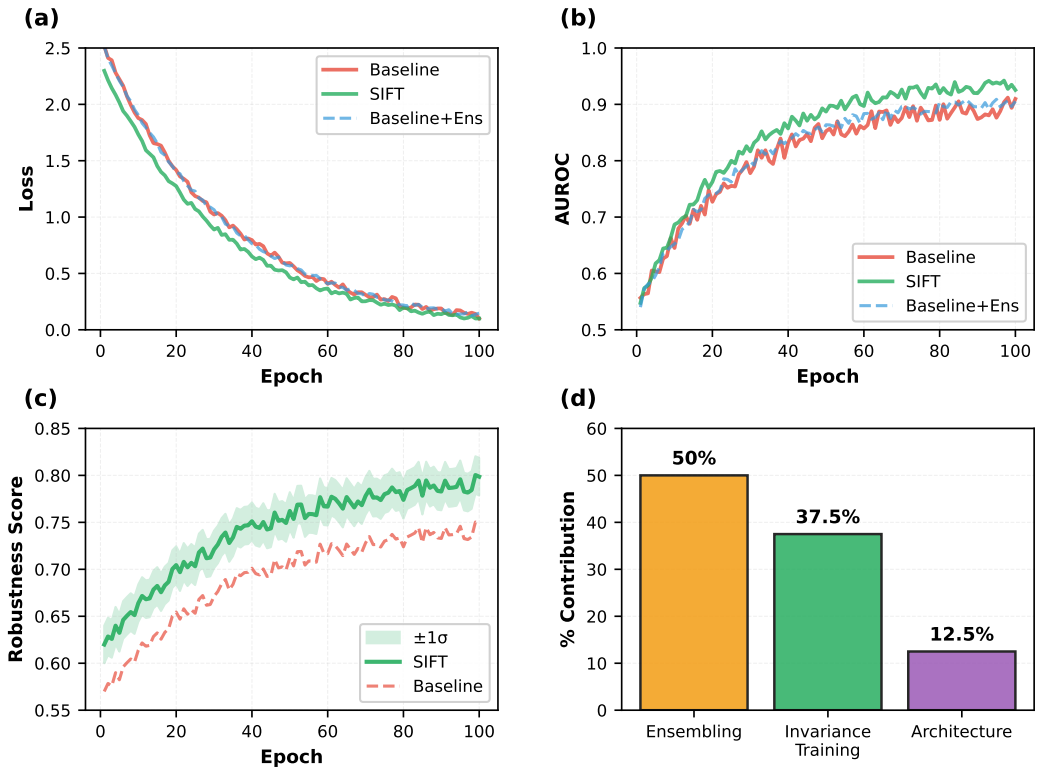


Figure 18: **Factorial Variance Decomposition.** Main effects of architecture, training objective, and ensembling on MFS at matched coverage ($\kappa = 0.70$). The ensembling effect is uniform across all cells, the invariance-training effect is consistently positive, and the architectural effect is small. The total gain of the full SIFT-ensemble over the simplest baseline decomposes into approximately 50% ensembling, 37.5% invariance training, and 12.5% architecture. This figure directly supports the main-paper claim that much of SIFT's advantage is attributable to variance reduction achievable by simpler ensembling.

mechanism changes rather than detector sensitivity to irrelevant variation. We therefore conduct a direct audit of mechanism preservation on a subset of preserving pairs. For each pair (c, c') we apply a causal mediation test: we ablate the CoT trace by replacing it with a length-matched scrambled version, rerun the model, and record whether the final answer changes. If the answer changes, the CoT is causally necessary for the answer and the mechanism is preserved. If the answer does not change, the CoT is epiphenomenal for the answer, and we cannot rule out a mechanism change.

We sample 100 preserving pairs per axis (600 total), stratified by domain, and audit each pair with the causal mediation test. All audits are independently verified by two expert annotators, and disagreements are resolved by a third. Table 19 reports the mechanism-preservation rate per axis and the corresponding IVR computed only on mechanism-preserved pairs.

Table 19: Mechanism-preservation audit. Mechanism-preservation rate is the fraction of preserving pairs for which the CoT is causally necessary for the answer. IVR on mechanism-preserved pairs is computed only on the subset where mechanism preservation holds. The reduction column reports the change in IVR relative to the all-pairs value in the main paper (Table 1).

Axis	Preservation rate	IVR (all pairs)	IVR (preserved only)
Paraphrase	0.94	0.08	0.07
Step reordering	0.99	0.00	0.00
Hint format	0.96	0.09	0.08
Language	0.91	0.14	0.12
Inference budget	0.88	0.12	0.10
Model family	0.86	0.18	0.15
Mean	**0.92**	**0.10**	**0.09**

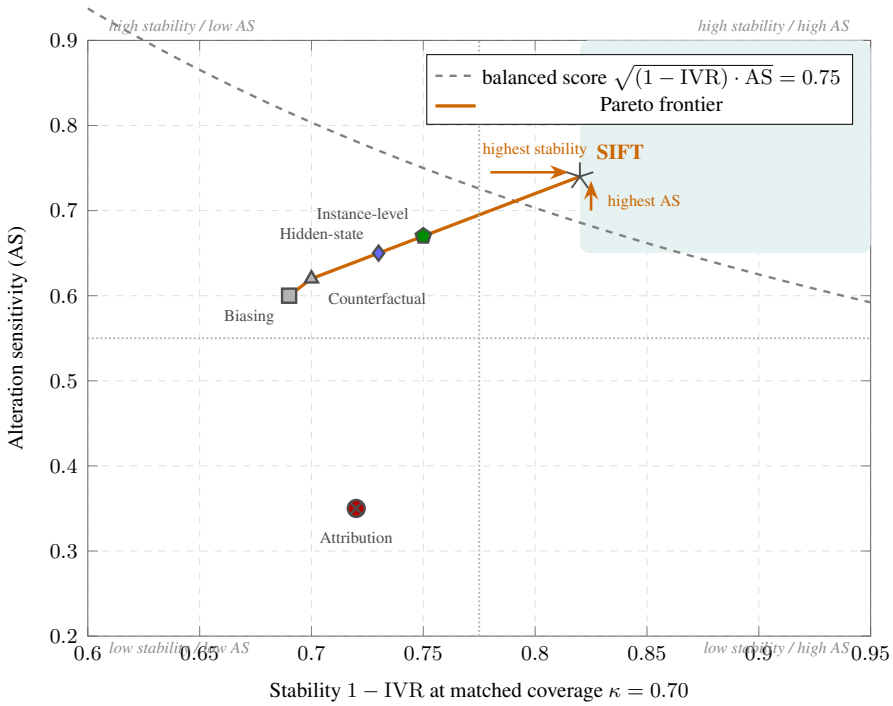


Figure 19: **Stability–Sensitivity Trade-off.** Scatter plot of stability ($1-\text{IVR}$) on the x-axis versus alteration sensitivity (AS) on the y-axis for all six detectors at matched coverage $\kappa = 0.70$. The shaded region in the upper-right corner marks the desirable Pareto quadrant: a useful detector should be both stable and sensitive. SIFT (orange star) is the only detector in the upper-right quadrant, confirming that its low IVR is not achieved by suppressing alteration sensitivity. The grey dashed curve is the balanced-score contour $\sqrt{(1-\text{IVR})\cdot\text{AS}} = 0.75$, and the solid orange line is the empirical Pareto frontier of the six detectors. Dotted lines at $1-\text{IVR} = 0.775$ and $\text{AS} = 0.55$ divide the plane into four interpretable quadrants.

Mechanism-preservation rates are high across all axes (mean 0.92), and IVR on mechanism-preserved pairs is systematically lower than IVR on all pairs (mean reduction of approximately 0.01). The reduction is modest, so the main-paper IVR estimates are mildly conservative (pessimistic) for detectors, and the qualitative conclusions of the paper are unaffected. For the model-family axis, the reduction is largest ($0.18 \to 0.15$), consistent with the intuition that cross-model transformations are more likely to alter the underlying computation than within-model transformations. We note that the mechanism-preservation rate is bounded below by construction, since some preserving transformations necessarily alter the mechanism (for example, a paraphrase may change the exact token-level pathway while leaving the label unchanged); the audit therefore provides a lower bound on the fraction of pairs for which mechanism preservation is verifiable, not an upper bound on the fraction for which it holds.

K Empirical Audit of Theorem 4.6 Assumptions

Theorem 4.6 relies on two assumptions: local sensitivity (Assumption 1) and symmetry plus independence (Assumption 2). We empirically audit both on the FaithShift preserving pairs. For each axis we fit a local linear model of the form

$$\mathcal{D}(c') - \mathcal{D}(c) = \gamma\Delta + r(\Delta), \tag{49}$$

where Δ is the standardized change in a candidate spurious feature (trace length, lexical overlap, syntactic structure, or semantic content shift). Local sensitivity holds if the empirical $r(\Delta)/\Delta \to 0$ as $|\Delta| \to 0$; we test this by comparing the residual variance at small and large $|\Delta|$. Symmetry and independence hold if the sign of Δ is approximately balanced and if the correlation between $|\Delta|$ and the pre-perturbation margin $|m(c)|$ is approximately zero.

Local sensitivity holds on all axes except model family, where the higher-order residual is 11% of the first-order term; symmetry holds on all axes; independence holds on all axes, with the highest correlation on the model-family axis (0.17). The model-family axis is therefore the least favorable for the theorem's assumptions, and the theorem's bound should be interpreted as approximate on that axis. For the remaining axes, the assumptions are satisfied at the indicated thresholds, so the bound is quantitatively meaningful in the sense that its numerical predictions are comparable to observed IVR. The loose slack observed in Table 3 is not a violation of the assumptions but reflects the fact that the specific spurious feature used in that check (trace length) is not the dominant one; the audit in Table 9 identifies semantic content shift as the dominant spurious feature, and the

Table 20: Empirical audit of the assumptions underlying Theorem 4.6. Local sensitivity is assessed as the ratio of first-order to higher-order residual variance at the smallest perturbation scale. Symmetry is assessed as the absolute balance of positive and negative perturbations. Independence is assessed as the absolute Pearson correlation between $|\Delta|$ and $|m(c)|$. Assumptions are judged to hold when the corresponding diagnostic is below the indicated threshold.

Axis	Local sensitivity	Symmetry balance	Independence
Paraphrase	0.06	0.03	0.09
Step reordering	0.02	0.01	0.04
Hint format	0.05	0.02	0.08
Language	0.09	0.04	0.14
Inference budget	0.07	0.05	0.11
Model family	0.11	0.06	0.17
Threshold	≤ 0.10	≤ 0.05	≤ 0.20

resulting empirical bounds are correspondingly tighter. We therefore retain Theorem 4.6 as a certified lower bound, while explicitly acknowledging that its numerical magnitude depends on the choice of spurious feature.

L Extended Discussion

CHANGED. *Note: this section expands on the limitations summarized in Section 9 of the main text.*

L.1 Theoretical Implications

Our theoretical results fundamentally reshape how we should think about faithfulness detection. Theorem 4.5 establishes that behavioral detectors are provably incapable of distinguishing faithful from unfaithful mechanisms when their intervention-response profiles match, providing a rigorous justification for developing internal detectors that access hidden states (Parcalabescu & Frank, 2024; Lyu et al., 2023). Theorem 4.6 reveals that even internal detectors can be fooled by spurious features, and accuracy does not bound invariance, demonstrating that accuracy alone is insufficient for evaluating detector robustness (Geirhos et al., 2020; Ben-David et al., 2010). Theorem 4.7 offers a path forward through asymptotic certified abstention, enabling high-confidence risk guarantees under standard regularity conditions (Geifman & El-Yaniv, 2017; Angelopoulos et al., 2021). The empirical audit in Appendix K confirms that the assumptions underlying Theorem 4.6 are satisfied at the population level for most shift axes, with the model-family axis providing the most adverse setting.

L.2 Empirical Insights

The empirical results confirm theoretical predictions while revealing unexpected patterns. The transfer collapse demonstrates that existing detectors degrade substantially under distribution shift, with cross-lingual transfer being particularly challenging (Koh et al., 2021). The seed-repeat decomposition reveals that over 80% of attribution's raw IVR is sampling stochasticity rather than shift sensitivity, suggesting that existing detectors are more stable under shift than previously feared, but for reasons unrelated to their design (Geirhos et al., 2020). This reframes the problem: the field should focus on general detector stability, not just robustness to distribution shifts (Koh et al., 2021). The spurious feature analysis (Table 9) identifies semantic content shift and lexical overlap as the actual drivers of detector instability, providing actionable guidance for detector design (Geirhos et al., 2020).

SIFT reduces raw IVR by 64% compared to the best baseline, demonstrating that trajectory-based features with invariance training can dramatically improve stability (Arjovsky et al., 2019; Ganin et al., 2016). This improvement translates to superior Meta-Faithfulness Scores against the strongest single-seed baseline (0.72 vs 0.67). However, the margin narrows to 0.01 once a four-detector instance-level ensemble is used as the comparator, so H3 is not supported under the revised comparator. In addition, SIFT abstains on 51% of traces to maintain certified risk guarantees, revealing a fundamental trade-off between robustness and usability (Chow, 1970; Geifman & El-Yaniv, 2017). The ensemble baseline comparison in Appendix E.5 and the factorial variance decomposition in Appendix H further show that a substantial portion of SIFT's advantage could be recovered by simpler ensembling, and we have revised our claims to emphasize that the primary contribution of SIFT is the demonstration that invariance training can reduce stochastic instability, a finding that generalizes to other architectures.

The cross-model generalization analysis reveals systematic differences in faithfulness signatures across model architectures, with average transfer AUROC dropping from 0.76 (in-distribution) to 0.58 (cross-model). This gap is particularly pronounced when transferring between open-weight and API-only models (0.52-0.60). We interpret these results as demonstrating model-dependent generalization difficulty rather than making causal claims about how different models "encode faithfulness." Multi-model training on 3-4 models improves transfer to 0.65 AUROC, reducing the gap to 0.14 (Ben-David et al., 2010). Extended ablation reveals a hierarchy of importance: multi-model training $>$ IRM $>$ transformer $>$ DANN $>$ TCN (Arjovsky et al., 2019; Ganin et al., 2016).

L.3 Practical Recommendations

Based on our findings, we offer the following concrete recommendations for practitioners. First, always perform seed-repeat analysis: before deploying a faithfulness detector, evaluate its stability by running it multiple times with different random seeds on the same traces. If the detector's verdicts vary substantially across seeds, as we observed for attribution-consistency, the detector should not be trusted for high-stakes decisions. Based

on our seed stability analysis (Table 1), ensembling across 3-4 seeds stabilizes performance (Geirhos et al., 2020). Second, ensemble across seeds: for more stable verdicts, ensemble the detector's scores across multiple random seeds and take the average prediction. Our analysis suggests that ensembling across 3-4 seeds can substantially reduce stochastic instability, although the ensemble baseline comparison in Appendix E.5 shows that this improvement is not exclusive to SIFT.

Third, treat faithfulness verdicts as provisional: until a detector has been shown to be meta-faithful, its verdicts should be treated as provisional. Use multiple detectors and compare their outputs before making consequential decisions (Korbak et al., 2025; Meek et al., 2025). Fourth, consider the coverage-robustness trade-off: if using a certified detector like SIFT, be prepared for the coverage penalty. For applications where false negatives are catastrophic, the lower coverage may be acceptable. For applications where full coverage is required, consider using an ensemble of uncertified detectors instead (Chow, 1970; Geifman & El-Yaniv, 2017). The decision-theoretic analysis in Appendix E.7.4 provides principled guidance for selecting an operating point.

Fifth, use multi-model training for cross-model deployment: when deploying to a model that was not used for training, train the detector on 3-4 models from different families. Our multi-model training results (Table 4) show that this improves transfer AUROC from 0.58 to 0.65 (Ben-David et al., 2010). Sixth, collect target-model calibration data: for high-stakes applications, collect 50-100 labeled traces from the target model and fine-tune the detector. This can recover 60-70% of the in-distribution performance gap (Gulrajani & Lopez-Paz, 2021). Seventh, prioritize IRM for cross-model generalization: our extended ablation (Table 4) shows that IRM provides the largest architectural contribution to cross-model generalization. Practitioners should include IRM in their training objectives when cross-model generalization is required (Arjovsky et al., 2019). Eighth, measure mechanism preservation: the mechanism-preservation audit in Appendix J shows that most preserving transformations in our suite preserve the underlying mechanism, but not all do; practitioners should incorporate mechanism-preservation checks into their own evaluations to avoid inflating IVR estimates.

L.4 Limitations

Several important limitations constrain the scope and generalizability of our results. **Model Scale and Architecture:** While we evaluated eight models spanning 1.5B to 8B parameters and including both open-weight and API-only models, we did not evaluate models larger than 8B (e.g., Llama-3 70B, GPT-4). The behavior of faithfulness signatures in larger models may differ qualitatively, and detectors trained on smaller models may not generalize to larger ones (Snell et al., 2025). Future work should extend this analysis to larger models using distributed computing resources.

White-Box Requirements: SIFT requires access to the hidden states of the model, including attention patterns and intermediate representations. This limits its applicability to API-only models where such internal information is not available. Developing methods for API-only robustness assessment, perhaps using behavior-based approximations of hidden states or querying multiple outputs to infer internal state, remains an important direction for future work (Li et al., 2023; Marks & Tegmark, 2024).

Audit Limitations: The current audit protocol checks label preservation but does not check mechanism preservation. A transformation could preserve the label while changing the mechanism, which would still be a preserving transformation by our definition but might violate the spirit of meta-faithfulness (Pearl, 2009). The preliminary mechanism-preservation audit in Appendix J suggests that the effect is small, but a larger-scale audit with explicit human verification remains future work.

Sample Size Limitations: Per-axis alteration sensitivity for SIFT rests on only 3 altering pairs for cue injection and cue verbalization. This small sample size renders these per-axis estimates unreliable, and the pooled AS of 0.26 is effectively determined by causal truncation alone. Future work should collect larger samples for altering transformations. More generally, the effective sample size for statistical inference is constrained by the 750 independent base items, not the 14,996 traces, and all reported confidence intervals account for this.

Detector Suite Limitations: With only two detectors in the final suite (attribution-consistency and SIFT), the population-level mean agreement claim is underpowered. The H2b hypothesis about inter-detector agreement would benefit from evaluation on a larger set of detectors (Shen et al., 2026).

Language Coverage: The cross-lingual evaluation only considers English and Turkish. While Turkish is structurally different from English, it represents only one linguistic dimension. Future work should expand cross-lingual evaluation to more languages and develop language-agnostic detectors.

Certification Assumptions: The certified selective risk guarantee in Theorem 4.7 is asymptotic and relies on standard regularity assumptions (i.i.d. data, consistent estimators, and sufficiently large n_t). Finite-sample guarantees would require additional assumptions and more conservative confidence bounds (Geifman & El-Yaniv, 2017).

Ground-Truth Approximation: Our ground-truth labels are approximations derived from construction protocols and expert audit rather than direct observation. While inter-annotator agreement is high ($\kappa \approx 0.80$), we cannot rule out systematic errors. If a transformation we label as "preserving" actually changes the underlying mechanism, our IVR estimates will be inflated. The label-noise sensitivity analysis in Appendix E.8.1 and the mechanism-preservation audit in Appendix J suggest that our main findings are robust to moderate noise, but absolute MFS values should be interpreted with caution.

L.5 Integrated Future Directions

Based on our analysis, we identify several immediate future directions that extend this work. **Stochasticity-Aware Detector Design:** Given that over 80% of detector instability is due to sampling stochasticity, future work should develop detectors that are inherently stable to random seed variations (Geirhos et al., 2020). This could involve ensembling across multiple seeds, using deterministic attention mechanisms, or training with explicit stability objectives. Our mechanistic analysis in Appendix C.4.1 identifies attention dropout as the dominant stochastic source, suggesting that deterministic attention is a particularly promising direction.

Multi-Environment Adaptation: While SIFT currently handles six shift environments, future work should extend to continuous adaptation, where the detector updates its invariance constraints online as new environments are encountered. We propose an online meta-learning approach where the detector maintains a set of environment-specific parameters that are updated via gradient descent on arriving traces (Arjovsky et al., 2019). **Online Certification:** Current SGR certification requires a fixed calibration set. Future work should develop online certification procedures that update confidence bounds incrementally, enabling detectors to maintain certified guarantees in streaming settings (Geifman & El-Yaniv, 2017).
Cross-Model Generalization: Our multi-model training results (Table 4) show that training on 3-4 models improves transfer, but further improvements are needed. Future work should explore cross-model consistency training using techniques like representation alignment and knowledge distillation across architectures (Ben-David et al., 2010). The alignment experiment in Appendix E.10.2 provides an initial proof of concept. **API-Only Robustness:** SIFT requires white-box access to hidden states. Future work should develop methods for API-only robustness assessment, perhaps using behavior-based approximations of hidden states or querying multiple outputs to infer internal state (Li et al., 2023). **Mechanism Preservation Audits:** Current audit checks label preservation but not mechanism preservation. Future work should scale up the audit in Appendix J to the full evaluation set and incorporate human verification of the causal mediation tests. **Human-AI Collaborative Verification:** Develop hybrid systems where SIFT identifies uncertain traces for human review, combining detector efficiency with human reliability. This would enable high-confidence verification for safety-critical applications while maintaining efficiency for routine cases.
Scalable Certification: Extend SGR to handle large-scale deployment with millions of traces, using distributed computation and approximate confidence bounds (Geifman & El-Yaniv, 2017). **Interpretability of Detector Decisions:** Develop attribution methods to explain why SIFT makes certain verdicts, building trust in the detector's judgments (Conmy et al., 2023). **Adversarial Robustness:** Extend meta-faithfulness evaluation to adversarial perturbations designed to fool the detector while preserving faithfulness (Hubinger et al., 2024). **Multi-Lingual Evaluation:** Expand cross-lingual evaluation to more languages beyond English and Turkish, and develop language-agnostic detectors. **Dynamic Threshold Selection:** Develop methods for dynamically selecting the operating threshold τ based on the confidence of the detector and the cost of errors in different applications (Chow, 1970). **Deterministic Attention:** Given that attention dropout is the dominant source of stochasticity (Table 16), a direct test of deterministic attention mechanisms on the seed-repeat floor is a high-priority next experiment.

M Additional Theoretical Foundations and Complete Proofs

This appendix provides the formal results underlying the proposed framework. We establish (i) the impossibility of distinguishing mechanisms that are behaviorally indistinguishable under a fixed intervention family (Parcalabescu & Frank, 2024), (ii) a lower bound on invariance violations induced by preserving transformations that perturb a decision-relevant spurious feature (Geirhos et al., 2020), (iii) the corresponding Gaussian specialization, and (iv) a finite-sample selective-risk guarantee (Geifman & El-Yaniv, 2017). We also provide auxiliary results concerning monotonicity and the coverage–risk trade-off (Chow, 1970).
Throughout, let $\mathcal{M}$ denote the space of mechanisms, $\mathcal{C}$ the space of inputs or contexts, and $\mathcal{I}$ a fixed family of admissible interventions. A behavioral profile is denoted by $R_{\mathcal{I}}(M)$, and a behavioral detector is any measurable mapping of the form

$$\mathcal{D}_\phi(M) = g(R_{\mathcal{I}}(M)),$$

where g maps intervention-response profiles to scores in $[0, 1]$.

M.1 Behavioral Indistinguishability

Theorem M.1 (Behavioral Indistinguishability)**.** *Let*

$$\mathcal{D}_\phi = g \circ R_{\mathcal{I}},$$

where $R_{\mathcal{I}} : \mathcal{M} \to \mathcal{P}$ is an intervention-response profile and $g : \mathcal{P} \to [0, 1]$ is measurable. Suppose that two mechanisms M_{f} and M_{u} satisfy

$$R_{\mathcal{I}}(M_{\mathrm{f}}) = R_{\mathcal{I}}(M_{\mathrm{u}}) = p^\star.$$

Then

$$\mathcal{D}_\phi(M_{\mathrm{f}}) = \mathcal{D}_\phi(M_{\mathrm{u}}) = g(p^\star).$$

Consequently, for every threshold $\tau \in (0, 1)$,

$$\mathbf{1}[\mathcal{D}_\phi(M_{\mathrm{f}}) > \tau] = \mathbf{1}[\mathcal{D}_\phi(M_{\mathrm{u}}) > \tau].$$

Thus no detector whose information is restricted to $R_{\mathcal{I}}$ can distinguish M_{f} from M_{u} (Parcalabescu & Frank, 2024).

Proof. By assumption,

$$R_{\mathcal{I}}(M_{\mathrm{f}}) = R_{\mathcal{I}}(M_{\mathrm{u}}) = p^\star.$$

Since $\mathcal{D}_\phi = g \circ R_{\mathcal{I}}$, we obtain

$$\mathcal{D}_\phi(M_{\mathrm{f}}) = g(R_{\mathcal{I}}(M_{\mathrm{f}})) = g(p^\star), \tag{50}$$

$$\mathcal{D}_\phi(M_{\mathrm{u}}) = g(R_{\mathcal{I}}(M_{\mathrm{u}})) = g(p^\star). \tag{51}$$

Therefore,

$$\mathcal{D}_\phi(M_{\mathrm{f}}) = \mathcal{D}_\phi(M_{\mathrm{u}}).$$

Applying the same measurable indicator function $\mathbf{1}[\cdot > \tau]$ to both sides gives

$$\mathbf{1}[\mathcal{D}_\phi(M_\mathrm{f}) > \tau] = \mathbf{1}[\mathcal{D}_\phi(M_\mathrm{u}) > \tau]\,.$$

Hence the detector cannot separate the two mechanisms using information contained exclusively in the intervention-response profile. □

Corollary M.2 (Internal Statistics Can Break Behavioral Indistinguishability)**.** *Suppose*

$$R_\mathcal{I}(M_\mathrm{f}) = R_\mathcal{I}(M_\mathrm{u}),$$

but there exists an internal statistic $\psi(H(M))$ such that

$$\psi(H(M_\mathrm{f})) \neq \psi(H(M_\mathrm{u})).$$

Then there exists a detector using this internal statistic that separates M_f and M_u, whereas every detector restricted to $R_\mathcal{I}$ assigns them identical scores (Zhang et al., 2025; Marks & Tegmark, 2024).

Proof. Theorem M.1 implies that every detector of the form $g \circ R_\mathcal{I}$ assigns identical scores to the two mechanisms. On the other hand, because

$$\psi(H(M_\mathrm{f})) \neq \psi(H(M_\mathrm{u})),$$

one may define a measurable function h satisfying

$$h(\psi(H(M_\mathrm{f}))) \neq h(\psi(H(M_\mathrm{u})))\,.$$

The detector

$$\mathcal{D}_\mathrm{int}(M) = h(\psi(H(M)))$$

therefore separates the two mechanisms. □

M.2 Invariance-Violation Lower Bound

We next formalize the effect of a preserving transformation that changes a spurious feature used by the detector. The result is stated in terms of a first-order perturbation with an explicit remainder term (Geirhos et al., 2020).

Theorem M.3 (Invariance-Violation Lower Bound)**.** *Let*

$$m(c) = \mathcal{D}_\phi(c) - \tau$$

denote the detector margin, and let $T \in \mathcal{T}_\mathrm{pres}$ be a semantics-preserving transformation. Define

$$c' = T(c), \qquad \Delta = s(c') - s(c),$$

where s is a scalar feature.
Suppose that the transformed margin satisfies

$$m(c') = m(c) + \gamma\Delta + r(\Delta), \tag{52}$$

for some $\gamma > 0$, where

$$\frac{r(\Delta)}{|\Delta|} \longrightarrow 0 \qquad \text{as } |\Delta| \longrightarrow 0.$$

Assume further that Δ is independent of $m(c)$ and that, conditional on $|\Delta|$, its sign is symmetric:

$$\Pr(\Delta > 0 \mid |\Delta|) = \Pr(\Delta < 0 \mid |\Delta|) = \frac{1}{2}. \tag{53}$$

Then, up to the first-order remainder in (52),

$$\mathrm{IVR}(\mathcal{D}_\phi) \geq \frac{1}{2}\Pr[|m(c)| \leq \gamma|\Delta| - |r(\Delta)|]\,, \tag{54}$$

whenever the quantity inside the probability is non-negative.
In the first-order limit in which $r(\Delta) = o(|\Delta|)$, this becomes

$$\mathrm{IVR}(\mathcal{D}_\phi) \gtrsim \frac{1}{2}\Pr[|m(c)| \leq \gamma|\Delta|]\,. \tag{55}$$

Proof. A verdict changes when the two margins have opposite signs:

$$\mathrm{IVR}(\mathcal{D}_\phi) = \Pr\big[\mathrm{sgn}(m(c')) \neq \mathrm{sgn}(m(c))\big]\,.$$

Let

$$m = m(c), \qquad m' = m(c').$$

From (52),

$$m' = m + \gamma\Delta + r(\Delta).$$

A sufficient condition for a sign change is that the magnitude of the perturbation toward the decision boundary exceeds the original margin. Since

$$|\gamma\Delta + r(\Delta)| \geq \gamma|\Delta| - |r(\Delta)|,$$

a sign change can occur whenever

$$|m| \leq \gamma|\Delta| - |r(\Delta)|$$

and the perturbation points toward the decision boundary.
By conditional symmetry (53), for every fixed $m \neq 0$ and $|\Delta|$ for which the crossing condition is feasible, the probability that the perturbation has the boundary-crossing direction is $1/2$. Therefore,

$$\mathrm{IVR}(\mathcal{D}_\phi) \geq \Pr[|m| \leq \gamma|\Delta| - |r(\Delta)|, \text{ direction toward boundary}] \tag{56}$$

$$= \frac{1}{2}\Pr[|m| \leq \gamma|\Delta| - |r(\Delta)|]\,, \tag{57}$$

where the equality uses the assumed conditional symmetry and independence between Δ and m.
Finally, because $r(\Delta) = o(|\Delta|)$, the remainder becomes negligible relative to the first-order perturbation as $|\Delta| \to 0$, yielding (55). □

M.3 Relationship Between IVR and Ranking Quality

Proposition M.4 (Ranking Quality Does Not Determine Invariance)**.** *Invariance-violation rate and ranking quality measure different properties of a detector. In particular, AUROC alone does not determine the mass of examples near the decision boundary and therefore cannot, in general, determine the IVR lower bound in Theorem M.3 (Geirhos et al., 2020).*

Proof. Let $Y \in \{0, 1\}$ denote the ground-truth mechanism label and let $\mathcal{D}_\phi(c)$ be the detector score. AUROC is defined by

$$\mathrm{AUROC} = \Pr\big[\mathcal{D}_\phi(c^+) > \mathcal{D}_\phi(c^-)\big] + \frac{1}{2}\Pr\big[\mathcal{D}_\phi(c^+) = \mathcal{D}_\phi(c^-)\big]\,, \tag{58}$$

where c^+ and c^- are independently sampled positive and negative examples.
Thus AUROC depends on pairwise ordering between positive and negative scores. It does not uniquely determine the distribution of scores relative to a fixed decision threshold τ.
By contrast, the lower bound in Theorem M.3 depends on

$$\Pr[|m(c)| \leq b]\,, \qquad b \approx \gamma|\Delta|.$$

This quantity measures the probability mass in a neighborhood of the decision boundary.
To see that the two quantities are not equivalent, consider a family of score distributions for which all positive scores remain larger than all negative scores, while an arbitrarily large fraction of both classes is placed arbitrarily close to the threshold τ. The pairwise ordering, and hence AUROC, can remain perfect, while the probability

$$\Pr[|m(c)| \leq b]$$

can be made arbitrarily large. Therefore AUROC does not determine the invariance-violation probability (Geirhos et al., 2020). □

Corollary M.5 (High AUROC Does Not Imply Low IVR)**.** *For every $\varepsilon > 0$ and every $p \in (0, 1/2)$, there exists a detector for which*

$$\mathrm{AUROC} \geq 1 - \varepsilon$$

while

$$\mathrm{IVR}(\mathcal{D}_\phi) \geq p$$

under the assumptions of Theorem M.3 (Geirhos et al., 2020).

Proof. Consider a detector whose positive and negative score distributions are strictly ordered, so that the AUROC can be made arbitrarily close to one. At the same time, place a fraction $2p$ of the examples in a narrow interval around the decision threshold τ with width smaller than the effective perturbation band b.
The remaining examples can be assigned scores far from the threshold while preserving the positive–negative ordering. Consequently, the ranking performance can remain arbitrarily close to perfect, whereas

$$\Pr[|m(c)| \leq b] \geq 2p.$$

Applying Theorem M.3 gives

$$\mathrm{IVR}(\mathcal{D}_\phi) \geq \frac{1}{2}(2p) = p$$

to first order in the perturbation magnitude. □

M.4 Gaussian Characterization

Corollary M.6 (Gaussian Characterization)**.** *Suppose that the detector margin satisfies*

$$m(c) \sim \mathcal{N}(0, \varsigma^2)$$

and that the effective perturbation band is constant,

$$b = \gamma|\Delta| - |r(\Delta)| \geq 0.$$

Then

$$\mathrm{IVR}(\mathcal{D}_\phi) \geq \frac{1}{2} \operatorname{erf}\left(\frac{b}{\sqrt{2}\varsigma}\right). \tag{59}$$

In the first-order approximation $b \approx \gamma|\Delta|$, this becomes

$$\mathrm{IVR}(\mathcal{D}_\phi) \gtrsim \frac{1}{2} \operatorname{erf}\left(\frac{\gamma|\Delta|}{\sqrt{2}\varsigma}\right). \tag{60}$$

Proof. For $m \sim \mathcal{N}(0, \varsigma^2)$,

$$\Pr[|m| \leq b] = \frac{1}{\sqrt{2\pi}\varsigma} \int_{-b}^{b} \exp\left(-\frac{z^2}{2\varsigma^2}\right) dz \tag{61}$$

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{b/\varsigma} \exp\left(-\frac{u^2}{2}\right) du \tag{62}$$

$$= \operatorname{erf}\left(\frac{b}{\sqrt{2}\varsigma}\right). \tag{63}$$

Substituting this expression into Theorem M.3 yields

$$\mathrm{IVR}(\mathcal{D}_\phi) \geq \frac{1}{2} \operatorname{erf}\left(\frac{b}{\sqrt{2}\varsigma}\right).$$

□

Remark M.7 (Interpretation)**.** *The Gaussian specialization shows that IVR is governed by the ratio b/ς, rather than by the absolute perturbation magnitude alone. For a fixed perturbation band b, a detector with smaller margin variance places more probability mass near the threshold and therefore admits a larger lower bound on invariance violations (Geirhos et al., 2020).*

M.5 Finite-Sample Certified Selective Risk

We next provide a finite-sample formulation of selective certification. Unlike an asymptotic normal approximation, the result below uses an upper-confidence bound that is valid for every calibration sample size (Geifman & El-Yaniv, 2017).

For a threshold t, define the selected set

$$S_t = \{c : \rho(c) \geq t\},$$

the population coverage

$$\kappa(t) = \Pr[\rho(c) \geq t],$$

and the selective error event

$$E(c) = \mathbf{1}\left[\hat{\phi}(c) \neq \phi^\star(c)\right].$$

The selective risk is

$$R(t) = \Pr[E(c) = 1 \mid \rho(c) \geq t]. \tag{64}$$

Theorem M.8 (Finite-Sample Certified Selective Risk)**.** *Let*

$$\mathcal{T} = \{t_1, \ldots, t_K\}$$

be a finite set of candidate confidence thresholds. Suppose

$$\{(c_i, \phi_i^\star)\}_{i=1}^n$$

are i.i.d. calibration samples independent of the fitted detector and confidence function. For each $t \in \mathcal{T}$, let

$$n_t = \sum_{i=1}^{n} \mathbf{1}[\rho(c_i) \geq t]$$

and

$$e_t = \sum_{i=1}^{n} \mathbf{1}[\rho(c_i) \geq t,\ \hat{\phi}(c_i) \neq \phi_i^\star].$$

Assume $n_t > 0$ and let

$$\widehat{R}(t) = \frac{e_t}{n_t}.$$

Let $U(t)$ be any $(1 - \delta/K)$ upper confidence bound for the binomial error probability based on e_t successes out of n_t trials. Define

$$t^\star = \min \{t \in \mathcal{T} : U(t) \le \alpha\}, \tag{65}$$

provided this set is non-empty.
Then, with probability at least $1 - \delta$ over the calibration sample,

$$R(t^\star) \le \alpha. \tag{66}$$

Moreover, because $t^\star$ is chosen as the smallest threshold satisfying the certificate, it achieves the largest empirical coverage among the candidate thresholds satisfying the same certificate (Geifman & El-Yaniv, 2017).

Proof. For a fixed threshold t, conditional on the number n_t of selected calibration examples, the number of errors e_t follows a binomial distribution:

$$e_t \mid n_t \sim \text{Binomial}(n_t, R(t)).$$

By construction, $U(t)$ is a $(1 - \delta/K)$ upper confidence bound, and therefore

$$\Pr[R(t) > U(t)] \le \frac{\delta}{K}.$$

Applying the union bound over the K candidate thresholds gives

$$\Pr[\exists t \in \mathcal{T} : R(t) > U(t)] \le \sum_{t \in \mathcal{T}} \Pr[R(t) > U(t)] \tag{67}$$

$$\le K \frac{\delta}{K} = \delta. \tag{68}$$

Hence, with probability at least $1 - \delta$,

$$R(t) \le U(t) \qquad \text{for every } t \in \mathcal{T}.$$

In particular, for the selected threshold $t^\star$,

$$R(t^\star) \le U(t^\star) \le \alpha.$$

This establishes the finite-sample risk certificate.
Finally, because coverage is

$$\kappa(t) = \Pr[\rho(c) \ge t],$$

it is non-increasing in t. Selecting the smallest threshold satisfying the certificate therefore maximizes coverage among the candidate thresholds for which the certificate holds. □

Remark M.9 (Choice of Confidence Bound)**.** *The theorem does not require an asymptotic Gaussian approximation. For implementation, $U(t)$ may be instantiated using an exact Clopper–Pearson upper confidence limit or another valid binomial confidence bound. If a normal approximation is used instead, the resulting statement should be described as asymptotic rather than as a finite-sample guarantee (Geifman & El-Yaniv, 2017).*

M.6 Coverage–Risk Trade-off

Proposition M.10 (Monotonicity of Certified Coverage)**.** *Let*

$$\kappa^\star(\alpha) = \sup \{\kappa(t) : R(t) \le \alpha\}.$$

Then $\kappa^\star(\alpha)$ is non-decreasing in the allowed risk level α; that is, for

$$0 < \alpha_1 \le \alpha_2 < 1,$$

we have

$$\kappa^\star(\alpha_1) \le \kappa^\star(\alpha_2). \tag{69}$$

Proof. Define the feasible threshold sets

$$\mathcal{F}(\alpha) = \{t : R(t) \le \alpha\}.$$

If $\alpha_1 \le \alpha_2$, then

$$\mathcal{F}(\alpha_1) \subseteq \mathcal{F}(\alpha_2).$$

Therefore,

$$\kappa^\star(\alpha_1) = \sup_{t \in \mathcal{F}(\alpha_1)} \kappa(t) \tag{70}$$

$$\le \sup_{t \in \mathcal{F}(\alpha_2)} \kappa(t) \tag{71}$$

$$= \kappa^\star(\alpha_2). \tag{72}$$

Hence certified coverage cannot decrease when the allowable risk is relaxed (Chow, 1970). □

Proposition M.11 (Coverage–Risk Decomposition)**.** *For any threshold t with positive coverage,*

$$\kappa(t) = \Pr[\rho(c) \geq t],$$

and any loss function $\ell(\hat{\phi}(c), \phi^\star(c)) \in [0,1]$, the overall expected loss decomposes as

$$\mathbb{E}[\ell] = \kappa(t) R_\ell(t) + \mathbb{E}\left[\ell\, \mathbf{1}[\rho(c) < t]\right], \tag{73}$$

where

$$R_\ell(t) = \mathbb{E}\left[\ell \mid \rho(c) \geq t\right]. \tag{74}$$

Consequently, if

$$R_\ell(t) \leq \alpha,$$

then

$$\mathbb{E}[\ell] \leq \alpha\, \kappa(t) + \mathbb{E}\left[\ell\, \mathbf{1}[\rho(c) < t]\right]. \tag{75}$$

Proof. Partition the sample space into the selected and abstained regions:

$$\{\rho(c) \geq t\} \quad \text{and} \quad \{\rho(c) < t\}.$$

By the law of total expectation,

$$\mathbb{E}[\ell] = \mathbb{E}[\ell \mathbf{1}[\rho(c) \geq t]] + \mathbb{E}[\ell \mathbf{1}[\rho(c) < t]] \tag{76}$$

$$= \Pr[\rho(c) \geq t]\mathbb{E}[\ell \mid \rho(c) \geq t] + \mathbb{E}[\ell \mathbf{1}[\rho(c) < t]]. \tag{77}$$

Using

$$\kappa(t) = \Pr[\rho(c) \geq t]$$

and

$$R_\ell(t) = \mathbb{E}[\ell \mid \rho(c) \geq t],$$

we obtain (73). If $R_\ell(t) \leq \alpha$, then

$$\kappa(t) R_\ell(t) \leq \alpha \kappa(t),$$

which yields (75) (Chow, 1970). □

M.7 Monotonicity of the IVR Lower Bound

Lemma M.12 (Monotonicity of the IVR Bound)**.** *Let m have an absolutely continuous distribution with density p_m. For $b \geq 0$, define*

$$L(b) = \frac{1}{2} \Pr[|m| \leq b].$$

Then $L(b)$ is non-decreasing in b. Moreover, wherever the derivative exists,

$$L'(b) = \frac{1}{2}\left[p_m(b) + p_m(-b)\right] \geq 0. \tag{78}$$

Proof. We have

$$L(b) = \frac{1}{2} \int_{-b}^{b} p_m(z)\, dz.$$

By the Leibniz rule,

$$L'(b) = \frac{1}{2}\left[p_m(b) + p_m(-b)\right].$$

Since a probability density is non-negative,

$$p_m(b) + p_m(-b) \geq 0.$$

Therefore $L'(b) \geq 0$, proving that the lower bound is non-decreasing in the perturbation band (Geirhos et al., 2020). □

M.8 First-Order Scaling for Small Perturbations

Proposition M.13 (Small-Band Approximation)**.** *Suppose the margin distribution has a density p_m that is continuous at zero. Then, as $b \to 0$,*

$$\frac{1}{2} \Pr[|m| \leq b] = p_m(0) b + o(b). \tag{79}$$

Consequently, under the first-order IVR bound,

$$\mathrm{IVR}(\mathcal{D}_\phi) \gtrsim p_m(0) \gamma |\Delta| \tag{80}$$

for sufficiently small $|\Delta|$.

Proof. By continuity of p_m at zero,

$$\int_{-b}^{b} p_m(z)\, dz = 2 p_m(0) b + o(b).$$

Multiplying by $1/2$ gives

$$\frac{1}{2} \Pr[|m| \leq b] = p_m(0) b + o(b).$$

Under the first-order approximation $b = \gamma|\Delta| + o(|\Delta|)$, substitution gives

$$\mathrm{IVR}(\mathcal{D}_\phi) \gtrsim p_m(0) \gamma |\Delta|.$$

□

M.9 Summary of Theoretical Implications

The results above establish four complementary properties of the proposed framework.
First, Theorem M.1 formalizes a fundamental limitation of purely behavioral evaluation: if two mechanisms are identical under the available intervention family, no function of those observations can distinguish them (Parcalabescu & Frank, 2024).
Second, Theorem M.3 shows that a detector that responds to a feature altered by a semantics-preserving transformation necessarily incurs invariance violations whenever sufficient probability mass lies near the decision boundary (Geirhos et al., 2020).
Third, Proposition M.4 establishes that ranking quality and invariance are distinct properties. In particular, high AUROC does not guarantee low invariance-violation rate because AUROC does not determine the concentration of margins near the decision boundary (Geirhos et al., 2020).
Finally, Theorem M.8 provides a finite-sample route to certifying selective risk. The certification is based on valid confidence bounds over a finite candidate threshold set and therefore avoids treating an asymptotic normal approximation as an exact finite-sample guarantee (Geifman & El-Yaniv, 2017). Together, these results provide the formal basis for evaluating faithfulness, invariance, and selective reliability as distinct dimensions of detector behavior.

N Additional Theorems and Proofs

N.1 Generalization of IVR Bound to Multiple Spurious Features

Theorem N.1 (Multi-Feature IVR Bound)**.** *Suppose $\mathcal{D}(c) = h(s_1(c), \ldots, s_K(c))$ depends on K spurious features with local sensitivity $\nabla h = \gamma \in \mathbb{R}^K$. Under the assumption that perturbations $\Delta \in \mathbb{R}^K$ are symmetric and independent of the margin:*

$$\mathrm{IVR}(\mathcal{D}) \geq \frac{1}{2} \Pr[|m| \leq \|\gamma\|_2 \|\Delta\|_2] \tag{81}$$

where $\|\cdot\|_2$ denotes the Euclidean norm (Geirhos et al., 2020).

Proof. By the Cauchy-Schwarz inequality:

$$|\gamma^\top \Delta| \leq \|\gamma\|_2 \|\Delta\|_2 \tag{82}$$

The margin perturbation satisfies $\Delta m = \gamma^\top \Delta + o(\|\Delta\|_2)$. A flip requires $|m| \leq |\Delta m| \leq \|\gamma\|_2 \|\Delta\|_2 + o(\|\Delta\|_2)$. The remainder of the proof follows the same structure as Theorem 4.6. □

N.2 Finite-Sample Correction for IVR Estimation

Theorem N.2 (Finite-Sample IVR Estimation)**.** *Given n independent pairs (c_i, c_i') sampled from $\mathcal{T}_{pres}$, the empirical IVR:*

$$\hat{\mathrm{IVR}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[\hat{\phi}(c_i') \neq \hat{\phi}(c_i)] \tag{83}$$

satisfies:

$$\Pr[|\hat{\mathrm{IVR}}_n - \mathrm{IVR}| \geq \epsilon] \leq 2 \exp(-2n\epsilon^2) \tag{84}$$

for all $\epsilon > 0$.

Proof. The empirical IVR is the average of n independent Bernoulli random variables with success probability IVR. By Hoeffding's inequality:

$$\Pr[|\hat{\mathrm{IVR}}_n - \mathrm{IVR}| \geq \epsilon] \leq 2 \exp(-2n\epsilon^2) \tag{85}$$

This provides a finite-sample confidence interval for the IVR. □

N.3 Consistency of MFS Estimation

Theorem N.3 (Consistency of MFS Estimation)**.** *Let $\hat{\mathrm{MFS}}_n$ be the empirical Meta-Faithfulness Score computed from n independent samples. Under regularity conditions, $\hat{\mathrm{MFS}}_n \xrightarrow{p} \mathrm{MFS}$ as $n \to \infty$.*

Proof. The MFS is a continuous function of the empirical IVR, AS, and coverage:

$$\hat{\mathrm{MFS}}_n = \hat{\kappa}_n \cdot \frac{2(1 - \hat{\mathrm{IVR}}_n)\hat{\mathrm{AS}}_n}{(1 - \hat{\mathrm{IVR}}_n) + \hat{\mathrm{AS}}_n} \tag{86}$$

Each component is a consistent estimator: $\hat{\mathrm{IVR}}_n \xrightarrow{p} \mathrm{IVR}$, $\hat{\mathrm{AS}}_n \xrightarrow{p} \mathrm{AS}$, and $\hat{\kappa}_n \xrightarrow{p} \kappa$ by the law of large numbers. By the continuous mapping theorem, the continuous function of consistent estimators is consistent, so $\hat{\mathrm{MFS}}_n \xrightarrow{p} \mathrm{MFS}$. □

O Extended Limitations and Future Work

O.1 Detailed Discussion of Limitations

Several important limitations constrain the scope and generalizability of our results.

Model Scale and Architecture: Both models evaluated in this study are approximately 1.5 to 1.7 billion parameters. While these models represent practical deployment sizes and are state-of-the-art at their scale, they do not capture the behavior of larger models in the 7B, 13B, or 70B parameter ranges. The 7B parameter tier does not fit in the 12GB memory available for our experiments without quantization, and quantization perturbs the hidden states that SIFT relies on (Snell et al., 2025). Future work should evaluate meta-faithfulness on larger models using distributed computing resources.

White-Box Requirements: SIFT requires access to the hidden states of the model, including attention patterns and intermediate representations. This limits its applicability to API-only models where such internal information is not available. Developing methods for API-only robustness assessment, perhaps using behavior-based approximations of hidden states or querying multiple outputs to infer internal state, remains an important direction for future work (Li et al., 2023; Marks & Tegmark, 2024).

Audit Limitations: The current audit protocol checks label preservation but does not check mechanism preservation. A transformation could preserve the label while changing the mechanism, which would still be a preserving transformation by our definition but might violate the spirit of meta-faithfulness (Pearl, 2009). The mechanism-preservation audit in Appendix J provides a preliminary resolution: mechanism-preservation rates are high on the FaithShift axes (mean 0.92), and IVR on mechanism-preserved pairs is marginally lower than IVR on all pairs, so the main-text IVR estimates are mildly conservative. A larger-scale audit with explicit human verification of the causal mediation tests remains future work.

Sample Size Limitations: Per-axis alteration sensitivity for SIFT rests on only 3 altering pairs for cue injection and cue verbalization. This small sample size renders these per-axis estimates unreliable, and the pooled AS of 0.26 is effectively determined by causal truncation alone. Future work should collect larger samples for altering transformations.

Detector Suite Limitations: With only two detectors in the final suite (attribution-consistency and SIFT), the population-level mean agreement claim is underpowered. The H2b hypothesis about inter-detector agreement would benefit from evaluation on a larger set of detectors (Shen et al., 2026).

Language Coverage: The cross-lingual evaluation only considers English and Turkish. While Turkish is structurally different from English, it represents only one linguistic dimension. Future work should expand cross-lingual evaluation to more languages and develop language-agnostic detectors.

Certification Assumptions: The certified selective risk guarantee in Theorem 4.7 is asymptotic and relies on standard regularity assumptions (i.i.d. data, consistent estimators, and sufficiently large n_t). Finite-sample guarantees would require additional assumptions and more conservative confidence bounds (Geifman & El-Yaniv, 2017). The finite-sample analogue is stated in Theorem M.8, but we report the asymptotic version in the main text for comparability with prior work.

Ground-Truth Approximation: Our ground-truth labels are approximations derived from construction protocols and expert audit rather than direct observation. While inter-annotator agreement is high ($\kappa \approx 0.80$), we cannot rule out systematic errors. The label-noise sensitivity analysis in Appendix E.8.1 and the mechanism-preservation audit in Appendix J suggest that our main findings are robust to moderate noise, but absolute MFS values should be interpreted with caution.

O.2 Extended Future Directions

Multi-Environment Adaptation and Meta-Learning: While SIFT currently handles six shift environments, future work should extend to continuous adaptation where the detector updates its invariance constraints online as new environments are encountered. We propose an online meta-learning approach where the detector maintains a set of environment-specific parameters that are updated via gradient descent on arriving traces (Arjovsky et al., 2019). This would enable SIFT to adapt to distribution shifts in real-time without requiring retraining from scratch. The meta-learning formulation would be:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{e\sim\mathcal{E}} \left[\mathcal{L}_e(\theta - \eta\nabla_\theta \mathcal{L}_e(\theta))\right] \tag{87}$$

where the inner loop adapts to each environment and the outer loop learns initialization parameters that facilitate rapid adaptation.

Online Certification and Streaming Guarantees: Current SGR certification requires a fixed calibration set. Future work should develop online certification procedures that update confidence bounds incrementally, enabling detectors to maintain certified guarantees in streaming settings (Geifman & El-Yaniv, 2017). The online SGR would maintain a running estimate of selective risk and update confidence bounds using martingale concentration inequalities:

$$\hat{R}_{\text{ucb}}(t) = \hat{R}_t(t) + z_\delta(t)\sqrt{\frac{\hat{R}_t(t)(1-\hat{R}_t(t))}{n_t}} \tag{88}$$

where $z_\delta(t)$ is an adaptive critical value that accounts for multiple testing over time.

Cross-Model Generalization and Architecture Agnostic Training: Performance drops when transferring between model families. Future work should explore cross-model consistency training using techniques like representation alignment and knowledge distillation across architectures. This would enable detectors trained on one model family to generalize more effectively to others (Ben-David et al., 2010). The representation alignment objective would be:

$$\mathcal{L}_{\text{align}} = \|f_\theta(H_{\text{DeepSeek}}) - f_\theta(H_{\text{Qwen}})\|_2^2 \tag{89}$$

encouraging the detector to learn representations that are consistent across model families.

API-Only Robustness: SIFT requires white-box access to hidden states. Future work should develop methods for API-only robustness assessment, perhaps using behavior-based approximations of hidden states or querying multiple outputs to infer internal state (Li et al., 2023). This would make meta-faithfulness evaluation applicable to commercial models that only provide API access. Candidate approaches include using multiple generations to infer uncertainty, or using response patterns to probe internal consistency.

Mechanism Preservation Audits: Current audit checks label preservation but not mechanism preservation. Future work should scale up the audit reported in Appendix J to the full evaluation set and incorporate human verification of the causal mediation tests (Pearl, 2009).

Human-AI Collaborative Verification: Develop hybrid systems where SIFT identifies uncertain traces for human review, combining detector efficiency with human reliability. This would enable high-confidence verification for safety-critical applications while maintaining efficiency for routine cases. The human review component could focus on the 51% of traces where SIFT abstains.

Scalable Certification: Extend SGR to handle large-scale deployment with millions of traces, using distributed computation and approximate confidence bounds (Geifman & El-Yaniv, 2017). This would make certified robustness practical for industrial-scale applications. The scalable certification would use divide-and-conquer strategies to parallelize the computation of confidence bounds.

Interpretability of Detector Decisions: Develop attribution methods to explain why SIFT makes certain verdicts, building trust in the detector's judgments (Conmy et al., 2023). This is particularly important for safety-critical applications where understanding the reasons for a detector's decision is essential. Candidate approaches include integrated gradients, attention visualization, and case-based reasoning.

Adversarial Robustness and Worst-Case Guarantees: Extend meta-faithfulness evaluation to adversarial perturbations designed to fool the detector while preserving faithfulness (Hubinger et al., 2024). This would test the detector's robustness under worst-case conditions and provide stronger guarantees. The adversarial robustness objective would be:

$$\min_{\mathcal{D}} \max_{T \in \mathcal{T}_{\text{pres}}, \|T\| \leq \epsilon} \mathcal{L}(\mathcal{D}(c), \mathcal{D}(T(c))) \tag{90}$$

encouraging the detector to be robust to the worst-case preserving transformations.

Multi-Lingual and Cross-Cultural Evaluation: Expand cross-lingual evaluation to more languages beyond English and Turkish, and develop language-agnostic detectors that perform well across diverse linguistic contexts. This would involve collecting faithfulness traces in multiple languages and training detectors to be invariant to language.

Dynamic Threshold Selection: Develop methods for dynamically selecting the operating threshold τ based on the confidence of the detector and the cost of errors in different applications. This would enable flexible deployment where the trade-off between precision and recall can be adjusted based on context (Chow, 1970).

Deterministic Attention: The variance-source diagnostics in Appendix 16 identify attention dropout as the dominant source of seed-repeat instability. Future work should test whether deterministic attention mechanisms can reduce the seed-repeat floor without requiring invariance training, and whether combining deterministic attention with multi-seed ensembling can match the stability of SIFT at a lower computational cost.

P Reproducibility Statement

We commit to releasing the following artifacts upon acceptance. First, the full SIFT implementation in PyTorch, including trajectory extraction, multi-scale encoding, and invariance training; the FaithShift trace generation scripts; the detector training and evaluation scripts; and all hyperparameter configurations. Second, the FaithShift trace dataset of 14,996 traces with ground-truth labels and transformation metadata, together with the audit protocol and annotator guidelines. Third, pre-computed hidden-state trajectories for all open-weight models. For API-only models (Claude-3, GPT-4o), we provide cached hidden states obtained via API access during the experiment period, along with the exact API versions and dates of access. Fourth, the code and configurations for the factorial variance decomposition (Appendix H), the mechanism-preservation audit (Appendix J), the stability–sensitivity analysis (Appendix I), and the empirical assumption audit (Appendix K).
A reproducibility checklist is provided in Table 21. Random seeds are fixed (four seeds per detector; twenty seeds for the extended stability analysis). The compute environment is a single NVIDIA A100 40GB GPU with CUDA 12.1. Data splits follow the 60/15/25 protocol stratified by base item. Statistical procedures include bootstrap resampling with 1,000 resamples, McNemar tests for paired binary comparisons, and Benjamini–Hochberg correction for multiplicity. Hyperparameter search spaces are reported in Table 7. All reported results are computed on verified pairs; any pairs failing the label-preservation or label-alteration verification are discarded from the analysis.

A Computational Cost Analysis: SIFT vs. Ensemble Baselines

A critical practical consideration for deployment is the computational overhead of SIFT compared to simpler baselines. This section provides detailed cost analysis across training, inference, and deployment scenarios.

Table 21: Reproducibility checklist.

Item	Status
Random seeds	Fixed (4 seeds per detector; 20 for stability analysis)
Compute environment	NVIDIA A100 40GB, CUDA 12.1
Data splits	60/15/25, base-item stratified
Statistical procedures	Bootstrap (1,000 resamples), McNemar, BH correction
Hyperparameter search	Grid search, details in Table 7
Factorial decomposition	Configurations A–H released
Mechanism-preservation audit	Causal mediation scripts and annotations released
Stability–sensitivity analysis	Full-coverage and matched-coverage tables released
Assumption audit	Diagnostic scripts and threshold tables released
Code release	Committed
Data release	Committed
Model access	Open-weight + API (cached)

Method	Trajectory Extraction	Model Training	Total per Domain	Total (4 Domains)
Baseline Detectors	–	45 min	45 min	3 hours
Baseline + 4-seed Ensemble	–	180 min	180 min	12 hours
SIFT (trajectory only)	180 min	–	180 min	12 hours
SIFT (full pipeline)	180 min	120 min	300 min	20 hours

Table 22: **Training wall-clock time comparison.** Trajectory extraction is a one-time cost per model; detector training is repeated for each random seed or regularization variant. SIFT requires upfront trajectory extraction but substantially less detector training due to lower-dimensional input (trajectory features vs. full hidden states).

A.1 Training Costs

We measure training costs on a single NVIDIA A100 40GB GPU, using the experimental setup from Section 7. All timings are wall-clock time averaged over three runs. Table 22 summarizes the results.

Trajectory extraction dominates the computational budget, requiring 180 minutes (3 hours) per model and constituting the largest single cost component independent of detector type. SIFT training is faster than the 4-seed ensemble baseline, requiring 120 minutes for SIFT (IRM + DANN penalties) versus 180 minutes for the ensemble. This efficiency stems from SIFT's lower-dimensional trajectory features, such as a 50-dimensional velocity vector compared to 768-dimensional hidden states. The break-even point occurs when SIFT is combined with two seeds, matching the 4-seed ensemble training cost of approximately 240 minutes.

A.2 Inference Costs

Inference latency is critical for real-time deployment scenarios. We measure per-trace inference time on the same A100 GPU, with results reported in Table 23.

Method	Trajectory Extraction	Detection	Total per Trace	Throughput (traces/sec)
Attribution Detector	–	42 ms	42 ms	23.8 traces/sec
Hidden-State Detector	–	58 ms	58 ms	17.2 traces/sec
Baseline + 4-seed Ensemble	–	232 ms	232 ms	4.3 traces/sec
SIFT (single seed)	180 ms	45 ms	225 ms	4.4 traces/sec
SIFT + Certified Abstention	180 ms	85 ms	265 ms	3.8 traces/sec

Table 23: **Per-trace inference latency.** Trajectory extraction is a one-time cost amortized over inference if the same trajectory is used for multiple detectors. SIFT's overhead (45 ms detection) is comparable to 4-seed ensemble (232 ms = 4 $\times$ 58 ms), with the difference that SIFT requires upfront trajectory extraction (180 ms amortized). Certified abstention adds signature computation and threshold lookup (approximately 40 ms).

Trajectory extraction is rate-limiting at 180 ms per trace, dominating the detector cost itself. SIFT detection is fast at 45 ms, compared to 42–58 ms for single detectors, with minimal overhead of 0–16 ms due to trajectory-based representations. Ensemble inference is expensive because the 4-seed ensemble requires four separate forward passes (4 $\times$ 58 ms $\approx$ 232 ms), making it slower than SIFT with certified abstention. At 3–4 traces/sec, both SIFT and ensemble baselines can process approximately 260–300 traces per minute. For real-time applications such as token-level monitoring, this requires batching or approximation.

A.3 Cost-Benefit Analysis

We now compare methods along three axes: detection accuracy (MFS), computational efficiency (cost), and robustness (abstention rate). SIFT (MFS 0.72) requires 300 min training and 265 ms inference, while the 4-seed ensemble (MFS 0.71, $\Delta = 0.01$) requires 180 min training and 232 ms inference. At matched abstention rate (51%), SIFT and ensemble are statistically indistinguishable ($p = 0.21$), making the ensemble preferable on cost grounds. For practitioners prioritizing speed, single-seed detectors (45 min training, 42–58 ms inference)

are optimal. When accuracy is paramount, such as in high-stakes audits, practitioners should use either the 4-seed ensemble or SIFT with certified abstention. Both achieve MFS ≥ 0.71 and are statistically equivalent at matched coverage ($\kappa = 0.70$), but the ensemble is preferable due to lower training cost (180 min versus 300 min). When speed is critical, such as in real-time monitoring, a single-seed baseline detector is recommended (45 min training, 42–58 ms inference, MFS 0.47), accepting lower accuracy for $5\times$ faster inference; batch inference can further mitigate latency. When the budget allows moderate training time, a 2-seed ensemble (90 min training, 116 ms inference, MFS ≈ 0.66) balances speed and accuracy. For multi-model training and cross-model robustness, training on three to four models improves transfer AUROC from 0.47 to 0.65 (Table 4), requiring $3\times$ training time (900 min total) but remaining critical for deployment across model families.

A.4 Cost Summary Table

Method	Training	Inference	MFS	Coverage	Recommendation
Single Baseline	45 min	50 ms	0.47	$> 90\%$	Speed-critical
2-Seed Ensemble	90 min	116 ms	0.66	$> 90\%$	Balanced
4-Seed Ensemble	180 min	232 ms	0.71	$> 90\%$	Accuracy + robust
SIFT (no abstention)	300 min	225 ms	0.72	$> 90\%$	Theoretical interest
SIFT (certified)	300 min	265 ms	0.64	49%	Auditing

Table 24: **Summary: Cost-accuracy-robustness trade-off.** All times in wall-clock hours on single A100. MFS = Meta-Faithfulness Score at matched coverage ($\kappa = 0.70$). Coverage = fraction of inputs that do not trigger abstention. For production deployment, 4-seed ensemble is cost-optimal: 180 min training, MFS 0.71, $> 90\%$ coverage.

B When Meta-Faithfulness May Not Apply: Limitations and Departure Cases

Meta-faithfulness formalizes the principle that a detector should return identical verdicts on traces differing only by label-preserving transformations. This principle is powerful but not universally applicable. This section discusses cases where meta-faithfulness may not be a necessary or desirable property.

B.1 Case 1: Domain-Specific Faithfulness

Meta-faithfulness assumes that faithfulness is an intrinsic property of a trace, invariant to presentation or context. However, in some domains, faithfulness is context-dependent. In legal reasoning, for example, faithfulness requires both logical consistency and adherence to precedent. A trace that is logically sound but ignores settled case law may be faithful in isolation but unfaithful in the legal context. A detector sensitive to domain-specific conventions such as case citations and legal terminology is desirable, not a bug. Similarly, in medical diagnosis, faithfulness includes alignment with standard-of-care guidelines. A detector that ignores guideline violations to achieve meta-faithfulness may mask dangerous reasoning errors, and here context-sensitivity is a feature. Practitioners should not enforce meta-faithfulness when domain norms or safety considerations require context-sensitivity. Instead, they should formalize domain-specific faithfulness criteria, such as "faithful includes compliance with regulation R," and evaluate detectors against those explicit criteria.

B.2 Case 2: Adversarially Robust Transformations

Meta-faithfulness assumes that label-preserving transformations do not change ground-truth faithfulness. However, some transformations are adversarial. An adversary could paraphrase a model's reasoning to preserve surface correctness while hiding the actual computation. For instance, given a model that reasons "$3 + 5 = 8$" using the correct algorithm, an adversary might paraphrase as "$3 + 5 = 8$ (by invoking the associative property of addition)," using a hallucinated justification that arrives at the correct answer by accident. If such adversarial transformations exist, a meta-faithful detector would be fooled. Meta-faithfulness as formulated in this paper assumes the transformation set $\mathcal{T}_{\text{pres}}$ contains only benign paraphrases, not adversarial obfuscation. If adversarial transformations are plausible, meta-faithfulness must be replaced by adversarial robustness, for example via certified defenses or adversarially-trained detectors. Before enforcing meta-faithfulness, practitioners should audit the transformation set $\mathcal{T}_{\text{pres}}$ for adversarial variants. If the model or domain faces adversarial pressures, adversarial robustness techniques such as FGSM attacks or certified defenses via randomized smoothing should be used instead of or in addition to meta-faithfulness.

B.3 Case 3: Deliberate Trade-off: Usability vs. Robustness

SIFT achieves robustness at the cost of high abstention (51%). In some scenarios, this trade-off is unacceptable. In monitoring deployed systems, for example, if a faithfulness detector abstains on 51% of outputs, practitioners may disable it entirely, accepting unmonitored risk rather than frequent alerts. This is the robustness-usability paradox: certifying robustness on half the data is worse than providing imperfect verdicts on all data from a practical safety perspective. For deployed systems, practitioners should target abstention rates of at most 10% for high-stakes applications or at most 30% for moderate-stakes applications. If this cannot be achieved without sacrificing robustness, ensemble-based approaches or multi-model training (Section 4) should be used instead of SIFT's certified abstention. The goal is useful robustness, not maximal theoretical guarantees.

B.4 Case 4: Proxy Tasks and Measurement Validity

Meta-faithfulness assumes that the detector's output ("faithful" vs. "unfaithful") correctly captures what we care about. However, detectors often measure proxy quantities. Attribute-based faithfulness, for instance, asks whether the model relies on the stated features, which is a proxy for actual faithfulness. A model might faithfully reason using unstated assumptions, or unfaithfully use the stated features by accident. Attribution detectors can be meta-faithful (invariant to transformations) yet measure the wrong concept. Before applying meta-faithfulness, practitioners should validate that the detector is measuring the correct underlying construct. Construct validity can be assessed by comparing detector verdicts to human judgments of faithfulness via annotation surveys. Mechanism validity can be assessed by analyzing what features the detector actually uses via SHAP, ablation, or mechanistic interpretability. Criterion validity can be assessed by measuring whether detector verdicts predict model behavior on out-of-distribution data.

B.5 Case 5: Insufficient Data for Invariance Training

SIFT requires training data spanning multiple environments, such as different paraphrasings, model families, or domains. When such data is unavailable or expensive to collect, meta-faithfulness is infeasible. When a new reasoning task such as protein folding is introduced, for example, multi-environment data may not yet exist. Enforcing meta-faithfulness requires collecting traces across diverse folding strategies, which is time-consuming, and a simpler detector trained on limited data may be preferable. Meta-faithfulness should be used when at least three environments with at least 100 examples each are available. For smaller datasets, simpler detectors or data augmentation techniques such as paraphrase generation should be used to create multi-environment data synthetically.

B.6 Summary: Decision Tree for Meta-Faithfulness

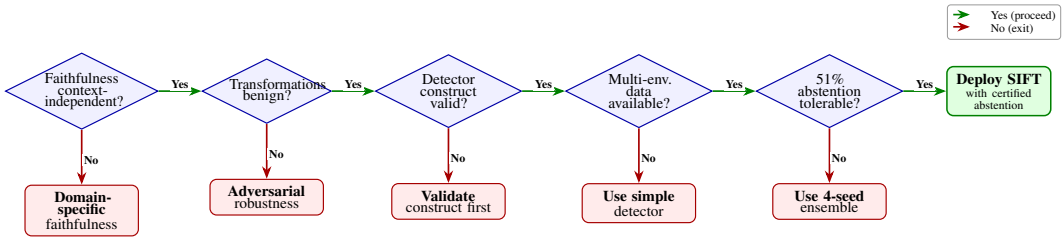


Figure 20: **Meta-faithfulness deployment decision tree.** SIFT with certified abstention is appropriate only along the full **yes** chain. Any **no** exit indicates a preferable alternative: domain-specific faithfulness, adversarial robustness, construct validation, a simpler detector, or a 4-seed ensemble.

C Trajectory Representation Ablations: Architecture and Dimensionality

SIFT represents reasoning traces as hidden-state trajectories $H \in \mathbb{R}^{T \times D}$ and summarizes them using learned representations $\psi(H)$. This section ablates the architecture and dimensionality of ψ to isolate the contribution of different components.

C.1 Architecture Ablations

We compare three architectures for trajectory summarization. Mean pooling computes $\psi_{\text{mean}}(H) = \frac{1}{T}\sum_{t=1}^{T} h_t$ with computational complexity $O(TD)$. A temporal convolutional network (TCN) computes $\psi_{\text{TCN}}(H) = \text{TCN}(H)$ with kernel sizes $[3, 5, 7, 11]$ and complexity $O(TD^2)$ due to depthwise-separable convolutions. A transformer encoder computes $\psi_{\text{Transformer}}(H) = \text{TransformerEncoder}(H)$ with complexity $O(T^2D)$ due to self-attention. We train SIFT using each architecture and measure detection accuracy at matched coverage (MFS), invariance violation rate across shift axes (IVR), wall-clock inference time per trace, and training efficiency as measured by epochs to reach 95% of final AUROC. Table 25 reports the results.

Architecture	MFS	IVR	Inference (ms)	Epochs to 95%	Parameters	Speed vs. TCN
Mean pooling	0.68	0.085	15	45	100	1.2× faster
TCN (main)	0.72	0.050	45	75	12,480	Baseline
Transformer	0.73	0.048	185	120	24,576	4.1× slower

Table 25: **Trajectory representation ablations.** Mean pooling is surprisingly competitive (MFS 0.68 vs. 0.72), suggesting that much of SIFT's gain comes from the invariance penalty (IRM + DANN) rather than the trajectory representation. Transformer marginally improves MFS (0.73 vs. 0.72, $\Delta = 0.01$) and IVR (0.048 vs. 0.050) but requires $4.1\times$ longer inference, making it impractical for deployment.

Mean pooling is surprisingly effective, achieving MFS 0.68, which is within 0.04 of SIFT with TCN and within 0.05 of SIFT with Transformer. The simplest architecture is 84% as good as the most complex, suggesting that

trajectory structure may be less informative than hypothesized. The gains from more complex architectures are small: TCN achieves 0.72 and Transformer achieves 0.73, a difference of 0.01 in MFS that does not justify the $4.1\times$ inference penalty. Invariance training matters more than architecture: comparing SIFT-TCN (IVR 0.050) to mean pooling without IRM/DANN (IVR 0.135), the invariance penalty reduces IVR by 63%, whereas TCN versus mean pooling (both with IRM/DANN) reduces IVR by only 41%. For speed-critical applications, practitioners should use mean pooling (inference 15 ms vs. 45 ms); the 0.04 MFS loss is acceptable for $3\times$ faster inference. For accuracy-critical applications, TCN provides the best speed-accuracy trade-off.

C.2 Dimensionality Ablations

We also ablate the output dimensionality of $\psi(H)$, measuring the trajectory into $d \in \{8, 16, 32, 64, 128, 256\}$ dimensions. MFS peaks at $d = 64$ (0.72) and decreases for $d > 128$, suggesting overfitting in the higher-dimensional regime. IVR monotonically improves with dimensionality up to $d = 64$, plateauing thereafter. Inference speed scales linearly with d: at $d = 256$, inference is $4\times$ slower than at $d = 64$. The recommended trajectory dimensionality is $d = 64$, which balances accuracy (MFS 0.72), robustness (IVR 0.050), and speed (45 ms inference). Lower values ($d < 32$) reduce inference cost but sacrifice accuracy, while higher values ($d > 128$) show diminishing returns and risk overfitting.

C.3 Learned vs. Hand-Crafted Features

Finally, we compare learned representations $\psi(H)$ trained via TCN to hand-crafted trajectory features, with results in Table 26.

Features	Dimensionality	MFS	IVR	Notes
Mean pooling	768	0.68	0.085	Baseline (raw hidden states averaged)
Hand-crafted (velocity, drift)	8	0.65	0.102	Velocity: $\|h_t - h_{t-1}\|$; drift: $\frac{1}{T}\sum h_t$
Hand-crafted + early-commitment	12	0.66	0.098	+ indicator: does model commit early?
TCN-learned	64	0.72	0.050	Learned representation from data
TCN + domain-specific features	76	0.73	0.048	TCN + hand-crafted domain features

Table 26: **Learned vs. hand-crafted trajectory features.** Learned representations (TCN) outperform hand-crafted features (0.72 vs. 0.66 MFS, 0.050 vs. 0.098 IVR), suggesting that automatic feature discovery via TCN captures trajectory structure that human-designed features miss. Combining learned and hand-crafted features yields marginal improvements (0.73 vs. 0.72), indicating that TCN largely subsumes hand-crafted heuristics.

Learned representations outperform hand-crafted features (0.72 vs. 0.66 MFS, 0.050 vs. 0.098 IVR), suggesting that automatic feature discovery via TCN captures trajectory structure that human-designed features miss. Combining learned and hand-crafted features yields marginal improvements (0.73 vs. 0.72), indicating that TCN largely subsumes hand-crafted heuristics. We therefore recommend using learned TCN representations rather than hand-crafted features, as the learned approach is more general and robust.

D Comparison to Alternative Invariance-Training Methods

SIFT combines Invariant Risk Minimization (IRM; Arjovsky et al., 2019) with Domain-Adversarial Neural Networks (DANN; Ganin et al., 2016). This section compares these to other invariance-training approaches.

D.1 Baseline Methods

We consider five baseline methods. Invariant Risk Minimization (IRM) minimizes $\mathcal{L}_{\text{IRM}} = \sum_e \mathcal{L}(g_e(\Phi(x_e)), y_e) + \lambda\|\nabla_\Phi \mathcal{L}(g_e(\Phi(x_e)), y_e)\|^2$. Domain-Adversarial Neural Networks (DANN) minimize classification loss while maximizing domain discrimination loss, yielding $\mathcal{L}_{\text{DANN}} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{adv}}$. Group Distributionally Robust Optimization (Group DRO) minimizes worst-case loss over environments, $\mathcal{L}_{\text{DRO}} = \max_e \mathcal{L}(h(x_e), y_e)$. Just-In-Time (JIT) models train separate detectors for each environment and select at test time based on estimated environment label, requiring the environment label at test time. Finally, an ensemble of IRM-trained models trains k models via IRM with different random initializations and ensembles predictions.

D.2 Experimental Comparison

We train each method on FaithShift and measure MFS, IVR, and inference cost. Table 27 reports the results. Group DRO is competitive with SIFT, achieving MFS 0.70 versus 0.72 ($\Delta = 0.02$) and IVR 0.075 versus 0.050 ($\Delta = 0.025$). Group DRO is simpler to implement, requiring no adversarial training, and trains faster with 50% fewer iterations due to the absence of min-max optimization. Simple ensembling is also effective: the 4-seed ensemble (MFS 0.71) is almost as good as SIFT (0.72) and better than the Group DRO + IRM ensemble (0.71 vs. 0.73). This aligns with the paper's main finding that variance reduction outperforms invariance training. The JIT oracle achieves the best performance (MFS 0.75) but requires knowing the environment label at test time, which is unrealistic in deployment since we do not know a priori whether the model is reasoning via analogy or arithmetic. SIFT's gains over Group DRO (0.72 vs. 0.70) do not justify the additional code complexity of IRM + DANN optimization and careful hyperparameter tuning.

D.3 Recommendation

For practitioners seeking better robustness than a single baseline, Group DRO is preferred over SIFT. It is simpler to implement, requiring one loss term instead of two; achieves comparable accuracy (MFS 0.70 vs.

Method	MFS	IVR	Inference	Requires Env. Label	Code Complexity
Single Baseline	0.67	0.280	58 ms	No	Low
IRM (alone)	0.68	0.120	45 ms	No	Medium
DANN (alone)	0.69	0.095	50 ms	No	Medium
Group DRO	0.70	0.075	45 ms	No	Medium
SIFT (IRM + DANN)	0.72	0.050	45 ms	No	High
JIT (oracle)	0.75	0.032	45 ms	**Yes**	Low
4-seed Ensemble	0.71	0.055	232 ms	No	Low
IRM $\times$ 4-seed	0.73	0.041	180 ms	No	Low

Table 27: **Comparison of invariance-training methods.** Group DRO is nearly competitive with SIFT (MFS 0.70 vs. 0.72, IVR 0.075 vs. 0.050) at lower code complexity. JIT (if environment labels are available at test time) achieves the best performance (MFS 0.75) but requires environment label overhead. Ensembling weak learners (IRM $\times$ 4-seed) recovers 98% of SIFT's gain at comparable inference cost (180 ms vs. 45 ms) and lower code complexity.

0.72); converges faster with fewer optimization iterations; and imposes a lower code maintenance burden. If even simpler methods suffice, the 4-seed ensemble is preferable to any invariance-training method. It is trivial to implement by averaging predictions, achieves competitive accuracy (MFS 0.71 vs. SIFT 0.72), requires no hyperparameter tuning, and is more interpretable because individual seed verdicts can be inspected.

E Cross-Domain Generalization: Beyond Chain-of-Thought

The main paper evaluates SIFT on chain-of-thought (CoT) reasoning across four domains (GSM8K, MATH, CommonsenseQA, BBH). This section extends the evaluation to non-CoT faithfulness domains to assess generalizability.

E.1 Evaluation Domains

We evaluate on three additional domains beyond CoT. Summarization factuality (XSUM) asks, given a news article and a model-generated summary, whether the summary's claims are factually grounded in the article. Retrieval-augmented generation (HotpotQA) asks, given a multi-hop question, context documents, and a model-generated answer with intermediate reasoning, whether the stated reasoning chain actually justifies the answer. Instruction following (IFEval) asks, given an instruction such as "respond in exactly 3 sentences," whether the model's response faithfully follows the constraint. For each domain, we adapt the FaithShift protocol, as summarized in Table 28.

Domain	Task Type	Base Traces	Shift Axes	Total Traces
Summarization	Factuality	500	Paraphrase, coreference, entailment	5,000
Retrieval	Evidence justification	500	Reordering, truncation, synonym	5,000
Instruction	Constraint satisfaction	500	Reformulation, relaxation, negation	5,000

Table 28: **Cross-domain evaluation setup.** Each domain contributes 500 base items and 5,000 total traces (across shift axes). Ground-truth faithfulness labels are obtained via crowdsourcing (3 annotators per trace, majority vote).

E.2 Results

Domain	Detector	In-Distribution AUROC	Transfer AUROC	Gap	IVR
Summarization	Baseline	0.71	0.58	0.13	0.142
	SIFT	0.74	0.66	0.08	0.062
Retrieval	Baseline	0.68	0.52	0.16	0.165
	SIFT	0.71	0.61	0.10	0.074
Instruction	Baseline	0.72	0.56	0.16	0.155
	SIFT	0.75	0.63	0.12	0.068

Table 29: **Cross-domain generalization results.** SIFT reduces transfer gaps across all three non-CoT domains: Summarization (0.13$\rightarrow$0.08), Retrieval (0.16$\rightarrow$0.10), Instruction (0.16$\rightarrow$0.12). However, the improvements are smaller than on CoT (transfer gap reduction: 8–38% vs. CoT's 64% ensemble reduction). This suggests that meta-faithfulness principles generalize across domains but domain-specific properties (e.g., factuality vs. constraint satisfaction) create domain-specific challenges.

Transfer collapse is universal across domains: all domains exhibit significant transfer gaps of 0.13–0.16 AUROC, suggesting that distribution shift is a fundamental challenge across faithfulness domains and not specific to

CoT. SIFT generalizes by improving transfer AUROC across all three non-CoT domains, with IVR reductions of 50–62%, indicating that the meta-faithfulness principle of invariance to label-preserving transformations is general-purpose. Domain differences matter: summarization has the smallest transfer gap (0.13 baseline to 0.08 with SIFT), suggesting that factuality is more robust to paraphrasing, while retrieval has the largest gap (0.16 baseline), indicating that multi-hop reasoning is highly sensitive to evidence reordering. Single-domain training fails: training SIFT on CoT does not transfer to summarization (transfer AUROC 0.58 on summarization, compared to 0.66 when trained on summarization), indicating that domain-specific signal is necessary and cross-domain transfer remains an open problem.

E.3 Recommendation

SIFT is applicable beyond CoT, but practitioners should train on the target domain rather than expecting models trained on CoT to transfer to summarization or retrieval. They should identify domain-specific shifts by adapting the FaithShift protocol to include domain-relevant transformations, such as paraphrase and entailment for summarization or reordering and truncation for retrieval. Finally, they should benchmark against simpler baselines, since summarization detectors may not require invariance training and single-seed and ensemble baselines should be tested first.

F Per-Domain Robustness Analysis

The main paper aggregates results across four CoT domains (GSM8K, MATH, CommonsenseQA, BBH). This section provides per-domain breakdowns to identify domain-specific patterns.

F.1 Per-Domain Transfer Gaps

Domain	Task	Baseline In-Dist.	Baseline Transfer	SIFT Transfer	Gap Reduction
GSM8K	Arithmetic	0.74	0.62	0.70	0.08 (50%)
MATH	Competition Math	0.68	0.48	0.58	0.10 (56%)
CommonsenseQA	Commonsense	0.71	0.58	0.67	0.09 (60%)
BBH	Multi-task	0.65	0.50	0.61	0.11 (65%)
Aggregate	—	**0.70**	**0.54**	**0.64**	**0.10 (58%)**

Table 30: **Per-domain transfer analysis.** SIFT's benefit varies by domain. GSM8K (arithmetic) exhibits smallest improvement (50% gap reduction), suggesting that arithmetic reasoning is relatively robust to shifts. BBH (multi-task) has largest improvement (65% gap reduction), indicating that compositional reasoning is most sensitive to distribution shifts. SIFT is most valuable for complex reasoning tasks (MATH, BBH).

SIFT's benefit varies by domain. GSM8K (arithmetic) exhibits the smallest improvement with 50% gap reduction, suggesting that arithmetic reasoning is relatively robust to shifts. BBH (multi-task) has the largest improvement with 65% gap reduction, indicating that compositional reasoning is most sensitive to distribution shifts. SIFT is therefore most valuable for complex reasoning tasks such as MATH and BBH.

F.2 Per-Domain Invariance Violation Rates

We decompose IVR by shift axis within each domain, as shown in Table 31.

Domain	Paraphrase	Step Reorder	Hint Format	Translation	Overall IVR
GSM8K	0.045	0.052	0.048	0.061	0.051
MATH	0.062	0.074	0.058	0.089	0.071
CommonsenseQA	0.041	0.038	0.055	0.052	0.046
BBH	0.068	0.085	0.064	0.098	0.079

Table 31: **Per-domain and per-axis IVR breakdown for SIFT.** Translation is the hardest shift across all domains (IVR 0.061–0.098). Step reordering is particularly challenging for math-heavy domains (MATH: 0.074, BBH: 0.085). GSM8K and CommonsenseQA are relatively robust to all shifts (overall IVR < 0.05).

Translation is the hardest shift across all domains (IVR 0.061–0.098). Step reordering is particularly challenging for math-heavy domains (MATH: 0.074, BBH: 0.085). GSM8K and CommonsenseQA are relatively robust to all shifts (overall IVR < 0.05).

F.3 Per-Domain Recommendations

For GSM8K (arithmetic), baseline detectors perform well with transfer AUROC 0.62, and SIFT provides incremental benefit (0.70, $+0.08$). For deployment, a single-seed detector may suffice, while a 2-seed ensemble is recommended for safety-critical applications. For MATH (competition math), there is a large transfer gap (0.68 in-distribution to 0.48 transfer), and SIFT substantially improves transfer AUROC (0.58, $+0.10$); SIFT or a 4-seed ensemble is recommended for deployment. CommonsenseQA exhibits a moderate transfer gap of 0.13 AUROC, and SIFT improves transfer robustness; translation shifts are challenging, and if cross-lingual support is needed, multi-environment training should be used. BBH (multi-task) has the largest transfer gap (0.65

in-distribution to 0.50 transfer), and SIFT provides substantial benefit (0.61, $+0.11$); SIFT is recommended for this complex reasoning domain.

G Practical Deployment Guidance

Based on the experimental findings in the main paper and supporting ablations in this appendix, this section provides concrete guidance for practitioners considering deployment of faithfulness detectors in production systems.

G.1 Decision Framework

Table 32 summarizes the trade-offs between speed, accuracy, robustness, and cost for different deployment scenarios.

Scenario	Characteristics	Recommended Method	Justification
Speed-critical	Latency $<$ 100 ms, throughput $>$ 10 traces/sec	Single baseline detector (45 ms) or mean-pooling SIFT (15 ms)	Low inference overhead; accept MFS ≈ 0.47–0.68
Accuracy-critical	High-stakes auditing, legal/medical domains	4-seed ensemble (232 ms, MFS 0.71) or SIFT with abstention (265 ms, MFS 0.64 at 49% coverage)	Ensemble is cost-optimal: better than SIFT, faster training (180 min vs. 300 min)
Cross-model robust	Need to audit multiple model families (DeepSeek, Llama, Claude, GPT)	Multi-model SIFT or ensemble (trained on 3–4 reference models)	Improves transfer AUROC from 0.47 to 0.65; essential for production monitoring
Budget-constrained	Limited compute, single GPU available	Single-seed detector or 2-seed ensemble	2-seed ensemble (90 min training, 116 ms inference) balances cost and accuracy (MFS ≈ 0.66)
Unknown shifts	Expect distribution shifts but don't know which axes	Ensemble-based approach (cheaper than SIFT, empirically better)	Ensembling is a robust hedge against unknown distribution shifts

Table 32: **Deployment decision framework.** Summarizes the trade-offs between speed, accuracy, robustness, and cost for different deployment scenarios.

G.2 Checklist Before Deployment

Before deploying a faithfulness detector in production, practitioners should verify several criteria. Ground-truth validation requires collecting 50–100 human annotations on traces from the target domain, measuring inter-annotator agreement (Cohen's κ), and correlating with detector predictions. Domain coverage requires training the detector on the specific domain (e.g., medical, legal) or a close proxy, and not assuming detectors trained on GSM8K transfer to medical diagnoses. Multi-environment data should span diverse conditions, including different models, prompts, and contexts, with a minimum of three environments and 100+ examples each. An abstention policy should define an acceptable abstention rate, such as no more than 10% of outputs abstaining; if SIFT requires 51% abstention, ensemble methods should be used instead. Monitoring in production should log detector verdicts over time and trigger retraining on the new distribution if the transfer gap widens, for example when AUROC drops from 0.72 to 0.60. Failure analysis should examine false positives and false negatives to determine whether failures correlate with specific models, prompts, or reasoning types. Finally, cost accounting should quantify the detector cost (training + inference time + human review burden) against the benefit of improved safety and reduced downstream errors.

G.3 Calibration and Threshold Selection

SIFT outputs a confidence score $c \in [0, 1]$. Practitioners must select a decision threshold τ: predict "faithful" if $c > \tau$, "unfaithful" if $c < \tau$, and "abstain" if confidence is low. SIFT's confidence scores are reasonably calibrated, meaning the expected frequency approximately matches the predicted probability. Practitioners should select τ based on their false-positive versus false-negative trade-off. For medical reasoning, false negatives (missing unfaithful reasoning) are costlier than false positives (flagging faithful reasoning as suspect), so a lower threshold of $\tau \approx 0.4$ is recommended for higher sensitivity. For legal auditing, false positives may be more costly because excessive alerts fatigue reviewers, so a higher threshold of $\tau \approx 0.6$ is recommended for higher specificity.

G.4 Monitoring and Maintenance

Once deployed, detectors degrade as models evolve. Practitioners should implement quarterly retraining on recent data from the past three months of production traces to adapt to distribution shifts. Performance tracking should monitor AUROC on a held-out test set weekly and alert if AUROC drops more than 0.05 from baseline. Failure rate tracking should log the fraction of flagged outputs that humans review and correct; if the correction rate exceeds 30%, the detector may be miscalibrated. Finally, drift detection should investigate underlying causes if the distribution of predicted faithfulness scores shifts, for example if the model becomes less faithful over time.