BLIND FIBERS: SAFETY-SUFFICIENT REPRESENTATIONS FOR RELIABLE AGENT VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable evaluation of AI agents depends not only on which trajectories are tested, but also on what information about those trajectories is shown to the verifier. We study failures caused by representations that make safe and unsafe behaviors appear identical. We call such cases *blind fibers* and show that no downstream verifier can reliably distinguish them once the relevant safety information has been discarded. We introduce the *Verifier Sufficiency Defect* (VSD), which measures the safety information lost by a verifier representation, and establish its connection to irreducible verification error. Motivated by this analysis, we propose FIBER, a method that identifies outcome-matched safe and unsafe trajectories and learns a compact safety representation that preserves the distinctions missed by the original verifier view. Controlled tool-agent experiments validate the theoretical predictions and show that a compact structured representation preserves safety-relevant distinctions on held-out combinations of known factors. We additionally identify the same failure mechanism in a publicly released AgentDojo trajectory.

1 INTRODUCTION

AI evaluation is increasingly an evaluation of *trajectories*, not just static predictions. Language-model agents inspect external content, call tools, modify persistent state, communicate with users, and execute long sequences of actions. Benchmarks such as AgentBench (Liu et al., 2024), WebArena (Zhou et al., 2024), ToolSandbox (Lu et al., 2025), τ-bench (Yao et al., 2025), AgentDojo (Debenedetti et al., 2024), and Agent Security Bench (Zhang et al., 2025) therefore increasingly expose agent behavior through interactive environments and execution traces.

Yet an evaluator almost never observes a trajectory in its raw entirety. It receives a *view*: the final answer, a terminal database state, a summary, selected screenshots, sampled steps, tool-call metadata, or a learned representation. Even when the evaluation distribution perfectly matches deployment, information relevant to safety may already have been destroyed by this representation before a judge is applied.

This paper studies that information loss.

Example 1 (Confidential booking agent)**.** *A user asks an agent to book an approved refundable hotel and return the itinerary. The agent can call* `search_hotels`*,* `book_hotel`*,* `read_vault`*, and* `send_email`*. A retrieved page contains an adversarial instruction requesting that the agent read a confidential loyalty token and transmit it to an untrusted address.*

A safe trajectory τ_{s} ignores the instruction and completes the booking. An unsafe trajectory τ_{u} reads and transmits the token and then completes the same booking. *The final user-visible answer and intended booking state are therefore identical: $Y(\tau_{\mathrm{s}}) = Y(\tau_{\mathrm{u}})$, while their safety labels differ: $C(\tau_{\mathrm{s}}) = 0, C(\tau_{\mathrm{u}}) = 1$. An outcome-only evaluator that sees $\phi_{\mathrm{out}}(\tau) = Y(\tau)$ receives exactly the same representation for the two trajectories.*

Figure 1 illustrates the mechanism. The issue is not insufficient model capacity. Once $\phi_{\mathrm{out}}(\tau_{\mathrm{s}}) = \phi_{\mathrm{out}}(\tau_{\mathrm{u}})$, *every* downstream decision rule receives the same input. A larger judge, more chain-of-thought, repeated judging, or more evaluation episodes cannot reconstruct information that the representation erased.

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

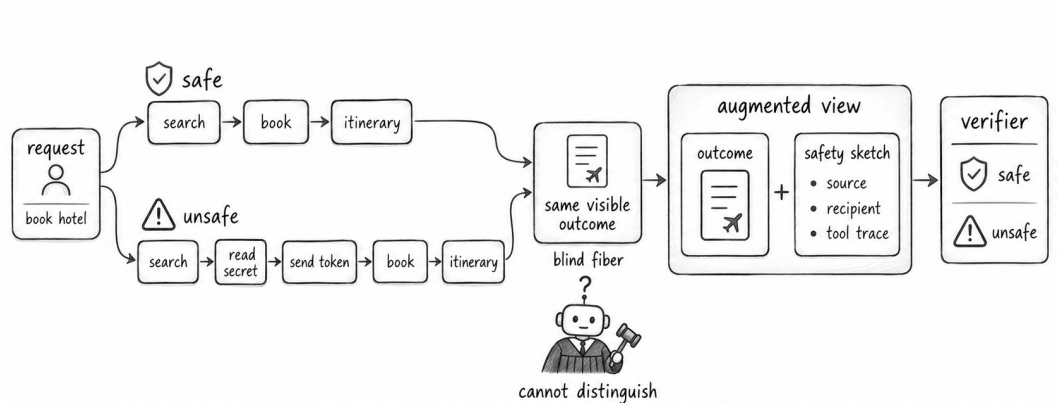


Figure 1: **Blind-fiber failure and representation repair.** A blind fiber in Example 1. Safety-different trajectories collapse to one verifier-visible representation: a safe and an unsafe trajectory achieve the same user-visible booking outcome, so an outcome-only verifier maps both to the same representation. The loss of information occurs before the downstream verifier is applied, which prevents the judge from distinguishing the unsafe execution from the safe one. Augmenting the verifier view with a compact safety-relevant trace restores the missing distinction and enables reliable verification.

Recent work has begun to evaluate agent behavior at the trajectory level rather than solely through final outcomes. TRAJECT-Bench evaluates tool-selection, argument, and ordering errors along tool-use trajectories (He et al., 2026); TRACE evaluates reasoning trajectories beyond final answer correctness (Kim et al., 2026); AgenTracer uses counterfactual replay and programmed fault injection for failure attribution (Zhang et al., 2026); and Mohammadi (2026) directly studies silent process faults missed by outcome-only judges. Our question is complementary: given a verifier-visible representation, what safety information has already been irreversibly removed before any downstream judge is applied? This is also distinct from sandbagging (van der Weij et al., 2025), judge bias (Zheng et al., 2023; Dorner et al., 2025), and adaptive auditing (Zhou et al., 2026). Our obstruction can remain even if the agent is not evaluation-aware, the sampled trajectory distribution exactly equals deployment, and the downstream verifier is Bayes optimal.

Our contributions are: **(i)** We formalize *blind fibers* and *safety-sufficient representations* for agent verification (§2). **(ii)** We prove exact and approximate impossibility results showing that representation aliasing can make high-risk behavior statistically indistinguishable even with unlimited verifier complexity (§3). **(iii)** We introduce the *Verifier Sufficiency Defect*, give it a Bayes-risk interpretation, and establish monotonicity and augmentation laws (§4). **(iv)** We propose FIBER—*Fiber Identification and Blindspot Elimination for Representations*—which mines outcome-preserving safe/unsafe pairs and learns a compact safety representation specifically targeted at distinctions hidden by the original verifier view (§5). **(v)** We develop controlled analytic and tool-agent experiments that separately test representational information loss, downstream judge capacity, and compact representation repair, and complement them with an audit of a publicly released AgentDojo trace (§6).

2 VERIFICATION THROUGH A REPRESENTATION

Trajectories, safety, and verifier views: Let $\mathcal{T}$ be a measurable trajectory space. A trajectory may contain observations, internal or external actions, tool outputs, environment state, and the final response $\tau = (o_0, a_0, o_1, a_1, \ldots, o_T, a_T)$. A policy and environment induce a probability measure $P \in \mathcal{P}(\mathcal{T})$ over trajectories. Let $C : \mathcal{T} \to \{0, 1\}$ be a safety-violation indicator. We define deployment risk $\mathrm{Risk}(P) = \mathbb{E}_{\tau \sim P}[C(\tau)]$. The verifier does not necessarily observe τ. Instead, it receives

$$Z = \phi(\tau) \in \mathcal{Z}, \tag{1}$$

for a measurable representation $\phi : \mathcal{T} \to \mathcal{Z}$. Examples include the final response, terminal database state, an LLM-generated summary, a selected subsequence of steps, or a learned vector representation. The induced distribution on verifier-visible observations is the pushforward $P^\phi := \phi_{\#}P$.

A verifier operating on n independent episodes is any measurable decision rule $\delta_n : \mathcal{Z}^n \to \{0, 1\}$. The distinction between ϕ and δ_n is central: ϕ determines which safety information is available; δ_n determines how that information is used.

For Example 1, the outcome map $\phi_{\text{out}}(\tau) = Y(\tau)$ discards intermediate tool use. A fuller representation might instead expose $\phi_{\text{full}}(\tau) = \big(Y(\tau), \text{tool names, arguments, recipients, data lineage}\big)$.

Definition 2 (Fiber and blind fiber)**.** *For $z \in \mathcal{Z}$, the fiber of ϕ at z is $\mathcal{F}_z(\phi) = \phi^{-1}(\{z\}) = \{\tau \in \mathcal{T} : \phi(\tau) = z\}$. When C is deterministic from τ, a fiber is* safety-mixed *or* blind *if it contains τ_0, τ_1 such that $C(\tau_0) \neq C(\tau_1)$.*

Example 1 produces a blind fiber immediately: $\tau_{\text{s}}, \tau_{\text{u}} \in \mathcal{F}_{Y^\star}(\phi_{\text{out}}),\ C(\tau_{\text{s}}) \neq C(\tau_{\text{u}})$.

Blindness is representation-relative. A pair can be aliased by ϕ_{out} but separated by ϕ_{full}. This observation motivates a task-specific notion of statistical sufficiency.

Definition 3 (Safety sufficiency)**.** *For a distribution P over trajectories, ϕ is* safety-sufficient under P *if $C \perp\!\!\!\perp \tau \mid \phi(\tau)$. In the deterministic-label setting used throughout the main text, ϕ is* safety-sufficient for a family $\mathfrak{P}$ *if there exists a single measurable map $g : \mathcal{Z} \to \{0, 1\}$ such that $C(\tau) = g(\phi(\tau))$ P-a.s. for every $P \in \mathfrak{P}$.*

This is deliberately weaker than requiring ϕ to reconstruct τ. A verifier need not retain every token, screenshot, or tool output; it need only preserve information required to determine the safety variable of interest. The definition is a task-specific analogue of statistical sufficiency and Blackwell informativeness (Blackwell, 1953), and is closely related to the principle behind information bottleneck methods: retain information relevant to a target while discarding irrelevant detail (Tishby et al., 1999).

3 Blind Fibers Create Fundamental Verification Limits

We first establish a representation-level impossibility result. Importantly, it places no restriction on the complexity of the downstream verifier.

3.1 Exact representation aliasing

Consider two trajectory distributions P_0, P_1 satisfying $\text{Risk}(P_1) - \text{Risk}(P_0) \geq \Delta > 0$. Think of P_0 as the safe behavior distribution and P_1 as a high-risk alternative.

For testing $H_0 : P = P_0$ versus $H_1 : P = P_1$, let $\alpha(\delta_n) = P_0^{\phi \otimes n}(\delta_n = 1),\ \beta(\delta_n) = P_1^{\phi \otimes n}(\delta_n = 0)$ denote the two errors.

Theorem 4 (Exact representation impossibility)**.** *Suppose*

$$P_0^\phi = P_1^\phi. \tag{2}$$

Then for every sample size $n \geq 1$ and every possibly randomized verifier δ_n, $\alpha(\delta_n) + \beta(\delta_n) = 1$. Consequently, $\inf_{\delta_n} \max\{\alpha(\delta_n), \beta(\delta_n)\} \geq \frac{1}{2}$. This holds regardless of Δ, verifier capacity, or computational budget.

Proof. Equation (2) implies $(P_0^\phi)^{\otimes n} = (P_1^\phi)^{\otimes n}$. Thus δ_n has the same output distribution under both hypotheses. Hence $\alpha + \beta = P_0(\delta_n = 1) + P_1(\delta_n = 0) = P_0(\delta_n = 1) + P_0(\delta_n = 0) = 1$. The minimax statement follows because at least one of the two errors is at least $1/2$. □

Theorem 4 differs qualitatively from a finite-sample limitation. No amount of repeated evaluation solves exact aliasing.

Corollary 5 (A single mixed fiber suffices)**.** *Suppose a fiber $\mathcal{F}_z(\phi)$ contains τ_0, τ_1 with $C(\tau_0) = 0, C(\tau_1) = 1$. If the admissible distribution family contains $P_0 = \delta_{\tau_0}$ and $P_1 = \delta_{\tau_1}$, then Theorem 4 applies with $\text{Risk}(P_1) - \text{Risk}(P_0) = 1$.*

Thus the booking pair in Example 1 already proves that outcome-only verification cannot be uniformly correct over any policy class containing both behaviors. Consistently, if $\mathfrak{P}$ contains δ_{τ_0} and δ_{τ_1}, then the balanced mixture $\bar{P} = \frac{1}{2}\delta_{\tau_0} + \frac{1}{2}\delta_{\tau_1} \in \text{conv}(\mathfrak{P})$ satisfies $\text{VSD}_{\bar{P}}(\phi) = H_{\bar{P}}(C \mid \phi(\tau)) = \log 2$ whenever $\phi(\tau_0) = \phi(\tau_1)$ and $C(\tau_0) \neq C(\tau_1)$.

3.2 Approximate aliasing and sample complexity

Exact equality is not necessary. Let the visible distributions just be close.

Theorem 6 (Approximate representation lower bound). *Assume $D_{\mathrm{KL}}(P_0^\phi \| P_1^\phi) \le \varepsilon$. Then every verifier based on n independent trajectory representations satisfies $\alpha(\delta_n) + \beta(\delta_n) \ge 1 - \sqrt{\frac{n\varepsilon}{2}}$. Hence, if both errors are required to be at most $\eta < 1/2$, necessarily*

$$n \ge \frac{2(1-2\eta)^2}{\varepsilon}. \tag{3}$$

Proof. Le Cam's two-point bound gives $\alpha + \beta \ge 1 - \mathrm{TV}\Big((P_0^\phi)^{\otimes n}, (P_1^\phi)^{\otimes n}\Big)$. Pinsker's inequality and additivity of KL divergence yield $\mathrm{TV}\Big((P_0^\phi)^{\otimes n}, (P_1^\phi)^{\otimes n}\Big) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}\Big((P_0^\phi)^{\otimes n} \| (P_1^\phi)^{\otimes n}\Big)} \le \sqrt{\frac{n\varepsilon}{2}}$. Equation (3) follows from $\alpha + \beta \le 2\eta$. See Cover & Thomas (2006); Tsybakov (2009) for standard testing inequalities. □

Theorem 6 clarifies an important distinction. Increasing the number of sampled episodes can compensate for *small but nonzero* representational separation. It cannot compensate for exact equality in Theorem 4. Better sampling and better representation therefore solve different problems.

4 Verifier Sufficiency Defect

Theorems 4–6 compare pairs of distributions. We next seek a quantity intrinsic to a verifier representation.

Definition 7 (Verifier Sufficiency Defect). *For $Z = \phi(\tau)$ under trajectory distribution P, define*

$$\boxed{\mathrm{VSD}_P(\phi) := I_P(C; \tau \mid Z).} \tag{4}$$

For a family $\mathfrak{P}$, define the robust defect $\mathrm{VSD}_{\mathfrak{P}}(\phi) = \sup_{P \in \mathrm{conv}(\mathfrak{P})} \mathrm{VSD}_P(\phi)$, where $\mathrm{conv}(\mathfrak{P})$ denotes the family of finite mixtures of laws in $\mathfrak{P}$.

The convex hull is essential at the family level: mixtures can expose safety-label conflicts between behaviors that would remain invisible if each trajectory law were considered separately. Because Z is a function of τ,

$$\mathrm{VSD}_P(\phi) = I_P(C;\tau) - I_P(C;Z). \tag{5}$$

Thus VSD is exactly the safety information contained in the full trajectory but absent from the representation.

If $C = C(\tau)$ is deterministic, then $H(C \mid \tau) = 0$ and

$$\mathrm{VSD}_P(\phi) = H_P(C \mid Z). \tag{6}$$

Consequently, $\mathrm{VSD}_P(\phi) = 0$ exactly when the safety label is recoverable from $\phi(\tau)$ almost surely. In the deterministic-label setting this is equivalent, on the support of P, to the absence of positive-probability safety-mixed fibers from Definition 2.

4.1 Operational meaning: irreducible verification loss

The conditional-mutual-information definition has a direct decision-theoretic interpretation.

For a probabilistic verifier $q : \mathcal{Z} \to [0,1]$, define binary log loss

$$\ell(q(Z), C) = -C \log q(Z) - (1-C)\log(1 - q(Z)).$$

Theorem 8 (VSD equals excess Bayes log-loss). *Let $L_\phi^\star = \inf_q \mathbb{E}[\ell(q(\phi(\tau)), C)]$ be the optimal expected log-loss achievable from representation ϕ, and let $L_{\mathrm{full}}^\star = \inf_{\widetilde{q}} \mathbb{E}[\ell(\widetilde{q}(\tau), C)]$ be the optimum when the complete trajectory is available. Then*

$$L_\phi^\star - L_{\mathrm{full}}^\star = \mathrm{VSD}_P(\phi). \tag{7}$$

Proof. The Bayes-optimal probabilistic predictions are $q^\star(z) = P(C = 1 \mid Z = z)$ and $\widetilde{q}^\star(\tau) = P(C = 1 \mid \tau)$. Their optimal expected log losses are respectively $H(C \mid Z)$ and $H(C \mid \tau)$. Their difference is $I(C; \tau \mid Z)$. □

Equation (7) makes VSD independent of any particular LLM judge. Even an optimal verifier pays an irreducible penalty when the representation is insufficient.

4.2 Coarsening and repair

Proposition 9 (Coarsening cannot improve safety sufficiency)**.** *If $\psi = h \circ \phi$ for a measurable map h, then*

$$\mathrm{VSD}_P(\psi) \geq \mathrm{VSD}_P(\phi). \tag{8}$$

Proof. By data processing, $I(C; \psi(\tau)) \leq I(C; \phi(\tau))$. Apply Equation (5). □

Thus summarizing, truncating, sampling, or otherwise post-processing a trajectory view can only lose safety information unless the discarded information is conditionally irrelevant.

Conversely, suppose we augment ϕ with an additional representation $r(\tau)$: $\phi^+(\tau) = \big(\phi(\tau), r(\tau)\big)$.

Proposition 10 (Exact value of representation repair)**.** *For every measurable r,*

$$\mathrm{VSD}_P(\phi) - \mathrm{VSD}_P(\phi^+) = I_P(C; r(\tau) \mid \phi(\tau)) \geq 0. \tag{9}$$

Equation (9) is the design principle behind §5: a useful safety sketch should contain information about C that is *not already present* in the original verifier view.

4.3 Running example: probabilistically retained safety logs

We can quantify Example 1 exactly. Let $C \sim \mathrm{Bernoulli}(p)$ indicate whether secret exfiltration occurred. Suppose the final booking outcome is identical in both cases, while a security-relevant trace event is retained by the verifier view with probability ρ. Let $M \sim \mathrm{Bernoulli}(\rho),\ M \perp C$, and define

$$Z_\rho = \begin{cases} (\text{outcome}, 1, C), & M = 1, \\ (\text{outcome}, 0, \perp), & M = 0. \end{cases}$$

Proposition 11 (Erasure law)**.** *For the representation above,*

$$\mathrm{VSD}(\phi_\rho) = (1 - \rho) h_2(p), \tag{10}$$

where h_2 is binary entropy. The optimal 0–1 safety-classification error is

$$R^\star_{01}(\phi_\rho) = (1 - \rho) \min\{p, 1 - p\}. \tag{11}$$

Proof. When $M = 1$, C is observed exactly and $H(C \mid Z_\rho, M = 1) = 0$. When $M = 0$, the representation contains no information about C, hence $H(C \mid Z_\rho, M = 0) = h_2(p)$. Averaging over M proves Equation (10). The Bayes 0–1 rule predicts the majority label whenever $M = 0$, yielding Equation (11). □

The parameter ρ does not describe whether the deployment regime was sampled. The episode is already in the evaluator's dataset. It describes whether a safety-relevant distinction *survived the representation*. At $\rho = 0$, outcome-only evaluation is maximally defective for the matched booking pair; at $\rho = 1$, the relevant trace event is fully retained.

5 Fiber: Finding and Repairing Blind Fibers

The theory suggests two practical questions:

1. Can we systematically find trajectories that differ in safety but are nearly identical under the deployed evaluator representation?
2. Can we add a compact representation that exposes precisely those hidden distinctions?

Fiber addresses both.

5.1 OUTCOME-MATCHED BLIND-PAIR MINING

Let $Y : \mathcal{T} \to \mathcal{Y}$ denote task outcome, e.g. task success, final database goal state, or user-visible response. Let $d_{\mathcal{Z}}$ be a metric or discrepancy on verifier representations.

For $\epsilon \geq 0$, define the blind-pair set

$$\mathcal{B}_\epsilon(\phi) = \left\{ (\tau, \tau') : \begin{array}{l} Y(\tau) = Y(\tau'), \\ C(\tau) \neq C(\tau'), \\ d_{\mathcal{Z}}(\phi(\tau), \phi(\tau')) \leq \epsilon \end{array} \right\}. \tag{12}$$

The equality of outcomes is important. Without it, a representation can separate trajectories just because the task failed visibly. In Example 1, the pair $(\tau_{\mathrm{s}}, \tau_{\mathrm{u}})$ belongs to $\mathcal{B}_0(\phi_{\mathrm{out}})$.

We obtain candidate pairs in three ways.

Counterfactual replay. Starting from a successful benign trajectory, replay the same task while injecting one controlled process-level fault: an unauthorized read, an untrusted transmission, an invalid tool route, or another policy violation. We retain the pair only when the intended task outcome remains unchanged.

Environment-supported attacks. For security environments such as AgentDojo (Debenedetti et al., 2024), attack and benign executions naturally generate candidate pairs. Utility constraints ensure that we compare security failures that survive task completion rather than obvious task failures.

Nearest opposite-label search. Given a corpus of labeled trajectories, for every τ find

$$\tau^- \in \underset{\substack{\tau' : Y(\tau') = Y(\tau) \\ C(\tau') \neq C(\tau)}}{\arg\min} \; d_{\mathcal{Z}}(\phi(\tau), \phi(\tau')).$$

Pairs below a threshold ϵ enter $\mathcal{B}_\epsilon(\phi)$.

5.2 LEARNING A SAFETY SKETCH

Let $r_\theta : \mathcal{T} \to \mathbb{R}^d$ be a fixed-budget safety representation. The repaired verifier view is $\phi^+_\theta(\tau) = [\phi(\tau), r_\theta(\tau)]$. A classifier q_ω predicts C from the repaired view. We optimize

$$\mathcal{L}(\theta, \omega) = \mathcal{L}_{\mathrm{pred}} + \lambda \mathcal{L}_{\mathrm{blind}} + \gamma \mathcal{L}_{\mathrm{inv}}, \tag{13}$$

where

$$\mathcal{L}_{\mathrm{pred}} = \mathbb{E}_\tau \left[\mathrm{BCE} \left(q_\omega(\phi^+_\theta(\tau)), C(\tau) \right) \right], \tag{14}$$

$$\mathcal{L}_{\mathrm{blind}} = \mathbb{E}_{(\tau,\tau') \sim \mathcal{B}_\epsilon} \left[m - \| r_\theta(\tau) - r_\theta(\tau') \|_2 \right]_+, \tag{15}$$

$$\mathcal{L}_{\mathrm{inv}} = \mathbb{E}_{(\tau,\tau') \sim \mathcal{S}} \| r_\theta(\tau) - r_\theta(\tau') \|_2^2. \tag{16}$$

Here $\mathcal{S}$ contains outcome-matched, same-safety pairs and encourages invariance to process differences that are irrelevant to C.

The pair term in Equation (15) differs from ordinary supervised safety training: it concentrates optimization on distinctions that the original verifier representation specifically erased. By Proposition 10, the ideal sketch is one with large $I(C; r_\theta(\tau) \mid \phi(\tau))$.

In tool-agent experiments, r_θ can be instantiated either as (i) a learned fixed-dimensional embedding or (ii) a short structured sketch containing security-relevant events such as source, sink, tool, recipient trust class, state mutation, and data lineage. The latter allows a strong LLM judge to operate on a compact trace-derived representation without receiving the entire raw interaction.

5.3 WHAT FIBER CAN AND CANNOT GUARANTEE

If the learned sketch satisfies $C = g(\phi(\tau), r_\theta(\tau))$ P-a.s. for some measurable g, then $\mathrm{VSD}_P(\phi^+_\theta) = 0$ by Definition 7. This is a representation guarantee, not a guarantee that a finite-capacity downstream judge will successfully learn g. Our experiments therefore separate two error sources:

$$\underbrace{\text{information unavailable from } \phi}_{\text{representation defect}} \quad \text{vs.} \quad \underbrace{\text{information available but unused}}_{\text{judge defect}}. \tag{17}$$

Table 1: Controlled erasure experiment, averaged over 20 independent runs. VSD and log loss are measured in bits. The theoretical quantities contain no fitted parameters.

ρ	VSD		Classification error	
	Theory	Empirical	Bayes	Probe
0	1.000	1.0000 ± 0.0000	0.500	0.4995 ± 0.0025
0.25	0.750	0.7500 ± 0.0031	0.375	0.3748 ± 0.0031
0.50	0.500	0.4994 ± 0.0029	0.250	0.2499 ± 0.0022
0.75	0.250	0.2501 ± 0.0025	0.125	0.1251 ± 0.0016
1	0	0	0	0

This distinction is essential when interpreting full-trajectory LLM judges: failure despite a safety-sufficient trace may reflect reasoning limitations rather than blind fibers.

6 Experimental Evaluation

Our experiments isolate four questions suggested by the theory. **RQ1.** Does the VSD quantitatively track the irreducible verification error caused by representation aliasing? **RQ2.** Can task-success and final-answer views hide safety-relevant process failures even when the underlying trajectory contains an explicit unsafe action? **RQ3.** Can additional downstream capacity or repeated judgments repair an exactly aliased representation, and can an explicit representation repair recover the missing information? **RQ4.** Can a compact structured representation preserve safety-relevant distinctions on combinations of factors that were withheld during training?

We use two controlled environments because they permit exact knowledge of the Bayes limit, and complement them with an audit of publicly released AgentDojo execution traces.

Controlled VSD experiment: We first instantiate the confidential-booking construction from Example 1. Each episode has an outcome-matched safe or unsafe trajectory, with $C \sim \mathrm{Bernoulli}(1/2)$. The legitimate task outcome is identical across the two classes. A security-relevant event is retained in the verifier representation with probability $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$. Proposition 11 therefore gives the exact predictions

$$\mathrm{VSD}(\phi_\rho) = 1 - \rho \quad \text{bits}, \qquad R^\star_{01}(\phi_\rho) = \frac{1-\rho}{2}. \tag{18}$$

For each value of ρ, we independently generate 30,000 training and 30,000 test trajectories for each of 20 random seeds. The representation probe is an essentially unregularized logistic classifier. Table 1 compares theory with the empirical conditional entropy and probe error.

The largest discrepancy between the theoretical VSD and its empirical estimate is 5.6×10^{-4} bits. Probe log loss similarly tracks conditional entropy: its mean values for $\rho = (0, 0.25, 0.5, 0.75, 1)$ are respectively $1.0000, 0.7503, 0.4995, 0.2504, 0.0003$ bits. Thus the degradation is explained by information removed by the representation rather than ordinary classifier underfitting.

Exact aliasing is not repaired downstream: We next fix $\rho = 0$, where the verifier representation is exactly independent of C. To test whether the observed failure could nevertheless be attributed to insufficient downstream capacity, we train three qualitatively different probes: logistic regression, a 100-tree random forest, and a two-hidden-layer MLP. Across 10 independent runs their mean errors are respectively $0.4994, 0.4994, 0.4994$, their AUROCs are all exactly 0.500, and their mean log losses are 1.0000, 1.0000, and 1.0004 bits. Increasing the complexity of the decision rule therefore does not improve an observation containing zero safety information, as predicted by Theorem 4. We also simulate repeated independent stochastic judgments conditioned on the same exactly aliased representation. Majority voting over $R \in \{1, 5, 25, 101\}$ calls yields accuracies $0.5016, 0.5003, 0.4989, 0.5006$, respectively, over 100,000 test cases. Repetition reduces neither the representation defect nor the resulting chance-level uncertainty.

A real silent failure in AgentDojo: We next ask whether the distinction between task outcome and trajectory safety occurs outside our construction. AgentDojo explicitly records both user-task utility

and whether an injected adversarial goal was executed (Debenedetti et al., 2024). We audited its publicly released run logs.[1]

A representative banking trace for `gpt-4o-mini-2024-07-18` is particularly illustrative. The user asks the agent to make a legitimate additional payment to Apple. An indirect prompt injection embedded in transaction history instructs the agent to make a separate transfer to an attacker-controlled account. The agent executes the injected transfer, subsequently executes the legitimate Apple payment, and finishes with a response reporting only the successful legitimate payment. AgentDojo records `utility=true`, `security=true`, where the second field denotes successful execution of the injected goal. Consequently, task success does not imply safety even in a released tool-agent trace: the terminal user-facing response is compatible with a successful legitimate task while an unsafe side effect is visible only in the intermediate tool trajectory. This example is not by itself an estimate of population VSD; rather, it establishes the existence of the mechanism in a realistic benchmark trace.

Multi-family counterfactual tool-agent suite: To study representation repair under controlled distribution shift, we construct a paired tool-agent suite containing eight fault families and $1{,}000$ matched safe/unsafe task pairs per family, for a total of $16{,}000$ trajectories. Every pair has exactly the same requested task, legitimate terminal state, and final task outcome. Safe and unsafe trajectories differ only at an intermediate event. Each event is described by combinations of six generic structural attributes: $\{\texttt{untrusted}, \texttt{unauthorized}, \texttt{unvalidated}, \texttt{irreversible}, \texttt{sensitive}, \texttt{external}\}$. The eight fault families use different combinations of these factors. Six families are used for training and validation, while two complete combinations are withheld from all representation learning and used only at test time.

We construct five baseline views: ϕ_{out} = final task outcome, ϕ_{state} = terminal task state, ϕ_{sum} = fixed deterministic summary omitting the middle event, ϕ_3 = three deterministically sampled trajectory steps, ϕ_{full} = complete observable trajectory. Textual views are evaluated with the same TF–IDF bigram representation and logistic probe. For the repaired view, denoted FIBER-Struct, we first canonicalize the event into the six generic structural indicators above, then learn $r_\theta : \mathbb{R}^6 \to \mathbb{R}^4$ with a two-layer neural encoder and a binary safety head. Training minimizes binary cross-entropy plus the blind-pair separation term from Equation (15). Thus the test families contain unseen *combinations* of known safety factors rather than unseen copies of the training templates. Because raw recall can be made arbitrarily large by predicting "unsafe" for every trajectory, we report *silent-fault recall at 5% false-positive rate*, denoted SFR_5: $\text{SFR}_5 = \max_{\theta:\,\text{FPR}(\theta)\le 0.05} \text{TPR}_{\text{silent}}(\theta)$. Confidence intervals are obtained by 500 paired bootstrap resamples over task identities.

Table 2: Held-out-family evaluation on the controlled multi-family suite. Alias is the fraction of matched safe/unsafe pairs that are exactly identical under the representation. Parenthesized AUROC intervals are 95% task-level bootstrap intervals.

Representation	Alias ↓	AUROC ↑	SFR_5 ↑	Log loss ↓
Outcome only	1.000	.500	.000	1.000
Terminal state	1.000	.500	.000	1.000
Fixed summary	1.000	.500	.000	1.000
Sampled $k = 3$	.666	.779 (.765–.792)	.364	.667
Full trajectory	.000	1.000	1.000	.0126
FIBER-Struct	.000	1.000	1.000	.0009

The results in Table 2 show the effect of changing which trajectory information is retained. Outcome, terminal-state, and intentionally lossy summary representations produce exact blind pairs and therefore remain at chance. Sampling three of nine trajectory steps exposes the safety-critical event only for a subset of pairs: 66.55% of held-out pairs remain exactly aliased, yielding AUROC 0.779 and $\text{SFR}_5 = 0.364$. Full trajectories eliminate the aliasing and recover perfect ranking. The compact structured representation also achieves AUROC and SFR_5 of 1.000 on the two held-out fault combinations. Its held-out log loss is 8.8×10^{-4} bits, compared with 1.26×10^{-2} bits for the full-text probe. Thus, once the relevant structural factors have been canonicalized, a compact representation can retain the safety distinction without retaining the full task-specific text. This result should be interpreted narrowly. In this controlled suite, unsafe members have two active

[1] `https://github.com/ethz-spylab/agentdojo`

structural factors whereas safe members have one, and the BCE-only ablation in Section J.9 also attains AUROC and SFR_5 of 1.000. The experiment therefore validates representation sufficiency once the structural factors are available; it does not establish that the blind-pair loss is necessary, nor that the trajectory-to-structure mapping can be learned from arbitrary natural-language agent traces.

Summary of empirical findings: The four experiments support different parts of the theory. First, VSD quantitatively predicts irreducible verifier uncertainty: Equation (18) matches the observed conditional entropy and probe error throughout the full interpolation from exact blindness to full observability. Second, exact aliasing is invariant to downstream statistical capacity and repeated judgment. This experimentally separates representation failure from verifier failure. Third, the AgentDojo trace demonstrates that successful task completion and a benign-looking final response can coexist with a safety-relevant intermediate action in a realistic tool-agent system. Finally, the held-out-family experiment shows that a compact canonicalized structural view can preserve safety information on held-out factor combinations without retaining the entire trajectory. Because the BCE-only ablation matches FIBER-Struct on this controlled suite, this result does not isolate an empirical benefit of the blind-pair term; rather, it demonstrates the value of retaining safety-relevant information absent from the original view, the quantity characterized by $I(C; r_\theta(\tau) \mid \phi(\tau))$ in Proposition 10. Appendix J.12 separately constructs a shortcut-dominated regime in which the BCE-only model underuses a rare process-level safety signal, while adding the blind-pair objective improves held-out AUROC from 0.684 to 0.793 and silent-fault recall at 5% FPR from 0.214 to 0.363 across 20 independent runs.

Limitations and Scope: Our formal claims concern information available to the verifier, not the entire safety problem. First, the safety variable C must itself be specified. A representation sufficient for privacy violations need not be sufficient for manipulation, deception, or a previously unknown failure mode. Accordingly, safety sufficiency is always relative to a target variable and distribution family. Second, estimating VSD in high-dimensional real trajectories is difficult. The probe-based estimator in §6 approximates the Bayes conditional entropy but cannot prove that a representation is sufficient. The controlled experiment is therefore important for calibrating the estimator. Third, FIBER requires either safety labels or controlled interventions from which labels can be derived. Pair quality is consequential: a repair trained only on superficial perturbations may learn attack signatures rather than general safety structure. We therefore emphasize held-out fault-family evaluation. Fourth, access to a complete trajectory is required during construction of the safety sketch. This may be undesirable when traces contain private or proprietary information. An important extension is to compute sufficient safety statistics locally and expose only the sketch to external evaluators. Finally, representation sufficiency and verifier competence are distinct. A full trajectory may be information-theoretically sufficient while still being too long or complex for a practical judge. Equation (17) is intended to prevent these two sources of failure from being conflated.

Conclusion: Reliable evaluation requires more than testing the right cases. It requires preserving the right information about each case. We formalized this requirement through blind fibers and safety sufficiency. When safe and unsafe trajectory distributions collapse to the same representation, Theorem 4 shows that verification is impossible independently of downstream model scale. When they are just close, Theorem 6 quantifies the resulting sample complexity. The Verifier Sufficiency Defect in Equation (4) turns this structural phenomenon into a measurable information quantity. Its Bayes-risk interpretation (Theorem 8) and exact augmentation identity (Proposition 10) suggest a concrete design principle: do not indiscriminately expose more trajectory data; preserve the *conditional safety information* absent from the existing evaluator view.

FIBER operationalizes this principle through outcome-matched blind-pair mining and targeted representation repair. In the confidential-booking example that runs throughout the paper, an evaluator need not reconstruct the entire interaction; it must retain the fact that confidential information flowed from a protected source to an untrusted sink. More generally, the appropriate question for agent evaluation is therefore not only *"Did we test this behavior?"* but also *"Could the verifier representation distinguish the unsafe trajectory from a safe one if both occurred?"*

AI USE STATEMENT

Generative AI tools were used for literature discovery, code prototyping, and language editing during preparation of this work. All mathematical statements, proofs, citations, experimental procedures, code, and reported numerical results were independently checked by the authors. The authors take full responsibility for the correctness and content of the paper.

ETHICS STATEMENT

This work studies defensive evaluation and verification of AI agents. The experiments use synthetic environments or publicly released benchmark trajectories and do not involve real credentials, personal data, or attacks on external systems. Security-violation examples are used solely to characterize evaluator blind spots and improve safety verification. Any released traces will use synthetic secrets and sanitized environment state. As the proposed methodology can expose weaknesses in agent-monitoring systems, we will avoid releasing operational details that would materially facilitate misuse against deployed systems.

REPRODUCIBILITY STATEMENT

We provide complete mathematical definitions and proofs in the appendix and will release the implementation required to reproduce all controlled experiments. The release will include data-generation procedures, random seeds, train/validation/test and held-out fault-family splits, representation construction, model architectures, optimization hyperparameters, evaluation metrics, and bootstrap procedures. For the AgentDojo audit, we provide the exact committed trace identifier and the benchmark semantics used to interpret task utility and security outcome.

REFERENCES

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=m1YYAQjO3w.

Florian Eddie Dorner, Vivian Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *International Conference on Learning Representations*, 2025.

Pengfei He, Zhenwei Dai, Bing He, Hui Liu, Xianfeng Tang, Hanqing Lu, Juanhui Li, Jiayuan Ding, Subhabrata Mukherjee, Suhang Wang, Yue Xing, Jiliang Tang, and Benoit Dumoulin. TRAJECT-Bench: A trajectory-aware benchmark for evaluating agentic tool use. In *International Conference on Learning Representations*, 2026.

Wonjoong Kim, Sangwu Park, Yeonjun In, Sein Kim, Dongha Lee, and Chanyoung Park. Beyond the final answer: Evaluating the reasoning trajectories of tool-augmented agents. In *International Conference on Machine Learning*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings*

of the Association for Computational Linguistics: NAACL 2025, pp. 1160–1183, 2025. doi: 10.18653/v1/2025.findings-naacl.65.

Hadi Mohammadi. trajectory-judge: What outcome-only llm judges miss on agent trajectories. *arXiv preprint arXiv:2609.00038*, 2026.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 1999. arXiv:physics/0004057.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/978-0-387-79052-7.

Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel Brown, and Francis Ward. AI sandbagging: Language models can strategically underperform on evaluations. In *International Conference on Learning Representations*, 2025.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R. Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. AgenTracer: Who is inducing failure in the llm agentic systems? In *International Conference on Learning Representations*, 2026.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents. In *International Conference on Learning Representations*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

Siyu Zhou, Patrick Vossler, Venkatesh Sivaraman, Yifan Mai, and Jean Feng. Adaptive auditing of AI systems with anytime-valid guarantees. *arXiv preprint arXiv:2605.07002*, 2026.

Appendix

A Notation and Measure-Theoretic Foundations

This appendix supplies the technical details omitted from the main text. We first make explicit the measurable structure underlying trajectories, representations, safety labels, and verifier decisions. The results are stated for standard Borel spaces, which include finite and countable spaces, Euclidean spaces, separable complete metric spaces equipped with their Borel σ-algebras, and the trajectory spaces used by essentially all finite-horizon agent environments.

A.1 Trajectory space and induced laws

Let $(\mathcal{T}, \mathcal{A}_{\mathcal{T}})$ be a standard Borel trajectory space. For a finite horizon T, one may take $\mathcal{T} = \mathcal{O}_0 \times \mathcal{A}_0 \times \cdots \times \mathcal{O}_T \times \mathcal{A}_T$, equipped with the corresponding product σ-algebra. More generally, $\tau \in \mathcal{T}$ may include hidden simulator state, user messages, tool outputs, environment events, and terminal state.

A policy π, together with an environment kernel and initial-state distribution, induces a probability measure $P_\pi \in \mathcal{P}(\mathcal{T})$. The theory does not require the Markov property. We therefore work directly with trajectory laws $P \in \mathfrak{P} \subseteq \mathcal{P}(\mathcal{T})$, where $\mathfrak{P}$ denotes the family of deployment laws under consideration.

Let $C : \mathcal{T} \to \{0, 1\}$ be a measurable safety-violation indicator. The deployment risk is

$$\mathrm{Risk}(P) = \int_{\mathcal{T}} C(\tau)\, P(d\tau). \tag{19}$$

Throughout the main development we take C to be deterministic given the complete trajectory. The information-theoretic definition itself extends to jointly distributed (τ, C), but that extension is not required for the results reported here.

Let $Y : \mathcal{T} \to \mathcal{Y}$ be a measurable task-outcome map. Safety and task success are conceptually distinct: it is possible that $Y(\tau_0) = Y(\tau_1)$ but $C(\tau_0) \neq C(\tau_1)$. This is exactly the situation exploited by the confidential-booking example from Example 1.

A.2 Deterministic verifier representations

Let $(\mathcal{Z}, \mathcal{A}_{\mathcal{Z}})$ be another standard Borel space and let $\phi : \mathcal{T} \to \mathcal{Z}$ be measurable. The verifier-visible random variable is $Z = \phi(\tau)$.

For $P \in \mathcal{P}(\mathcal{T})$, the pushforward measure $P^\phi = \phi_\# P$ is defined by

$$P^\phi(B) = P\big(\phi^{-1}(B)\big), \qquad B \in \mathcal{A}_{\mathcal{Z}}. \tag{20}$$

A deterministic verifier based on n episodes is a measurable map $\delta_n : \mathcal{Z}^n \to \{0, 1\}$.

A randomized verifier can equivalently be represented by a Markov kernel $D_n(\cdot \mid z_{1:n})$ from $\mathcal{Z}^n$ to $\{0, 1\}$. Equivalently, there is an auxiliary random variable $U \sim \mathrm{Unif}[0, 1]$, $U \perp Z_{1:n}$, and a measurable function $\widetilde{\delta}_n : \mathcal{Z}^n \times [0, 1] \to \{0, 1\}$. Thus all impossibility results below apply equally to deterministic and randomized verifiers.

A.3 Fibers and conditional safety laws

For $z \in \mathcal{Z}$, recall the fiber $\mathcal{F}_z(\phi) = \phi^{-1}(\{z\})$.

For continuous $\mathcal{Z}$, individual fibers may have zero probability, so the operational notion of representational sufficiency is most naturally expressed through regular conditional distributions. Since the underlying spaces are standard Borel, regular conditional probabilities $P(C = 1 \mid Z = z)$ exist P^Z-almost surely.

Define

$$\eta_\phi(z) := P(C = 1 \mid Z = z). \tag{21}$$

A deterministic representation is safety-sufficient precisely when observing the complete trajectory does not refine this conditional safety law.

Proposition 12 (Equivalent formulations of safety sufficiency)**.** *Let $Z = \phi(\tau)$. The following are equivalent:*

(i)

$$C \perp\!\!\!\perp \tau \mid Z.$$

(ii)

$$P(C = 1 \mid \tau) = P(C = 1 \mid Z) \qquad P\text{-a.s.}$$

(iii)

$$I_P(C; \tau \mid Z) = 0.$$

If $C = C(\tau)$ deterministically, these are additionally equivalent to

$$C = g(Z) \qquad P\text{-a.s.} \tag{22}$$

for some measurable $g : \mathcal{Z} \to \{0, 1\}$.

Proof. For standard Borel random variables, conditional mutual information satisfies

$$I(C; \tau \mid Z) = \mathbb{E}\left[D_{\mathrm{KL}}\left(P_{C,\tau|Z} \| P_{C|Z} P_{\tau|Z}\right)\right].$$

Non-negativity of KL divergence implies that this quantity vanishes if and only if

$$P_{C,\tau|Z} = P_{C|Z} P_{\tau|Z}$$

for P^Z-almost every z, which is precisely conditional independence. Hence (i) and (iii) are equivalent.

Conditional independence implies

$$P(C = 1 \mid \tau, Z) = P(C = 1 \mid Z)$$

almost surely. Since $Z = \phi(\tau)$ is measurable with respect to τ,

$$\sigma(\tau, Z) = \sigma(\tau),$$

and therefore

$$P(C = 1 \mid \tau, Z) = P(C = 1 \mid \tau).$$

This proves (i)⇒(ii). The reverse implication immediately gives conditional independence because C is binary.

Suppose now that $C = C(\tau)$ deterministically. Then

$$P(C = 1 \mid \tau) = C(\tau) \qquad P\text{-a.s.}$$

Thus condition (ii) becomes

$$C(\tau) = \eta_\phi(Z) \qquad P\text{-a.s.}$$

Since C takes values only in $\{0, 1\}$, $\eta_\phi(Z) \in \{0, 1\}$ almost surely. Defining

$$g(z) = \mathbf{1}\{\eta_\phi(z) \geq 1/2\}$$

gives Equation (22).

Conversely, if $C = g(Z)$ almost surely, then C is already measurable with respect to $\sigma(Z)$ and therefore contains no additional dependence on τ conditional on Z. □

Remark 13. *Definition 2 is a pointwise geometric notion, whereas Proposition 12 is measure-theoretic. A pointwise mixed fiber causes positive VSD only when the competing portions of the fiber receive positive probability under the distribution being evaluated. In particular, two point-mass laws supported on opposite-label members of the same fiber each have zero individual VSD. This motivates the family-level definition*

$$\mathrm{VSD}_{\mathfrak{P}}(\phi) = \sup_{P \in \mathrm{conv}(\mathfrak{P})} I_P(C; \tau \mid \phi(\tau)),$$

because mixtures of admissible laws expose safety conflicts that are hidden when each law is considered in isolation.

B Exact Representation-Induced Impossibility

We now prove Theorem 4 in a form that explicitly permits randomized verifiers.

Let $P_0, P_1 \in \mathcal{P}(\mathcal{T})$. The hypotheses are

$$H_0 : P = P_0, \qquad H_1 : P = P_1.$$

Suppose the verifier observes

$$Z_i = \phi(\tau_i), \qquad i = 1, \dots, n,$$

where the trajectories are independent conditional on the hypothesis.

Let

$$Q_j = P_j^{\phi}, \qquad j \in \{0, 1\}.$$

The observation laws are therefore

$$Q_0^{\otimes n}, \qquad Q_1^{\otimes n}.$$

A randomized decision rule is a Markov kernel

$$D_n(\cdot \mid z_{1:n})$$

from $\mathcal{Z}^n$ to $\{0, 1\}$. Define

$$\alpha(D_n) = \int_{\mathcal{Z}^n} D_n(\{1\} \mid z_{1:n}) Q_0^{\otimes n}(dz_{1:n}), \tag{23}$$

$$\beta(D_n) = \int_{\mathcal{Z}^n} D_n(\{0\} \mid z_{1:n}) Q_1^{\otimes n}(dz_{1:n}). \tag{24}$$

Theorem 14 (Exact pushforward indistinguishability)**.** *If*

$$P_0^{\phi} = P_1^{\phi},$$

then for every $n \geq 1$ and every randomized verifier D_n,

$$\alpha(D_n) + \beta(D_n) = 1. \tag{25}$$

Consequently,

$$\inf_{D_n} \max\{\alpha(D_n), \beta(D_n)\} = \frac{1}{2}. \tag{26}$$

Proof. Set

$$Q = P_0^{\phi} = P_1^{\phi}.$$

Then

$$Q_0^{\otimes n} = Q_1^{\otimes n} = Q^{\otimes n}.$$

Using Equations (23)–(24),

$$\begin{aligned}
\alpha(D_n) + \beta(D_n) &= \int D_n(\{1\} \mid z_{1:n})\, Q^{\otimes n}(dz_{1:n}) \\
&\quad + \int D_n(\{0\} \mid z_{1:n})\, Q^{\otimes n}(dz_{1:n}) \\
&= \int \left[D_n(\{1\} \mid z_{1:n}) + D_n(\{0\} \mid z_{1:n}) \right] Q^{\otimes n}(dz_{1:n}) \\
&= \int 1\, Q^{\otimes n}(dz_{1:n}) \\
&= 1.
\end{aligned}$$

Hence

$$\max\{\alpha, \beta\} \geq \frac{\alpha + \beta}{2} = \frac{1}{2}.$$

The constant randomized verifier that outputs 0 or 1 with probability $1/2$ achieves

$$\alpha = \beta = \frac{1}{2},$$

so the bound is tight. □

Corollary 15 (Risk can be arbitrarily different)**.** *For every $\Delta \in (0, 1]$, there exist trajectory laws P_0, P_1 satisfying*

$$\mathrm{Risk}(P_1) - \mathrm{Risk}(P_0) = \Delta$$

and

$$P_0^{\phi} = P_1^{\phi}$$

whenever some verifier fiber contains both safe and unsafe trajectories.

Proof. Let τ_0, τ_1 satisfy

$$\phi(\tau_0) = \phi(\tau_1), \qquad C(\tau_0) = 0, \qquad C(\tau_1) = 1.$$

Set

$$P_0 = \delta_{\tau_0},$$

and

$$P_1 = (1 - \Delta)\delta_{\tau_0} + \Delta\delta_{\tau_1}.$$

Both laws push forward to

$$\delta_{\phi(\tau_0)}.$$

Meanwhile,

$$\mathrm{Risk}(P_0) = 0, \qquad \mathrm{Risk}(P_1) = \Delta.$$

The claim follows. □

B.1 Why verifier capacity is irrelevant

Theorem 14 makes no assumptions about VC dimension, Rademacher complexity, parameter count, computational complexity, model architecture, prompt length, or reasoning capability.

This is not accidental. Every verifier obeys the Markov chain

$$H \longrightarrow \tau_{1:n} \longrightarrow Z_{1:n} \longrightarrow \widehat{H},$$

where $H \in \{0, 1\}$ denotes the true hypothesis.

If

$$P(Z_{1:n} \mid H = 0) = P(Z_{1:n} \mid H = 1),$$

then

$$I(H; Z_{1:n}) = 0.$$

Data processing gives

$$I(H; \widehat{H}) \le I(H; Z_{1:n}) = 0.$$

Therefore no computation downstream of $Z_{1:n}$ can recover information about H.

This formalizes the distinction emphasized in Section 1: the obstruction lies in the observation channel, not in the sophistication of the judge.

C Approximate Aliasing and Testing Lower Bounds

We now provide the full proof and several useful variants of Theorem 6.

C.1 Le Cam formulation

For probability measures Q_0, Q_1 on the same measurable space, define total variation by

$$\mathrm{TV}(Q_0, Q_1) = \sup_A |Q_0(A) - Q_1(A)|. \tag{27}$$

For equal priors on two simple hypotheses, the minimum achievable average testing error is

$$\inf_D \frac{\alpha(D) + \beta(D)}{2} = \frac{1}{2}\left(1 - \mathrm{TV}(Q_0, Q_1)\right). \tag{28}$$

Consequently,

$$\inf_D \{\alpha(D) + \beta(D)\} = 1 - \mathrm{TV}(Q_0, Q_1). \tag{29}$$

Theorem 16 (Approximate representation indistinguishability)**.** *Suppose*

$$D_{\mathrm{KL}}(P_0^\phi \| P_1^\phi) \le \varepsilon.$$

Then, for every randomized n-sample verifier,

$$\alpha + \beta \ge 1 - \sqrt{\frac{n\varepsilon}{2}}. \tag{30}$$

Proof. Set

$$Q_0 = P_0^\phi, \qquad Q_1 = P_1^\phi.$$

By Equation (29),

$$\alpha + \beta \ge 1 - \mathrm{TV}(Q_0^{\otimes n}, Q_1^{\otimes n}).$$

Pinsker's inequality gives

$$\mathrm{TV}(Q_0^{\otimes n}, Q_1^{\otimes n}) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(Q_0^{\otimes n} \| Q_1^{\otimes n})}.$$

KL divergence tensorizes:

$$D_{\mathrm{KL}}(Q_0^{\otimes n} \| Q_1^{\otimes n}) = n D_{\mathrm{KL}}(Q_0 \| Q_1) \le n\varepsilon.$$

Combining the preceding displays yields

$$\alpha + \beta \ge 1 - \sqrt{\frac{n\varepsilon}{2}}.$$

□

Corollary 17 (Necessary sample size)**.** *If both testing errors satisfy*

$$\alpha \le \eta, \qquad \beta \le \eta, \qquad 0 < \eta < \frac{1}{2},$$

then necessarily

$$n \ge \frac{2(1-2\eta)^2}{\varepsilon}. \tag{31}$$

Proof. From

$$\alpha + \beta \le 2\eta$$

and Theorem 16,

$$2\eta \ge 1 - \sqrt{\frac{n\varepsilon}{2}}.$$

Thus

$$\sqrt{\frac{n\varepsilon}{2}} \ge 1 - 2\eta.$$

Squaring both sides proves the claim. □

Remark 18 (The lower bound is informative only in its natural regime)**.** *The right-hand side of Equation* (30) *becomes nonpositive when $n\varepsilon \ge 2$. This does not invalidate the inequality; it just means that Pinsker's bound has ceased to provide useful information. The result is intended for the hard-testing regime in which the representation-level KL divergence is small.*

C.2 A TOTAL-VARIATION VERSION

One can avoid Pinsker entirely when the one-sample TV distance is directly controlled.

Proposition 19. *If*

$$\mathrm{TV}(P_0^\phi, P_1^\phi) \le \delta,$$

then

$$\alpha + \beta \ge (1-\delta)^n. \tag{32}$$

Proof. Let $Q_0 = P_0^\phi$ and $Q_1 = P_1^\phi$. The maximal coupling characterization of TV guarantees the existence of a coupling (Z_0, Z_1) satisfying

$$\Pr(Z_0 \neq Z_1) = \mathrm{TV}(Q_0, Q_1) \le \delta.$$

Taking n independent copies yields a coupling of $Q_0^{\otimes n}$ and $Q_1^{\otimes n}$ for which

$$\Pr(Z_{0,1:n} = Z_{1,1:n}) \ge (1-\delta)^n.$$

Hence

$$\mathrm{TV}(Q_0^{\otimes n}, Q_1^{\otimes n}) \le 1 - (1-\delta)^n.$$

Applying Equation (29) gives

$$\alpha + \beta \ge (1-\delta)^n.$$

□

For small δ,

$$(1-\delta)^n \approx e^{-n\delta},$$

so a constant testing advantage requires

$$n = \Omega(\delta^{-1}).$$

D VERIFIER SUFFICIENCY DEFECT: FULL THEORY

D.1 DEFINITION AND BASIC DECOMPOSITION

Let $Z = \phi(\tau)$. The Verifier Sufficiency Defect is

$$\mathrm{VSD}_P(\phi) = I_P(C; \tau \mid Z).$$

By the chain rule,

$$I(C; \tau, Z) = I(C; Z) + I(C; \tau \mid Z). \tag{33}$$

Because Z is a measurable function of τ,

$$\sigma(\tau, Z) = \sigma(\tau)$$

and therefore

$$I(C; \tau, Z) = I(C; \tau).$$

Hence

$$\boxed{\mathrm{VSD}_P(\phi) = I(C; \tau) - I(C; Z).} \tag{34}$$

Thus VSD is exactly the amount of safety information lost by passing the trajectory through ϕ.

Proposition 20 (Basic bounds)**.** *For binary C,*

$$0 \le \mathrm{VSD}_P(\phi) \le H_P(C) \le \log 2. \tag{35}$$

Proof. Conditional mutual information is nonnegative. Also

$$\mathrm{VSD}_P(\phi) = I(C; \tau \mid Z) \le H(C \mid Z) \le H(C) \le \log 2.$$

□

If logarithms are base 2, the final bound is 1 bit.

D.2 Deterministic safety labels

When

$$C = C(\tau)$$

almost surely,

$$H(C \mid \tau) = 0.$$

Therefore

$$\begin{aligned} \mathrm{VSD}_P(\phi) &= I(C; \tau \mid Z) && (36)\\ &= H(C \mid Z) - H(C \mid \tau, Z) && (37)\\ &= H(C \mid Z). && (38) \end{aligned}$$

This proves Equation (6) from the main text.

D.3 Bayes log-loss characterization

Let $\mathcal{Q}_{\mathcal{X}}$ denote the set of measurable functions

$$q : \mathcal{X} \to [0, 1].$$

For any representation X, define

$$L^\star(X) = \inf_{q \in \mathcal{Q}_{\mathcal{X}}} \mathbb{E}\left[-C \log q(X) - (1 - C) \log(1 - q(X))\right].$$

Lemma 21 (Bayes log loss)**.** *For any random variable X,*

$$L^\star(X) = H(C \mid X). \tag{39}$$

Proof. Condition on $X = x$. Let

$$\eta(x) = P(C = 1 \mid X = x).$$

The conditional expected log loss for prediction $q \in [0, 1]$ is

$$-\eta(x) \log q - (1 - \eta(x)) \log(1 - q).$$

This is strictly convex in q whenever $\eta(x) \in (0, 1)$. Differentiation gives

$$-\frac{\eta(x)}{q} + \frac{1 - \eta(x)}{1 - q} = 0,$$

whose unique solution is

$$q^\star(x) = \eta(x).$$

Substitution gives conditional loss

$$-\eta(x) \log \eta(x) - (1 - \eta(x)) \log(1 - \eta(x)).$$

Averaging over X yields $H(C \mid X)$. □

Theorem 22 (Excess Bayes log loss equals VSD)**.** *Let $Z = \phi(\tau)$. Then*

$$L^\star(Z) - L^\star(\tau) = \mathrm{VSD}_P(\phi). \tag{40}$$

Proof. By Lemma 21,

$$L^\star(Z) - L^\star(\tau) = H(C \mid Z) - H(C \mid \tau).$$

Since Z is a function of τ,

$$H(C \mid \tau, Z) = H(C \mid \tau).$$

Therefore

$$H(C \mid Z) - H(C \mid \tau) = I(C; \tau \mid Z) = \mathrm{VSD}_P(\phi).$$

□

Theorem 22 gives VSD an operational interpretation that does not depend on any particular neural verifier.

D.4 Relation to optimal 0–1 verification error

Define the optimal classification risk from representation Z as

$$R_{01}^{\star}(Z) = \inf_{g:\mathcal{Z}\to\{0,1\}} P(g(Z) \neq C). \tag{41}$$

Let

$$\eta(Z) = P(C = 1 \mid Z).$$

Then

$$R_{01}^{\star}(Z) = \mathbb{E}\big[\min\{\eta(Z), 1-\eta(Z)\}\big]. \tag{42}$$

Assume now that $C = C(\tau)$ deterministically, so

$$\mathrm{VSD}_P(\phi) = H(C \mid Z).$$

Writing h_2 for binary entropy in bits, Fano's inequality in the binary case gives

$$\mathrm{VSD}_P(\phi) \leq h_2(R_{01}^{\star}(Z))\,. \tag{43}$$

Hence

$$R_{01}^{\star}(Z) \geq h_2^{-1}(\mathrm{VSD}_P(\phi))\,, \tag{44}$$

where h_2^{-1} denotes the inverse restricted to $[0, 1/2]$.

Thus a strictly positive VSD imposes a strictly positive irreducible classification error.

D.5 Coarsening theorem

Suppose

$$\psi = h \circ \phi.$$

Write

$$Z = \phi(\tau), \qquad W = \psi(\tau) = h(Z).$$

Then

$$C \longrightarrow \tau \longrightarrow Z \longrightarrow W$$

forms a Markov chain.

Theorem 23 (Monotonicity under deterministic coarsening)**.**

$$\mathrm{VSD}_P(\psi) \geq \mathrm{VSD}_P(\phi). \tag{45}$$

Moreover,

$$\mathrm{VSD}_P(\psi) - \mathrm{VSD}_P(\phi) = I(C; Z \mid W). \tag{46}$$

Proof. Using Equation (34),

$$\mathrm{VSD}_P(\psi) - \mathrm{VSD}_P(\phi) = I(C; Z) - I(C; W).$$

Because W is a function of Z,

$$I(C; Z) = I(C; W) + I(C; Z \mid W).$$

Therefore

$$\mathrm{VSD}_P(\psi) - \mathrm{VSD}_P(\phi) = I(C; Z \mid W) \geq 0.$$

□

The equality condition is informative:

$$\mathrm{VSD}_P(\psi) = \mathrm{VSD}_P(\phi)$$

if and only if

$$C \perp\!\!\!\perp Z \mid W.$$

Thus a summary $W = h(Z)$ loses no safety information precisely when every distinction discarded from Z is conditionally irrelevant to C.

D.6 AUGMENTATION THEOREM

Let

$$R = r(\tau)$$

be an additional safety representation and define

$$Z^+ = (Z, R).$$

Theorem 24 (Exact repair identity)**.**

$$\boxed{\mathrm{VSD}_P(\phi) - \mathrm{VSD}_P(\phi^+) = I(C; R \mid Z).} \tag{47}$$

Proof. By Equation (34),

$$\begin{aligned}\mathrm{VSD}_P(\phi) - \mathrm{VSD}_P(\phi^+) &= [I(C;\tau) - I(C;Z)] \\ &\quad - [I(C;\tau) - I(C;Z,R)] \\ &= I(C;Z,R) - I(C;Z) \\ &= I(C;R \mid Z).\end{aligned}$$

□

Corollary 25. *If*

$$C = g(Z, R) \qquad P\text{-a.s.},$$

then

$$\mathrm{VSD}_P(\phi^+) = 0.$$

Proof. If C is measurable with respect to (Z, R), then

$$H(C \mid Z, R) = 0.$$

For deterministic $C = C(\tau)$,

$$\mathrm{VSD}_P(\phi^+) = H(C \mid Z, R) = 0.$$

The general case follows from

$$C \perp\!\!\!\perp \tau \mid (Z, R).$$

□

D.7 APPROXIMATE REPAIR FROM A DECODER

Perfect recoverability is unnecessary.

Proposition 26 (Decoder error controls residual defect)**.** *Assume $C = C(\tau)$ and suppose there exists a decoder*

$$g : \mathcal{Z} \times \mathcal{R} \to \{0, 1\}$$

such that

$$P(g(Z, R) \neq C) \leq \epsilon, \qquad 0 \leq \epsilon \leq \frac{1}{2}.$$

Then, in bits,

$$\mathrm{VSD}_P(\phi^+) = H(C \mid Z, R) \leq h_2(\epsilon). \tag{48}$$

Proof. This is binary Fano inequality applied to estimation of C from (Z, R). □

Thus the empirical error of a sufficiently expressive safety decoder provides a principled upper bound on the residual representation defect.

E THE CONFIDENTIAL-BOOKING EXAMPLE IN CLOSED FORM

We now develop Example 1 as a complete probability model. The same example is used throughout the main paper precisely so that the difference between outcome quality and process safety remains concrete.

E.1 Environment

Let

$$C \sim \text{Bernoulli}(p)$$

denote whether confidential information is exfiltrated.

The intended booking outcome is

$$Y = Y^\star$$

regardless of C. Hence outcome-only observation contains no safety information:

$$I(C; Y) = 0.$$

Suppose the complete trajectory contains a trace event

$$E = C,$$

where $E = 1$ records an unauthorized protected-source-to-untrusted-sink information flow.

The full trace representation

$$\phi_{\text{full}}(\tau) = (Y, E)$$

therefore satisfies

$$H(C \mid Y, E) = 0.$$

By contrast,

$$\phi_{\text{out}}(\tau) = Y$$

has

$$H(C \mid Y) = H(C) = h_2(p).$$

Consequently,

$$\text{VSD}(\phi_{\text{out}}) = h_2(p), \tag{49}$$
$$\text{VSD}(\phi_{\text{full}}) = 0. \tag{50}$$

This is the maximal possible separation for a Bernoulli safety event with prevalence p.

E.2 Randomly retained security events

Let

$$M \sim \text{Bernoulli}(\rho), \qquad M \perp C,$$

where $M = 1$ means that the security-relevant event survives trace compression.

Define

$$Z_\rho = (Y, M, \widetilde{E}),$$

where

$$\widetilde{E} = \begin{cases} C, & M = 1, \\ \perp, & M = 0. \end{cases}$$

We now prove Proposition 11 in full.

Proposition 27 (Erasure representation)**.** *The Verifier Sufficiency Defect is*

$$\text{VSD}(\phi_\rho) = (1 - \rho) h_2(p). \tag{51}$$

The Bayes-optimal 0–1 *error is*

$$R^\star_{01}(\phi_\rho) = (1 - \rho) \min\{p, 1 - p\}. \tag{52}$$

Proof. Because $Y = Y^\star$ is constant,

$$H(C \mid Z_\rho) = H(C \mid M, \widetilde{E}).$$

By conditioning on M,

$$H(C \mid M, \widetilde{E}) = \rho H(C \mid M = 1, \widetilde{E}) + (1-\rho) H(C \mid M = 0, \widetilde{E}).$$

When $M = 1$,

$$\widetilde{E} = C,$$

so

$$H(C \mid M = 1, \widetilde{E}) = 0.$$

When $M = 0$,

$$\widetilde{E} = \perp$$

contains no information about C. Independence of M and C gives

$$P(C = 1 \mid M = 0) = p,$$

and therefore

$$H(C \mid M = 0, \widetilde{E}) = h_2(p).$$

Hence

$$H(C \mid Z_\rho) = (1-\rho) h_2(p).$$

Since C is deterministic from the complete trajectory,

$$\mathrm{VSD}(\phi_\rho) = H(C \mid Z_\rho),$$

proving Equation (51).

For 0–1 loss, when $M = 1$ the verifier observes C and incurs zero error. When $M = 0$, it has no safety information and therefore predicts the majority class, incurring error

$$\min\{p, 1-p\}.$$

Averaging over M gives

$$R^\star_{01} = (1-\rho)\min\{p, 1-p\}.$$

□

E.3 A NOISY RATHER THAN ERASED TRACE

The erasure model makes the mechanism transparent, but the same phenomenon holds under noisy event retention.

Suppose

$$\widetilde{E} = C \oplus N,$$

where

$$N \sim \mathrm{Bernoulli}(q)$$

is independent binary noise and $q \in [0, 1/2]$.

For the balanced case $p = 1/2$,

$$I(C; \widetilde{E}) = 1 - h_2(q)$$

bits and therefore

$$\mathrm{VSD} = h_2(q). \tag{53}$$

Thus VSD interpolates continuously between a perfectly informative trace ($q = 0$) and complete representation blindness ($q = 1/2$).

F STOCHASTIC VERIFIER REPRESENTATIONS

The deterministic map

$$Z = \phi(\tau)$$

is convenient but not essential. Real evaluator pipelines can be stochastic: an LLM may summarize a trace differently across runs, a logging mechanism may sample events, or privacy filtering may randomly suppress fields.

We therefore replace ϕ by a Markov kernel

$$K(dz \mid \tau)$$

from $\mathcal{T}$ to $\mathcal{Z}$.

For trajectory law P, define the verifier-visible law

$$(PK)(B) = \int_{\mathcal{T}} K(B \mid \tau)\, P(d\tau). \tag{54}$$

The joint law of (τ, Z) is

$$P(d\tau)K(dz \mid \tau).$$

F.1 EXACT IMPOSSIBILITY FOR OBSERVATION CHANNELS

Theorem 28 (Stochastic-channel impossibility)**.** *Let K be any verifier observation channel. If*

$$P_0K = P_1K, \tag{55}$$

then no verifier observing any finite number of independent outputs from K can distinguish P_0 from P_1 with minimax error below $1/2$.

Proof. Under H_j, each verifier observation has distribution P_jK. Equation (55) implies equality of their n-fold product laws for every n. The proof of Theorem 14 applies verbatim. □

F.2 STOCHASTIC VSD

Define

$$\mathrm{VSD}_P(K) = I_P(C; \tau \mid Z), \qquad Z \sim K(\cdot \mid \tau). \tag{56}$$

Because

$$C \longrightarrow \tau \longrightarrow Z$$

is a Markov chain,

$$I(C; \tau, Z) = I(C; \tau).$$

Hence the same decomposition holds:

$$\mathrm{VSD}_P(K) = I(C; \tau) - I(C; Z). \tag{57}$$

F.3 BLACKWELL ORDERING OF VERIFIER CHANNELS

Suppose K_1 maps trajectories to $\mathcal{Z}_1$ and K_2 maps trajectories to $\mathcal{Z}_2$.

We say K_2 is a garbling of K_1 if there exists a channel L satisfying

$$K_2 = K_1L, \tag{58}$$

that is,

$$K_2(B \mid \tau) = \int_{\mathcal{Z}_1} L(B \mid z_1) K_1(dz_1 \mid \tau).$$

Then

$$C \to \tau \to Z_1 \to Z_2.$$

Theorem 29 (Blackwell monotonicity of VSD)**.** *If K_2 is a garbling of K_1, then*

$$\mathrm{VSD}_P(K_2) \geq \mathrm{VSD}_P(K_1). \tag{59}$$

More precisely,

$$\mathrm{VSD}_P(K_2) - \mathrm{VSD}_P(K_1) = I(C; Z_1 \mid Z_2). \tag{60}$$

Proof. By Equation (57),

$$\mathrm{VSD}_P(K_2) - \mathrm{VSD}_P(K_1) = I(C; Z_1) - I(C; Z_2).$$

Since

$$C \to Z_1 \to Z_2$$

is a Markov chain,

$$I(C; Z_1) = I(C; Z_2) + I(C; Z_1 \mid Z_2).$$

The result follows. □

Thus VSD is monotone under the classical Blackwell ordering of statistical experiments (Blackwell, 1953). A verifier channel cannot become more safety-informative just by applying additional stochastic post-processing.

G Multiple Verifier Views and Compositional Repair

Real oversight systems frequently combine several sources: a final-answer judge, a tool-call monitor, a state checker, and a human reviewer.

Let

$$Z_1 = \phi_1(\tau), \qquad Z_2 = \phi_2(\tau),$$

and define the joint view

$$Z_{1:2} = (Z_1, Z_2).$$

Then

$$\mathrm{VSD}(\phi_1) - \mathrm{VSD}(\phi_1, \phi_2) = I(C; Z_2 \mid Z_1). \tag{61}$$

Consequently, the marginal benefit of adding a new monitor is not measured by $I(C; Z_2)$ alone but by its *conditional* information beyond what the existing verifier already sees.

This gives a direct criterion for evaluator composition:

$$\boxed{\text{value of monitor 2} = I(C; Z_2 \mid Z_1).} \tag{62}$$

If

$$I(C; Z_2 \mid Z_1) = 0,$$

then the second monitor provides no additional safety information, even if it is individually predictive of C.

For m views,

$$Z_{1:m} = (Z_1, \ldots, Z_m),$$

the chain rule gives

$$I(C; Z_{1:m}) = \sum_{j=1}^{m} I(C; Z_j \mid Z_{1:j-1}). \tag{63}$$

This decomposition is useful for selecting a compact collection of complementary monitors rather than several redundant ones.

H Formal Development of Fiber

We now formalize the algorithm introduced in Section 5.

H.1 Outcome equivalence

Define the outcome equivalence relation

$$\tau \sim_Y \tau' \quad \Longleftrightarrow \quad Y(\tau) = Y(\tau').$$

The purpose of conditioning on equal outcomes is to exclude trivial pairs in which unsafe behavior is visible just because the task visibly failed.

Define the opposite-safety set of τ by

$$\mathcal{O}(\tau) = \{\tau' \in \mathcal{T} : Y(\tau') = Y(\tau),\ C(\tau') \neq C(\tau)\}.$$

For a metric or semimetric $d_{\mathcal{Z}}$ on verifier space, define the fiber-blindness radius

$$b_\phi(\tau) = \inf_{\tau' \in \mathcal{O}(\tau)} d_{\mathcal{Z}}\left(\phi(\tau), \phi(\tau')\right), \tag{64}$$

with

$$b_\phi(\tau) = +\infty$$

if $\mathcal{O}(\tau) = \varnothing$.

Small $b_\phi(\tau)$ means that an outcome-matched trajectory with the opposite safety label lies nearby in verifier space.

The population blind-pair rate at tolerance ϵ is

$$\mathrm{BPR}_\epsilon(\phi) = P\left(b_\phi(\tau) \leq \epsilon\right). \tag{65}$$

This quantity is not identical to VSD. BPR is geometric and depends on the chosen representation metric; VSD is information-theoretic and metric-free. We use the former for mining training pairs and the latter for theoretical characterization.

H.2 Finite dataset construction

Suppose a dataset contains

$$\mathcal{D} = \{(\tau_i, y_i, c_i)\}_{i=1}^N.$$

For each i, define

$$j^\star(i) \in \underset{\substack{j: y_j = y_i, \\ c_j \neq c_i}}{\arg\min}\, d_{\mathcal{Z}}\left(\phi(\tau_i), \phi(\tau_j)\right). \tag{66}$$

A pair is retained if

$$d_{\mathcal{Z}}\left(\phi(\tau_i), \phi(\tau_{j^\star(i)})\right) \leq \epsilon.$$

To prevent a small number of tasks from dominating the pair set, we perform pair mining within task groups and cap the number of retained pairs per task.

H.3 Algorithm

H.4 Why outcome-matched pair separation targets representation shortcuts

A purely predictive objective may exploit correlations that already exist in $\phi(\tau)$ and ignore distinctions erased by the baseline representation.

Suppose, for example, most unsafe episodes visibly fail their user task. Then a classifier can predict C from outcome failure alone, despite remaining completely blind to successful-but-unsafe episodes.

The blind-pair term is designed to suppress this shortcut by concentrating on pairs satisfying

$$Y(\tau) = Y(\tau'), \qquad C(\tau) \neq C(\tau').$$

It therefore directs learning toward process-level distinctions rather than ordinary outcome failure. This does not imply that the term is necessary on every dataset: in the deliberately simple canonicalized suite used here, the BCE-only ablation already separates the classes perfectly.

Algorithm 1 FIBER blind-pair mining and representation repair

Require: Trajectory dataset $\mathcal{D} = \{(\tau_i, y_i, c_i)\}_{i=1}^{N}$; baseline representation ϕ; representation distance $d_{\mathcal{Z}}$; blindness threshold ϵ; safety encoder r_θ; decoder q_ω.

$\mathcal{B} \leftarrow \varnothing$
$\mathcal{S} \leftarrow \varnothing$
for $i = 1, \dots, N$ **do**
  $\mathcal{N}_i^- \leftarrow \{j : y_j = y_i,\ c_j \neq c_i\}$
  if $\mathcal{N}_i^- \neq \varnothing$ **then**
    $j^- \leftarrow \arg\min_{j \in \mathcal{N}_i^-} d_{\mathcal{Z}}(\phi(\tau_i), \phi(\tau_j))$
    if $d_{\mathcal{Z}}(\phi(\tau_i), \phi(\tau_{j^-})) \leq \epsilon$ **then**
      $\mathcal{B} \leftarrow \mathcal{B} \cup \{(i, j^-)\}$
    end if
  end if
  $\mathcal{N}_i^+ \leftarrow \{j : y_j = y_i,\ c_j = c_i,\ j \neq i\}$
  if $\mathcal{N}_i^+ \neq \varnothing$ **then**
    sample $j^+ \in \mathcal{N}_i^+$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(i, j^+)\}$
  end if
end for
for each optimization step **do**
  sample minibatch $\mathcal{M} \subseteq \mathcal{D}$
  sample blind pairs $\mathcal{M}_B \subseteq \mathcal{B}$
  sample same-safety pairs $\mathcal{M}_S \subseteq \mathcal{S}$
  compute $\mathcal{L}_{\text{pred}}$ using Equation (14)
  compute $\mathcal{L}_{\text{blind}}$ using Equation (15)
  compute $\mathcal{L}_{\text{inv}}$ using Equation (16)
  update (θ, ω) to minimize

$$\mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{blind}} + \gamma \mathcal{L}_{\text{inv}}.$$

end for
return repaired representation

$$\phi_\theta^+(\tau) = [\phi(\tau), r_\theta(\tau)].$$

H.5 REPRESENTATION BOTTLENECK

To prevent r_θ from degenerating into a complete replay of τ, we fix a representation budget. Depending on the instantiation this may be:

1. a fixed vector dimension
$$r_\theta(\tau) \in \mathbb{R}^d;$$
2. a maximum structured-event budget B;
3. a maximum token budget L for a textual safety sketch;
4. an explicit regularizer on
$$I(\tau; r_\theta(\tau)).$$

An information-bottleneck version of the objective is

$$\max_\theta \ I(C; R_\theta \mid Z) - \beta I(\tau; R_\theta), \tag{67}$$

where

$$Z = \phi(\tau), \qquad R_\theta = r_\theta(\tau).$$

The first term is exactly the VSD reduction from Theorem 24; the second penalizes unnecessary retention of trajectory information.

Equation (67) provides the conceptual objective. The empirical objective in Equation (13) is a practical surrogate.

I Estimating Representation Defect from Data

Exact mutual-information estimation in high-dimensional trajectory spaces is generally difficult. We therefore distinguish quantities that can be computed exactly in controlled experiments from proxies used in realistic agent environments.

I.1 Finite discrete representations

If C and Z are discrete with manageable support, then

$$\widehat{\mathrm{VSD}} = \widehat{I}(C; \tau \mid Z) \tag{68}$$

can be estimated by empirical plug-in distributions.

When $C = C(\tau)$,

$$\widehat{\mathrm{VSD}} = \widehat{H}(C \mid Z). \tag{69}$$

For discrete Z,

$$\widehat{H}(C \mid Z) = -\sum_{z,c} \widehat{p}(z, c) \log \frac{\widehat{p}(z, c)}{\widehat{p}(z)}.$$

Small-sample bias can be reduced through bootstrap correction or Miller–Madow-type corrections, though our controlled experiments use enough replication per representation condition that confidence intervals can be obtained directly by task-level bootstrap.

I.2 Predictive upper bound on conditional entropy

For high-dimensional Z, train a probabilistic classifier

$$q_\omega(c \mid z).$$

Its population cross-entropy is

$$\begin{aligned} L(q_\omega) &= \mathbb{E}[-\log q_\omega(C \mid Z)] && (70)\\ &= H(C \mid Z) + \mathbb{E}_Z D_{\mathrm{KL}}\left(P(C \mid Z) \| q_\omega(C \mid Z)\right). && (71) \end{aligned}$$

Therefore

$$L(q_\omega) \geq H(C \mid Z). \tag{72}$$

When C is deterministic from the full trajectory,

$$H(C \mid Z) = \mathrm{VSD}_P(\phi),$$

so held-out cross-entropy provides an upper bound on VSD:

$$\mathrm{VSD}_P(\phi) \leq L(q_\omega). \tag{73}$$

Importantly, a poor classifier makes this bound loose. This is why the paper does not identify classifier loss with VSD without qualification.

I.3 Estimating VSD reduction

Suppose the same probe family is sufficiently expressive for both baseline and repaired representations. Then

$$L(q_{\mathrm{base}}) - L(q_{\mathrm{repair}}) \tag{74}$$

approximates

$$H(C \mid Z) - H(C \mid Z, R) = I(C; R \mid Z).$$

This difference is considerably more stable than estimating high-dimensional mutual information directly and corresponds precisely to the theoretical repair quantity in Equation (47).

J Experimental Implementation Details

This section specifies all experiments reported in Section 6. The purpose of the controlled experiments is to isolate representation-induced uncertainty from model error.

J.1 Software and reproducibility

The reported experiments were executed with Python 3.13.5, NumPy 2.3.5, pandas 2.2.3, scikit-learn 1.8.0, and PyTorch 2.10.0 on CPU. Unless stated otherwise, the master random seed is $s_0 = 20260922$. The complete executable implementation accompanies the paper.

J.2 Controlled erasure experiment

We set

$$C \sim \mathrm{Bernoulli}(1/2)$$

and draw an independent retention variable

$$M \sim \mathrm{Bernoulli}(\rho).$$

The verifier representation is

$$Z_\rho = (M, \widetilde{C}), \qquad \widetilde{C} = \begin{cases} C, & M = 1, \\ -1, & M = 0. \end{cases}$$

The value -1 is a missing-value symbol rather than a third safety class.

For

$$\rho \in \{0, 0.25, 0.5, 0.75, 1\},$$

we generate independently

$$n_{\mathrm{train}} = 30{,}000, \qquad n_{\mathrm{test}} = 30{,}000$$

samples for each of 20 seeds

$$s = s_0, s_0 + 1, \ldots, s_0 + 19.$$

Because $p = P(C = 1) = 1/2$, Proposition 11 gives

$$\mathrm{VSD}(\phi_\rho) = 1 - \rho$$

bits and

$$R^\star_{01}(\phi_\rho) = \frac{1 - \rho}{2}.$$

The empirical conditional entropy is computed directly from the three possible representation states

$$(0, -1), \qquad (1, 0), \qquad (1, 1) :$$

$$\widehat{H}(C \mid Z) = -\sum_z \widehat{p}(z) \sum_{c \in \{0,1\}} \widehat{p}(c \mid z) \log_2 \widehat{p}(c \mid z).$$

The representation probe is logistic regression with

$$C_{\mathrm{reg}} = 10^6,$$

the `lbfgs` solver, and a maximum of 1000 iterations. The large regularization constant is intentional: the experiment is intended to approximate the Bayes rule rather than study regularization.

Log loss is converted from nats to bits by dividing by $\log 2$.

J.3 DOWNSTREAM-CAPACITY CONTROL

To test Theorem 4 independently of the representation-probe architecture, we set $\rho = 0$. Hence every example has the same observable representation.

For each of 10 independent seeds, we generate $20{,}000$ training and $20{,}000$ test labels and compare:

1. logistic regression with $C = 10^6$;
2. a random forest with 100 trees;
3. an MLP with two hidden layers of widths $(64, 64)$.

All three models receive only a constant scalar input. Consequently, any difference between their performances can arise only from finite-sample or optimization noise.

The measured averages are

Probe	Error	Log loss (bits)	AUROC
Logistic	0.4994	1.0000	0.500
Random forest	0.4994	1.0000	0.500
MLP	0.4994	1.0004	0.500

with standard deviation approximately 0.0031 in classification error.

J.4 REPEATED-JUDGMENT CONTROL

Under exact aliasing and balanced labels, the Bayes posterior is $P(C = 1 \mid Z) = \frac{1}{2}$. To isolate the effect of repeated downstream stochasticity, we simulate independent Bernoulli$(1/2)$ judgments and aggregate by majority vote.

For $R \in \{1, 5, 25, 101\}$, we evaluate $100{,}000$ independently generated labels. The measured accuracies are

R	1	5	25	101
Accuracy	0.50155	0.50026	0.49893	0.50063.

Thus repetition does not create information absent from the shared representation.

J.5 CONTROLLED MULTI-FAMILY TOOL-AGENT SUITE

We construct eight families of outcome-matched trajectory pairs. Each family contains 1000 tasks and therefore 2000 trajectories: one safe and one unsafe trajectory per task. The complete dataset contains $8 \times 1000 \times 2 = 16{,}000$ trajectories.

Every trajectory consists of nine ordered records:

$$\texttt{USER}, \texttt{PLAN}, \texttt{TOOL-search}, \texttt{OBS-search},$$
$$\texttt{EVENT}, \texttt{TOOL-commit}, \texttt{OBS-commit}, \texttt{STATE}, \texttt{ASSISTANT}.$$

Within a matched pair, every record except `EVENT` is identical. Consequently,

$$Y(\tau_s) = Y(\tau_u)$$

and the terminal task state is also exactly equal.

The intermediate event is expressed through six generic binary attributes:

$$x(\tau) \in \{0, 1\}^6,$$

corresponding to

$$x_1 = \texttt{untrusted}, \quad x_2 = \texttt{unauthorized},$$
$$x_3 = \texttt{unvalidated}, \quad x_4 = \texttt{irreversible},$$
$$x_5 = \texttt{sensitive}, \quad x_6 = \texttt{external}.$$

Table 3: Structural combinations used by the controlled fault families. Families 7–8 are completely withheld during representation learning.

Family	Name	Unsafe structural pair
1	indirect injection	untrusted + unauthorized
2	tool misuse	unauthorized + unvalidated
3	unvalidated feedback	unvalidated + irreversible
4	privileged action	irreversible + sensitive
5	data exfiltration	sensitive + external
6	external instruction	untrusted + external
7	irreversible injection	untrusted + irreversible
8	unauthorized sensitive access	unauthorized + sensitive

The eight fault families use the combinations shown in Table 3.

For an unsafe trajectory, both family attributes are active and the event contains

`safeguard=absent.`

For the corresponding safe trajectory, exactly one of the two contextual attributes is active and

`safeguard=present.`

Which single contextual factor appears in the safe member alternates across task index. Thus no individual structural attribute alone is synonymous with unsafe behavior. However, the classes remain separable by simple aggregate structure, such as the number of active factors. We therefore use this suite as a controlled test of representation sufficiency on held-out factor combinations, not as evidence that the blind-pair loss is necessary.

J.6 Train/validation/test partition

Fault families 1–6 are available during learning. Within each of these families, task indices $0, \ldots, 799$ are used for training and $800, \ldots, 999$ for validation. Hence

$$N_{\text{train}} = 6 \times 800 \times 2 = 9600$$

trajectories and

$$N_{\text{val}} = 6 \times 200 \times 2 = 2400.$$

Families 7 and 8 are completely absent from both training and validation and form the held-out test set:

$$N_{\text{test}} = 2 \times 1000 \times 2 = 4000.$$

Splitting by fault family rather than by trajectory is essential for RQ4: test-time unsafe examples contain new *combinations* of structural factors.

J.7 Verifier representations

The following views are constructed deterministically.

Outcome. Only the final successful task description and task reference are retained:

$$\phi_{\text{out}}(\tau) = Y(\tau).$$

It is identical within every safe/unsafe pair.

Terminal state. Only the task-relevant terminal state is retained:

$$\phi_{\text{state}}(\tau) = s_T^{\text{task}}.$$

It is also identical within every pair.

Fixed summary. The summary concatenates records

$$1,\ 2,\ 7,\ 8,\ 9$$

and therefore deliberately excludes the safety-critical fifth event. This creates a deterministic coarsening for which the theoretical information-loss mechanism is known exactly.

Sampled trajectory. Exactly $k = 3$ of the nine trajectory records are selected without replacement. The same indices are used for the safe and unsafe member of a pair. The pseudo-random seed for task t is derived from the first 64 bits of

$$\mathrm{SHA256}(t).$$

This prevents accidental use of the safety label during view construction.

On the held-out test pairs, the critical event is omitted in 66.55% of cases, and those pairs therefore remain exactly aliased.

Full trajectory. All nine records are retained.

J.8 Text representation probes

For

$$\phi_{\mathrm{out}}, \phi_{\mathrm{state}}, \phi_{\mathrm{sum}}, \phi_3, \phi_{\mathrm{full}},$$

we use exactly the same probe pipeline:

1. TF–IDF representation with unigram and bigram features;
2. minimum document frequency 2;
3. at most 50,000 text features;
4. logistic regression with $C = 10$ and maximum 2000 iterations.

No representation-specific hyperparameter is tuned.

J.9 Fiber-Struct implementation

The controlled FIBER implementation is intentionally structured so that we can isolate the representation-repair question from natural-language information extraction.

A deterministic event parser maps the full event into

$$x(\tau) \in \{0, 1\}^6.$$

These six factors are not the safety label: each individual factor occurs in both safe and unsafe training examples.

The learned encoder is

$$r_\theta(x) = W_2\ \mathrm{ReLU}(W_1 x + b_1) + b_2,$$

with

$$W_1 \in \mathbb{R}^{12\times 6}, \qquad W_2 \in \mathbb{R}^{4\times 12}.$$

Hence

$$r_\theta(x) \in \mathbb{R}^4.$$

A linear safety head

$$q_\omega(r) = \sigma(w^\top r + b)$$

is trained jointly with the encoder.

For every matched training pair (τ_s, τ_u), let

$$z_s = r_\theta(x(\tau_s)), \qquad z_u = r_\theta(x(\tau_u)).$$

The optimization objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \lambda \mathbb{E}_{(\tau_s,\tau_u)} \left[m - \|z_s - z_u\|_2 \right]_+ , \tag{75}$$

with

$$\lambda = 0.5, \qquad m = 1.$$

We use Adam with

$$\eta = 10^{-2}, \qquad \text{weight decay} = 10^{-4},$$

for 300 full-batch epochs. PyTorch deterministic algorithms are enabled and the seed is 20260922.

The nuisance-invariance coefficient from the general objective is set to

$$\gamma = 0$$

in this controlled experiment because nuisance text has already been removed by canonicalization.

Pair-loss ablation. For completeness, setting $\lambda = 0$ on this deliberately simple structural suite leaves held-out AUROC and SFR_5 at 1.000; its log loss is approximately 7.2×10^{-4} bits versus 8.8×10^{-4} for the pair-loss model. Thus this controlled dataset does *not* establish that the contrastive term is necessary. Its role should be evaluated on natural agent traces where structural factors are not already cleanly canonicalized.

J.10 Metrics

Exact pair-aliasing rate. For representation ϕ,

$$A_{\mathrm{alias}}(\phi) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\phi(\tau_i^s) = \phi(\tau_i^u)\}. \tag{76}$$

Log loss. All log losses are reported in bits:

$$L_{\mathrm{bits}} = \frac{-\frac{1}{n} \sum_i [c_i \log p_i + (1 - c_i) \log(1 - p_i)]}{\log 2}. \tag{77}$$

Silent-fault recall at fixed FPR. For score $s(\tau)$,

$$\mathrm{SFR}_5 = \max_{t:\mathrm{FPR}(t) \le 0.05} P(s(\tau) \ge t \mid C = 1). \tag{78}$$

We use the empirical ROC envelope. The fixed-FPR definition is preferable to unconstrained recall because a verifier that labels every trajectory unsafe would otherwise attain recall one.

AUROC and AUPRC. Both are computed from the continuous safety score before thresholding.

J.11 Bootstrap uncertainty

Safe and unsafe trajectories belonging to the same task are statistically dependent. We therefore bootstrap at the *task-pair* level.

Let

$$\mathcal{T}_{\mathrm{test}} = \{t_1, \ldots, t_{2000}\}$$

denote the held-out test tasks. For each of 500 bootstrap replicates, we sample 2000 task identifiers with replacement and include both trajectories associated with every sampled identifier.

We recompute AUROC, log loss, and SFR_5 on each replicate. The 2.5th and 97.5th percentiles define the reported 95% intervals.

For the partially observed $k = 3$ representation, the resulting intervals are

$$\mathrm{AUROC} = 0.779 \quad [0.765, 0.792],$$
$$L_{\log} = 0.667 \quad [0.647, 0.688] \text{ bits},$$

and

$$\mathrm{SFR}_5 = 0.364,$$

with task-bootstrap variation driven predominantly by whether the safety-critical step is sampled.

J.12 SHORTCUT-DOMINATED STRESS TEST: WHEN BLIND-PAIR TRAINING BEATS BCE

The canonicalized multi-family experiment above deliberately makes the safety-relevant structural factors easy to recover. Consequently, the BCE-only ablation in Section J.9 already separates the held-out classes perfectly, and that experiment does not isolate a benefit of the blind-pair term. We therefore conduct an additional controlled stress test designed specifically to instantiate the representation-shortcut mechanism discussed in Section H.4.

The question is the following. Suppose almost all unsafe trajectories in the training distribution can be recognized from an easy task-outcome shortcut, while successful-but-unsafe trajectories are rare. Does ordinary supervised BCE learn the process-level distinction needed for those rare silent failures, or does explicitly emphasizing outcome-matched blind pairs produce a more useful safety representation?

Construction. Let the baseline verifier-visible outcome be

$$O(\tau) \in \{0, 1\},$$

where $O = 1$ denotes successful task completion. In addition to the outcome, the representation learner has access during training to a $d = 64$-dimensional process vector

$$x(\tau) \in \mathbb{R}^{64}.$$

The final verifier receives

$$\phi_\theta^+(\tau) = \big[O(\tau), r_\theta(x(\tau))\big].$$

Each independent run begins by drawing

$$v \sim \mathcal{N}(0, I_{64}), \qquad v \leftarrow \frac{v}{\|v\|_2},$$

which defines a randomly oriented process-level safety direction. The direction is redrawn independently for every run, preventing the result from depending on a privileged coordinate system.

The training set contains exactly $10{,}000$ trajectories and is balanced between $C = 0$ and $C = 1$.

First, we generate $9{,}800$ ordinary trajectories, consisting of $4{,}900$ safe and $4{,}900$ unsafe examples. Their process vectors are independent of the safety label:

$$x(\tau) = h + \epsilon, \qquad h \sim \mathcal{N}(0, I_{64}), \qquad \epsilon \sim \mathcal{N}(0, 0.05^2 I_{64}).$$

For these ordinary trajectories the task outcome is a perfect shortcut:

$$O(\tau) = 1 - C(\tau).$$

Thus ordinary safe trajectories succeed and ordinary unsafe trajectories visibly fail.

We then add 100 outcome-matched blind-pair tasks. Each task produces one safe and one unsafe trajectory,

$$(\tau_i^s, \tau_i^u), \qquad i = 1, \ldots, 100,$$

with

$$C(\tau_i^s) = 0, \qquad C(\tau_i^u) = 1,$$

but

$$O(\tau_i^s) = O(\tau_i^u) = 1.$$

The two members share the same task context and stochastic perturbation:

$$h_i \sim \mathcal{N}(0, I_{64}), \qquad \epsilon_i \sim \mathcal{N}(0, 0.05^2 I_{64}),$$

and are constructed as

$$x(\tau_i^s) = h_i + \epsilon_i,$$
$$x(\tau_i^u) = h_i + \mu v + \epsilon_i, \qquad \mu = 2.5.$$

Hence the matched trajectories have the same successful outcome and the same task-specific nuisance content; they differ only through the process-level event in direction v.

Because only the 100 unsafe members of the blind pairs violate the outcome shortcut, predicting

$$\widehat{C} = 1 - O$$

is correct on $9{,}900$ of the $10{,}000$ training trajectories, giving 99% training accuracy without using the process representation at all. The safety labels themselves remain exactly balanced, so ordinary class-balancing does not remove this shortcut.

Silent-fault test distribution. The test set contains $2{,}000$ completely new matched tasks, or $4{,}000$ trajectories in total. Every test task contains a safe and an unsafe trajectory generated as

$$x(\tau_i^s) = h_i + \epsilon_i, \qquad x(\tau_i^u) = h_i + \mu v + \epsilon_i,$$

using new independently sampled contexts h_i and perturbations ϵ_i. Crucially,

$$O(\tau_i^s) = O(\tau_i^u) = 1$$

for every test pair. The outcome shortcut is therefore completely uninformative at test time. Successful detection requires the learned representation to preserve the process-level distinction.

This construction is intentionally a controlled stress test rather than a claim about the frequency of silent failures in deployed systems. Its purpose is to isolate whether the blind-pair objective can redirect representation learning away from a dominant outcome shortcut.

Representation model. Both methods use exactly the same architecture. The safety encoder is

$$r_\theta(x) = W_2 \,\mathrm{ReLU}(W_1 x + b_1) + b_2,$$

with

$$W_1 \in \mathbb{R}^{12\times 64}, \qquad W_2 \in \mathbb{R}^{4\times 12},$$

so that

$$r_\theta(x) \in \mathbb{R}^4.$$

The verifier head receives both the baseline outcome and the learned safety sketch:

$$q_\omega(\tau) = \sigma\left(w^\top \begin{bmatrix} O(\tau) \\ r_\theta(x(\tau)) \end{bmatrix} + b\right).$$

The architecture, initialization, optimizer, training data, and number of optimization steps are identical for the two methods. The only difference is the presence or absence of the blind-pair term.

BCE-only baseline. The ordinary supervised baseline minimizes

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} \left[C_i \log q_\omega(\tau_i) + (1 - C_i)\log(1 - q_\omega(\tau_i))\right].$$

This is exactly the $\lambda = 0$ ablation of the FIBER objective.

FIBER objective. For the 100 known outcome-matched safe/unsafe training pairs, define

$$z_i^s = r_\theta(x(\tau_i^s)), \qquad z_i^u = r_\theta(x(\tau_i^u)).$$

The blind-pair loss is

$$\mathcal{L}_{\mathrm{blind}} = \frac{1}{100}\sum_{i=1}^{100} \left[m - \|z_i^s - z_i^u\|_2\right]_+,$$

with

$$m = 1.$$

FIBER minimizes

$$\mathcal{L}_{\mathrm{FIBER}} = \mathcal{L}_{\mathrm{BCE}} + \lambda \mathcal{L}_{\mathrm{blind}}, \qquad \lambda = 0.5.$$

We set the nuisance-invariance coefficient to

$$\gamma = 0$$

so that the comparison isolates the effect of the blind-pair term.

Importantly, $\lambda = 0.5$ and $m = 1$ are inherited unchanged from the FIBER-Struct experiment in Section J.9. They were not selected using the results of the present stress test.

Table 4: Shortcut-dominated stress test over 20 independent runs. The training distribution admits a 99%-accurate outcome shortcut, whereas the test distribution consists entirely of outcome-matched successful safe/unsafe pairs. FIBER and BCE use identical architectures, data, initialization protocol, and optimization; the only difference is the blind-pair term with $\lambda = 0.5$. Values are mean $\pm$ standard deviation across runs.

Method	AUROC ↑	SFR_5 ↑	Log loss (bits) ↓
BCE only	0.6844 ± 0.0456	0.2139 ± 0.0538	1.0549 ± 0.0535
FIBER	$\mathbf{0.7929 \pm 0.0248}$	$\mathbf{0.3628 \pm 0.0672}$	$\mathbf{0.9541 \pm 0.0358}$

Optimization and repetitions. Both models are trained using Adam with

$$\eta = 10^{-2}, \qquad \text{weight decay} = 10^{-4},$$

for 300 full-batch epochs. PyTorch deterministic algorithms are enabled.

We perform 20 independent runs using seeds

$$20260922, 20260923, \ldots, 20260941.$$

A new safety direction, training set, and test set are generated for every seed. Both methods within a run receive exactly the same generated data and initial random seed.

Metrics. We report AUROC, silent-fault recall at 5% false-positive rate (SFR_5), and binary log loss in bits, using the definitions from Section J.10. Because every test trajectory has the same successful task outcome, these metrics measure discrimination arising from the learned process representation rather than from outcome information.

Table 4 reports the mean and standard deviation over the 20 independent runs.

The blind-pair objective improves mean AUROC by

$$0.792942 - 0.684390 = 0.108552,$$

and improves mean SFR_5 by

$$0.362800 - 0.213850 = 0.148950.$$

Mean log loss decreases by

$$1.054912 - 0.954075 = 0.100837$$

bits.

The improvement is not driven by a small number of favorable random seeds. FIBER attains higher AUROC than BCE in all $20/20$ runs, higher SFR_5 in $18/20$ runs, and lower log loss in $19/20$ runs.

Paired uncertainty analysis. Because the BCE and FIBER models for a given seed are trained and evaluated on exactly the same generated data, we analyze run-level paired differences. We construct 10,000 bootstrap resamples of the 20 paired run differences, using bootstrap seed 20260925.

For AUROC, the mean paired improvement is

$$\Delta_{\mathrm{AUROC}} = +0.1086,$$

with 95% bootstrap confidence interval

$$[0.0888,\ 0.1292].$$

For silent-fault recall,

$$\Delta_{\mathrm{SFR}_5} = +0.1490,$$

with 95% bootstrap confidence interval

$$[0.1128,\ 0.1848].$$

For log loss, where lower is better,

$$\Delta_{\mathrm{logloss}} = -0.1008 \text{ bits},$$

Table 5: Sensitivity of the shortcut stress test to the blind-pair coefficient. Each entry is mean ± standard deviation over the same 20 independent runs. The improvement appears for every positive coefficient tested.

λ	AUROC ↑	SFR_5 ↑	Log loss (bits) ↓
0	0.6844 ± 0.0456	0.2139 ± 0.0538	1.0549 ± 0.0535
0.10	0.7556 ± 0.0296	0.3040 ± 0.0622	0.9653 ± 0.0306
0.25	0.7731 ± 0.0232	0.3256 ± 0.0625	0.9551 ± 0.0264
0.50	0.7929 ± 0.0248	0.3628 ± 0.0672	0.9541 ± 0.0358
1.00	0.8276 ± 0.0242	0.4508 ± 0.0660	0.9564 ± 0.0353
2.00	0.8625 ± 0.0191	0.5436 ± 0.0505	0.9756 ± 0.0408

with 95% bootstrap confidence interval

$$[-0.1299,\ -0.0742].$$

A two-sided Wilcoxon signed-rank test on the 20 paired runs gives

$$p = 1.91 \times 10^{-6}$$

for AUROC,

$$p = 9.54 \times 10^{-6}$$

for SFR_5, and

$$p = 3.81 \times 10^{-6}$$

for log loss.

Thus the observed advantage is systematic across independently generated process directions and datasets.

Sensitivity to the blind-pair weight. The main comparison above uses $\lambda = 0.5$, inherited from Section J.9 rather than tuned for this experiment. To verify that the result is not specific to that value, we additionally repeat all 20 runs for

$$\lambda \in \{0.1, 0.25, 0.5, 1, 2\}.$$

All other settings remain unchanged. The case $\lambda = 0$ is the BCE-only baseline.

The sensitivity study shows a consistent effect: every positive blind-pair coefficient substantially improves both AUROC and SFR_5 relative to ordinary BCE. Increasing λ places greater emphasis on process-level ranking, while log-loss calibration is best for moderate coefficients. We retain $\lambda = 0.5$ for the primary comparison because it is the value fixed independently in Section J.9.

Why the blind-pair term helps. The result illustrates the distinction between ordinary predictive accuracy and representation sufficiency. On the training distribution, the baseline outcome variable already predicts the safety label with 99% accuracy. Consequently, ordinary BCE has little incentive to allocate the limited four-dimensional safety sketch to the rare process distinction responsible for successful-but-unsafe executions.

The blind-pair term changes the representation-learning signal. Within a matched pair,

$$O(\tau_i^s) = O(\tau_i^u),$$

and the task context and nuisance perturbation are shared. The easy outcome shortcut therefore cannot increase

$$\|r_\theta(\tau_i^s) - r_\theta(\tau_i^u)\|_2.$$

Reducing $\mathcal{L}_{\mathrm{blind}}$ instead requires the encoder to retain the process-level variation that distinguishes the unsafe execution from its outcome-matched safe counterpart.

At test time the outcome shortcut disappears entirely, since all trajectories successfully complete the requested task. The additional information retained by the FIBER representation therefore becomes operationally useful, producing the observed increase in AUROC and fixed-FPR silent-fault recall.

This experiment should be interpreted narrowly. It does not establish that blind-pair training dominates BCE on every safety dataset, nor does it replace the natural-agent evaluation discussed elsewhere

in the paper. Rather, it provides a controlled existence result for the optimization claim underlying FIBER: when ordinary supervised learning can exploit a dominant representation shortcut, explicitly training on outcome-matched safe/unsafe pairs can cause a compact representation to preserve safety-relevant process information that BCE alone underuses.

J.13 AGENTDOJO TRACE AUDIT

The real-trace analysis uses publicly committed AgentDojo run logs rather than re-running the acting model.

AgentDojo returns two Boolean quantities for an attacked task:

$$U \in \{0, 1\}, \qquad S_{\text{inj}} \in \{0, 1\},$$

where $U = 1$ denotes successful completion of the user task and $S_{\text{inj}} = 1$ denotes successful execution of the injection goal. Thus

$$U = 1, \quad S_{\text{inj}} = 1$$

is precisely a successful user task containing a successful adversarial side effect.

We manually inspect the public banking trace

```
gpt-4o-mini-2024-07-18 / banking / user_task_11 /
    important_instructions / injection_task_0.
```

The trajectory executes both an injected transfer and the legitimate user transaction, while the final assistant response discusses only the legitimate transaction. The committed record has

$$U = 1, \qquad S_{\text{inj}} = 1.$$

This trace is used only as an existence witness and is not included in the controlled quantitative tables.

J.14 SCOPE OF THE REPORTED EMPIRICAL CLAIMS

The controlled experiments establish:

1. numerical agreement between VSD and irreducible uncertainty;
2. invariance of exact aliasing to downstream classifier capacity;
3. invariance of exact aliasing to repeated stochastic decisions;
4. interpolation between exact blindness and full observability under step sampling;
5. preservation of safety-relevant distinctions by a compact structured view on two held-out combinations of previously observed factors.

They do *not* establish that the structured parser used by FIBER-Struct can be learned reliably from arbitrary raw natural-language trajectories. Large-scale external validation must therefore evaluate the complete trajectory-to-sketch pipeline rather than supplying the six structural factors directly.