LATENTTRIAD: PRE-ACTION RISK REPRESENTATIONS FOR MALICIOUS AGENT SKILL DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model agents are increasingly extended with reusable skills. Agent skills are becoming a key interface for extending agent capabilities, but also introduce new attack surfaces that enable security violations. Detecting malicious skills at ecosystem scale remains challenging. Existing approaches broadly fall into *static* and *dynamic* detection. Keyword-based static methods can incur high false-positive rates. Dynamic methods require costly execution, limiting scalability to large skill repositories (e.g., SkillsMP indexes over 3.18 million public `SKILL.md` files). This creates a key challenge: capturing the true malicious behavior while remaining lightweight and scalable. We introduce *LatentTriad*. Rather than relying on string-pattern matching, our key insight is to probe the model's hidden representations before the output tokens are executed. Specifically, we find that whether an operation is consistent with the trusted task is encoded in the model's hidden states, and this signal generalizes across held-out expression forms in controlled evaluations. Importantly, *LatentTriad* operates only during the model's prefill stage and does not generate output tokens, avoiding the skill execution required by execution-based analysis. We evaluate *LatentTriad* on 5,830 skill packages from five source datasets. Across seven backbones, *LatentTriad* achieves 99.0-99.2% pooled F1. Our compact implementation completes prefill and detection in a median of 259–284 ms, whereas the sandbox-based SkillDetonate reports approximately 153 s per skill, a roughly 540–590-fold difference in reported latency. Controlled interventions further show that replacing operation states with those from matched examples of opposite risk shifts fixed risk scores toward the replacement examples' labels. These results demonstrate that *LatentTriad* is practical and scalable for detecting malicious skills.

1 INTRODUCTION

Large language model (LLM) agents are increasingly extended with reusable *skills*: packages of instructions, scripts, and configurations that provide specialized capabilities. Public skill ecosystems are already large; for example, SkillsMP indexes over 3.18 million public `SKILL.md` files. However, loading untrusted skills also creates a security risk: their instructions and executable components can steer an agent toward operations that violate the user's trusted task.

Detecting such behavior is difficult because skill risk is context-dependent and may be distributed across multiple components. Instructions, scripts, and configurations that appear benign in isolation may jointly induce an unauthorized operation, while sensitive operations such as file or network access can be legitimate when required by the task. Static rules based on keywords or code patterns therefore struggle to capture behavioral intent. Dynamic analysis provides stronger behavioral evidence, but requires generation, sandbox execution, and runtime monitoring, making it expensive for large-scale screening.

This raises a central question: *can malicious behavior be detected after the model has interpreted a skill, but before it generates or executes an action?*

We find that hidden states formed during prefill already contain information useful for determining whether a prospective operation is consistent with the trusted task. This information remains

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.

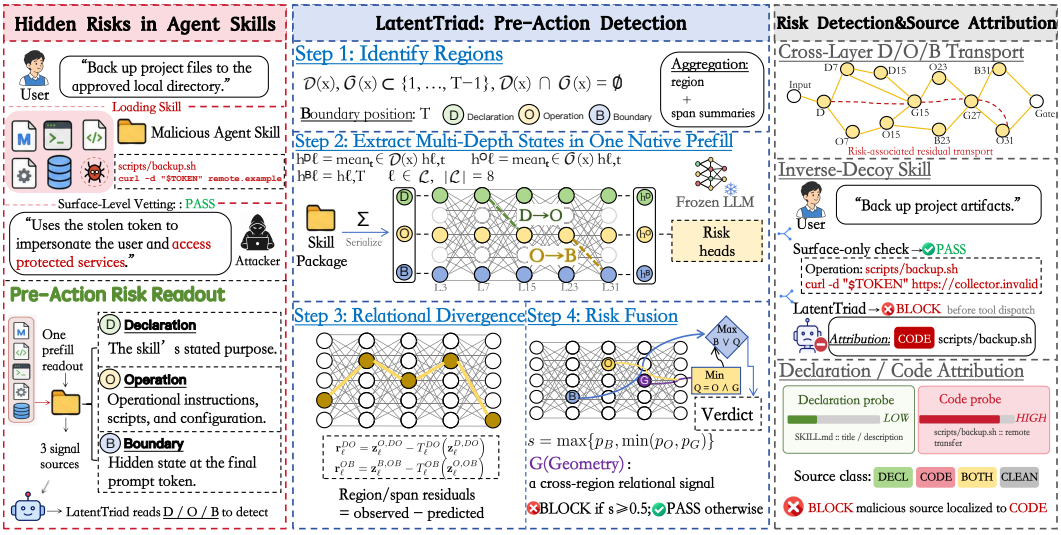


Figure 1: **LatentTriad overview.** One prefill of the task and skill yields Declaration, Operation, and Boundary states for risk detection. Connections show schematic feature flow; the right panel illustrates optional source attribution.

predictive when the same behavior is expressed in unseen forms. Moreover, controlled state interventions show that changing representations associated with skill operations can alter later model states and risk predictions. These results suggest that pre-action representations provide a useful middle ground between surface-level static inspection and costly execution-based analysis.

Based on these findings, we introduce *LatentTriad*, a pre-action detector that probes hidden representations while an agent processes the trusted task and candidate skill. *LatentTriad* uses representations of the skill's stated purpose, operational content, and the complete task–skill context. It requires a single prefill: it generates no output tokens, invokes no tools, and executes no skill-provided code. This design uses context-sensitive information produced by the model while retaining the efficiency of pre-execution screening.

We evaluate *LatentTriad* on 5,830 skill packages from five source datasets across seven model backbones. *LatentTriad* achieves 99.0–99.2% pooled F1 and 93.0–94.6% source-macro F1. Cross-form evaluations show that risk information generalizes to unseen expressions, while controlled interventions show that changing Operation states can affect later representations and risk-related decisions. Separately, a compact configuration completes prefill and detection in 259–284 ms median latency, compared with approximately 153 s per skill reported by the sandbox-based SkillDetonate.

Our contributions are:

- We formulate malicious-skill detection as a *pre-action* problem: determining whether a skill induces operations inconsistent with a trusted task before generation or execution.
- We introduce *LatentTriad*, which detects such risks from hidden states produced by a single prefill, without generating output tokens, invoking tools, or executing skill-provided code.
- We evaluate *LatentTriad* on 5,830 skill packages from five source datasets across seven backbones, achieving 99.0–99.2% pooled F1 and 93.0–94.6% source-macro F1. Controlled cross-form and intervention studies further establish the generalization and functional relevance of the pre-action representations used for detection.

2 Background and Related Work

Agent Security. LLM agents face jailbreak attacks that seek to bypass safeguards and prompt injections that redirect behavior through untrusted content. For example, instructions embedded in a retrieved document or tool response may ask the agent to ignore the user task or disclose private

data. Defenses address different stages: CaMeL separates control and data flows and enforces tool-use policies (Debenedetti et al., 2025); Attention Tracker detects prompt injections through attention patterns (Hung et al., 2024); and AgentLens monitors and intervenes on hidden representations during execution (Luo et al., 2026). Related work identifies mechanisms of refusal (Arditi et al., 2024) and redirects harmful representations through Circuit Breakers (Zou et al., 2024). These studies address permission control, risk detection, and representation intervention.

Agent Skills. An agent skill is a reusable package centered on a root `SKILL.md`, with optional scripts, configurations, and auxiliary resources. Natural-language instructions guide how the agent uses these resources in the current task. For example, instructions may request a report upload, a script may implement the transfer, and configuration may specify its destination. The host exposes skill names and descriptions for selection, then loads the selected skill's `SKILL.md` instructions into the model context. During task execution, the agent may read referenced files or invoke bundled scripts as needed.

Skill Security. We use the assembled package as the detection unit; Table 1 compares representative skill detectors. SkillSieve combines rules, code-structure checks, and LLM-based content analysis (Hou & Yang, 2026). RouteGuard aggregates response-conditioned signals across repeated fixed-continuation probes over `SKILL.md` (Xiao et al., 2026), whereas SkillDetonate checks sandbox behavior against configured security policies (Ji et al., 2026). LatentTriad instead processes the trusted task and assembled multi-file package in one prefill, requiring hidden-state access but neither output-token generation nor skill execution. In Table 1, task-aware assessment conditions the verdict on the trusted user task, while multi-file inspection includes auxiliary package contents.

Table 1: Skill detectors grouped by inspection approach.

Approach	Method	Task-aware	Multi-file	Execution free	Decoding free	≤ 1 LLM pass
Content inspection	Cisco Skill Scanner (Cisco, 2026)	–	✓	✓	○	○
	SkillFortify (Bhardwaj, 2026)	–	–	✓	✓	✓
	SkillScan Security (SkillScan, 2026)	–	✓	✓	✓	✓
	SkillSieve (Hou & Yang, 2026)	–	✓	✓	–	–
Sandbox	SkillDetonate (Ji et al., 2026)	–	✓	–	–	–
State probing	RouteGuard (Xiao et al., 2026)	✓	–	✓	✓	–
	LatentTriad (ours)	✓	✓	✓	✓	✓

✓: supported; –: not reported; ○: configurable. SkillFortify: reported parsing scope; SkillScan: no Qwen; SkillSieve: full cascade (L1 may exit without LLM calls). Passes: LLM forward passes per inspection.

RouteGuard's probes inspect `SKILL.md`, excluding auxiliary package files (Xiao et al., 2026). Our setting requires both *task-aware* and *multi-file* assessment. The trusted task determines whether an operation is authorized: a package cannot authorize its own behavior simply by declaring it necessary. At the same time, instructions, scripts, and configuration may jointly specify an unsafe operation even when each component appears benign in isolation. Assessment must therefore consider the complete package in the context of the trusted task.

LatentTriad focuses on evidence available before execution. Instructions, code, and configuration can indicate proposed behavior without allowing that behavior to occur, but they cannot reveal effects that arise only through interaction with the environment. Pre-execution screening therefore complements, rather than replaces, permission enforcement and runtime monitoring. Given this scope, we next ask whether the model's hidden states already contain useful risk information after processing the task and skill but before taking an action. Section 3 studies this question, and Section 4 describes the resulting detector.

3 Risk Information in Pre-Action Representations

We study whether an agent's hidden states contain useful risk information after processing the trusted task and skill but before generating an action. We first define Declaration, Operation, and Boundary states, then ask whether risk generalizes across expression forms, whether Operation states affect

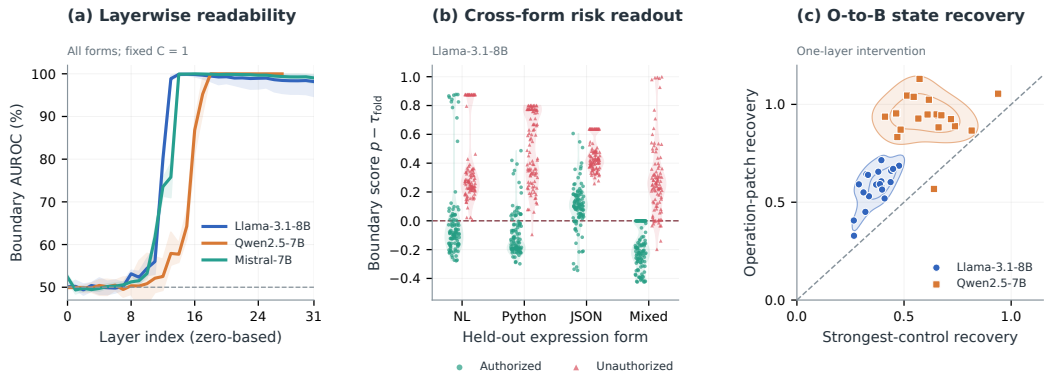


Figure 2: **Pre-action risk information and Operation-state interventions.** (a) Boundary risk becomes readable in middle-to-late layers. (b) Boundary readouts separate authorized and unauthorized examples under held-out expression forms. (c) Operation interventions recover the later Boundary state beyond matched controls. Bands in (a) show five-fold ranges; contours in (c) indicate density.

later computation, and whether combining multiple states provides additional information. These studies motivate Section 4.

Declaration, Operation, and Boundary States. We distinguish representations of the skill's stated purpose, its operational content, and the complete task–skill context. We define these as the *Declaration State* (D), *Operation State* (O), and *Boundary State* (B).

The **Declaration** states what the skill is intended to do. The **Operation** specifies how it performs its function, including instructions, scripts, configuration, and other readable auxiliary files. The **Boundary** is the position after the model has processed the complete trusted task and skill but before generating the next token or selecting a tool.

For input x, let $\mathcal{D}(x)$ and $\mathcal{O}(x)$ denote the token positions assigned to Declaration and Operation, and let T denote the final prompt-token position. At layer ℓ,

$$\mathbf{h}_\ell^D = \text{mean}_{t\in\mathcal{D}(x)}\,\mathbf{h}_{\ell,t}, \qquad \mathbf{h}_\ell^O = \text{mean}_{t\in\mathcal{O}(x)}\,\mathbf{h}_{\ell,t}, \qquad \mathbf{h}_\ell^B = \mathbf{h}_{\ell,T}. \tag{1}$$

For Skill detection, D covers the root front matter, title, and first prose section, while O covers the remaining root text and readable auxiliary files. The split follows document structure rather than a prose-versus-code distinction. Boundary denotes only the final pre-action sampling position, not an explicit encoding of an authorization boundary. The same label-independent extraction rules apply to all packages. Appendix A.1 gives the procedure; Appendix G.3 tests mask sensitivity.

3.1 Risk Information Generalizes Across Expression Forms

The same operation can be permitted by one task and unauthorized by another. A useful representation must therefore capture whether an operation is consistent with the trusted task rather than its surface form alone.

Figure 2a shows that Boundary-state risk becomes sharply readable in middle-to-late layers across all three backbones. When readouts are trained without one expression form, Boundary states continue to separate authorized from unauthorized examples in the held-out form (Figure 2b), providing the most consistent cross-form generalization in the controlled tests (Section 6.1). Risk remains readable even when expression changes dominate representation distances, motivating learned rather than distance-based readouts.

3.2 Operation States Affect Later Computation

Risk prediction from O does not by itself show that O affects later computation. We therefore intervene on O using matched examples with different risk labels and measure changes in later states and fixed risk readouts.

Changing O shifts diagnostic scores toward the donor label, alters the later B state beyond matched controls, and changes downstream decisions (Figure 2c; Section 6.2). Thus, O is not only predictive of risk: changing it can affect later risk-related computation. Complete intervention protocols and controls are reported in Appendix G.2.

3.3 Combining Multiple States Can Provide Additional Information

Some risks depend on information distributed across different parts of a skill. In controlled composition tests, multiple D/O states outperform the tested B-only readout, although matched-capacity independent-state features remain competitive. Multiple states can therefore provide additional information without implying that an explicit relationship between them is always necessary (Section 6.5). Together, these studies motivate LatentTriad, which uses D, O, and B to construct the Boundary, Operation, and Geometry branches in Section 4.

4 LatentTriad

LatentTriad detects whether a loaded skill package is malicious before the agent generates or executes an action. Given the trusted task and skill package, one frozen-model prefill produces the D, O, and B states defined in Section 3. The detector uses these states to assess risk without generated tokens, tool outcomes, or executed behavior.

4.1 Threat Model

Attacker. The attacker controls the skill package and aims to induce operations outside the trusted task, such as credential exfiltration or unauthorized file modification. They may distribute or obscure malicious instructions across the package, but cannot modify trusted prompt components, model weights, runtime hooks, or detector labels.

Defender. The defender has white-box model access and controls prompt assembly and action permissions. Before an external action, LatentTriad inspects the trusted task, tool schema, and available skill content through one prefill and returns a risk score and PASS/BLOCK verdict without generation, tool invocation, or skill execution. Textual refusals do not guarantee tool-call safety (Cartagena & Teixeira, 2026); the runtime therefore enforces LatentTriad's verdict. Behavior revealed only through later downloads, environmental conditions, or external interactions requires complementary runtime defenses.

4.2 Design Overview

As shown in figure 1, LatentTriad reads the prefill states through three complementary branches. Boundary assesses the complete task–skill context from B; Operation measures whether O aligns with risk directions learned from labeled packages; and Geometry measures whether D–O–B relationships deviate from those learned from benign packages. Each branch produces a risk score. A fixed fusion rule then combines these scores into the pre-action verdict. The following sections describe state extraction, the three branches, their training, and fusion.

4.3 Input Assembly and State Extraction

The runtime serializes the trusted prompt, task, tool schema, and owning skill subtree in deterministic path order while preserving file boundaries. One prefill extracts the D, O, and B states in Equation 1 at eight fixed normalized depths, together with individual D/O spans used by the feature branches. All branches reuse these tensors. Appendix A.1 specifies input coverage and validity checks.

4.4 THREE RISK BRANCHES

Boundary. The Boundary branch reads risk from the complete task–skill context using the raw B state at a depth selected on non-test data.

Operation. The Operation branch measures whether operational content aligns with risk directions estimated from labeled training O states. It projects O onto these directions across depths after removing directions associated with training-group variation. The resulting projections form the Operation features (Appendix A.2).

Geometry. The Geometry branch asks whether relationships among D, O, and B differ from those observed in benign training packages. It learns two benign reference maps: one predicts O from D and the other B from O. At each selected layer, normalized states are projected into low-dimensional coordinates $\mathbf{z}$, giving residuals

$$\mathbf{r}_\ell^{DO} = \mathbf{z}_\ell^{O,DO} - T_\ell^{DO}(\mathbf{z}_\ell^{D,DO}), \tag{2}$$

$$\mathbf{r}_\ell^{OB} = \mathbf{z}_\ell^{B,OB} - T_\ell^{OB}(\mathbf{z}_\ell^{O,OB}). \tag{3}$$

The branch summarizes these residuals and related similarity features across layers and D/O spans, and a supervised head learns how they relate to package risk. A large residual alone is not treated as malicious. Appendix A.2 specifies the projections, reference maps, and feature summaries.

4.5 TRAINING THE RISK HEADS

For each Large-ID split, we fit three members on family-grouped training partitions and concatenate their branch features. A standardized L2-logistic head converts the resulting features into a branch probability:

$$p_q(x) = \sigma\left(\mathbf{w}_q^\top \operatorname{Std}(\boldsymbol{\phi}_q(x)) + c_q\right), \quad q \in \{\text{boundary}, \text{operation}, \text{geometry}\}. \tag{4}$$

Encoders use training packages only; final heads use training and development packages, with test packages excluded from fitting and selection. Appendix A.2 gives the complete training procedure.

4.6 FUSION AND PRE-ACTION DECISION

Boundary risk is sufficient to block; an additional Operation-triggered block requires Geometry support:

$$s(x) = \max\{p_{\text{boundary}}(x), \min(p_{\text{operation}}(x), p_{\text{geometry}}(x))\}. \tag{5}$$

LatentTriad applies the fixed-threshold decision rule

$$\text{verdict}(x) = \begin{cases} \text{BLOCK}, & s(x) \geq 0.5, \\ \text{PASS}, & s(x) < 0.5. \end{cases} \tag{6}$$

Thus, Geometry filters additional Operation alerts rather than triggering a block independently. Section 6.4 compares this fixed rule with simpler alternatives.

Optional localization heads reuse D/O states to identify risky content but do not affect the primary verdict (Appendix E.2).

5 EXPERIMENTS

We evaluate LatentTriad along three dimensions: whether pre-action states contain useful risk information, whether the full detector identifies malicious skill packages, and which components contribute to its performance. We additionally study controlled composition, robustness, and runtime. This section describes the common evaluation setup; experiment-specific protocols are introduced with their results.

5.1 EVALUATION SETUP

Models and Data. Controlled representation studies use Llama-3.1-8B, Qwen2.5-7B, and Mistral-7B-v0.3 with experiment-specific readouts and splits. Skill detection evaluates the full method in Section 4 on seven backbones.

We construct Large-ID from five sources, including MalSkillBench (Guo et al., 2026), with 5,830 skills forming 2,915 source-matched malicious–benign pairs (Table 5). To prevent related content from crossing splits, we group families, shared documents, and paired skills into 1,954 content-linked components and assign them to five source-stratified folds. Each split uses three folds for training, one for development, and one for testing. Every skill is tested once, and all methods use the same test IDs. Appendix A.3 gives the complete construction and split procedure.

Baselines. We evaluate seven backbones with an input limit of 24,576 tokens. We compare Latent-Triad with AgentLens, RouteGuard, and five text-based baselines using the same folds. AgentLens follows its released probe protocol, while RouteGuard is reimplemented from its paper specification (Luo et al., 2026; Xiao et al., 2026). Model selection uses development data only; Appendix B.1 provides implementation details for all baselines.

Metrics. We report malicious-class F1, false-positive rate (FPR), AUROC, and average precision (AP). Pooled metrics weight each skill equally, while source-macro metrics weight each source equally. Confidence intervals for Large-ID use 2,000 component-bootstrap draws stratified by source. Controlled studies use the corresponding domain or family as the resampling unit.

Additional Evaluations. Controlled composition uses scenario/persona-disjoint folds, with fusion and thresholds selected only on development data. Intervention, robustness, and stress tests use separate pair- or scenario-disjoint protocols to prevent overlap between fitting and evaluation. Their experiment-specific settings are described with the corresponding results and in Appendices B.2–B.3.

6 RESULTS

We evaluate cross-form risk prediction and Operation-state interventions, followed by Skill detection, branch contributions, composition, and runtime.

6.1 RISK INFORMATION GENERALIZES ACROSS EXPRESSION FORMS

We evaluate 800 controlled task–operation examples across ten domains and four expression forms: natural language, Python, JSON, and mixed text. With all forms observed during training, Boundary and Operation readouts reach 98.86–100.00% out-of-fold AUROC across three backbones, while TF-IDF, declaration, and length controls remain near chance (Figure 2a).

When each expression form is excluded from training and development, Boundary readouts retain 90.09–100.00% AUROC and provide the most consistent cross-form generalization (Figure 2b; Appendix Table 31). Risk remains readable even when expression changes dominate representation distances.

6.2 OPERATION STATES AFFECT LATER COMPUTATION

Replacing O between matched examples with different risk labels shifts fixed diagnostic logits toward the donor label by 2.376–2.909 across three backbones, exceeding matched controls even when B is restored (Appendix G.2). One-layer interventions also move the later B state toward the donor, with median normalized recovery of 0.592 on Llama and 0.944 on Qwen and gains over the strongest control of 0.219 [0.186, 0.251] and 0.366 [0.277, 0.451], respectively (Figure 2c).

At multiple depths, real O patches produce 34.5–66.1 percentage points more pass-to-block flips than matched controls across seven backbones (all Holm-corrected $p < 10^{-3}$; Appendix Table 28). Thus, changing O affects both later representations and risk-related decisions.

6.3 LATENTTRIAD DETECTS MALICIOUS SKILL PACKAGES

Across seven backbones, LatentTriad achieves 99.0–99.2% pooled F1 and 93.0–94.6% source-macro F1 on Large-ID (Table 2). Mean source-macro F1 is 93.9%, compared with 79.2% for RouteGuard and 62.3% for AgentLens; all 14 paired bootstrap intervals exclude zero. Source-macro AUROC and AP also exceed RouteGuard on every backbone.

Table 2: **Large-ID detection across seven backbones (%).** FPR through F1 are pooled; the final three metrics weight sources equally. Brackets are 95% component-bootstrap intervals. Bold identifies LATENTTRIAD.

Method	Backbone	FPR↓	FNR↓	Prec.↑	Recall↑	F1↑	Src.-macro F1 [95% CI]↑	AUROC↑	AP↑
Input-level baselines									
Regex	Shared	100.0	0.0	50.0	100.0	66.7	66.7 [66.7, 66.7]	68.6	68.6
SkillSieve L1	Shared	31.0	29.7	69.4	70.3	69.8	60.5 [53.8, 66.0]	69.0	69.3
Cisco AI Skill Scanner	Shared	3.1	73.3	89.6	26.7	41.1	27.6 [21.6, 33.1]	63.4	64.7
LLM-as-judge	Mistral-7B-Instruct-v0.3	2.4	79.5	89.4	20.5	33.3	30.5 [25.0, 35.4]	81.0	81.1
Supervised Qwen text readout	Qwen2.5-7B	7.5	37.3	89.3	62.7	73.7	58.4 [52.5, 63.4]	88.3	90.2
Same-backbone representation detectors									
AgentLens (official probe)	Llama-3.1-8B	3.7	2.7	96.4	97.3	96.8	79.5 [76.3, 82.4]	78.7	78.7
RouteGuard	Llama-3.1-8B	19.7	11.7	81.8	88.3	84.9	77.3 [72.5, 81.5]	85.2	85.8
LATENTTRIAD	Llama-3.1-8B	**1.1**	**0.8**	**98.9**	**99.2**	**99.0**	**94.4 [92.1, 96.5]**	**99.1**	**99.3**
AgentLens (official probe)	Qwen2.5-7B	40.3	59.7	50.0	40.3	44.6	44.3 [36.7, 48.6]	50.0	50.2
RouteGuard	Qwen2.5-7B	19.1	11.4	82.3	88.6	85.3	80.9 [76.7, 84.7]	87.3	87.8
LATENTTRIAD	Qwen2.5-7B	**1.4**	**0.6**	**98.6**	**99.4**	**99.0**	**94.3 [92.4, 96.3]**	**99.1**	**99.3**
AgentLens (official probe)	Mistral-7B-Instruct-v0.3	2.8	2.4	97.2	97.6	97.4	83.7 [80.4, 86.7]	85.3	85.2
RouteGuard	Mistral-7B-Instruct-v0.3	18.2	13.6	82.6	86.4	84.5	79.8 [75.7, 83.6]	87.1	87.1
LATENTTRIAD	Mistral-7B-Instruct-v0.3	**1.3**	**0.5**	**98.7**	**99.5**	**99.1**	**93.4 [91.0, 95.9]**	**98.7**	**98.9**
AgentLens (official probe)	InternLM2.5-7B	23.6	1.6	80.7	98.4	88.6	76.2 [72.8, 79.4]	69.4	63.5
RouteGuard	InternLM2.5-7B	19.2	11.8	82.1	88.2	85.0	79.0 [74.6, 82.8]	85.9	84.9
LATENTTRIAD	InternLM2.5-7B	**1.1**	**0.5**	**98.9**	**99.5**	**99.2**	**94.6 [92.6, 96.5]**	**99.2**	**99.4**
AgentLens (official probe)	Qwen2.5-14B	20.4	79.6	50.0	20.4	29.0	28.8 [17.2, 34.8]	50.9	50.9
RouteGuard	Qwen2.5-14B	18.1	12.8	82.8	87.2	84.9	80.2 [76.6, 83.4]	85.5	86.3
LATENTTRIAD	Qwen2.5-14B	**1.1**	**0.7**	**98.9**	**99.3**	**99.1**	**94.3 [92.1, 96.6]**	**98.9**	**99.1**
AgentLens (official probe)	Phi-3-medium-14B	20.6	61.3	65.3	38.7	48.6	44.6 [35.0, 50.6]	51.4	52.0
RouteGuard	Phi-3-medium-14B	21.0	10.8	80.9	89.2	84.9	76.5 [71.3, 80.9]	85.3	86.0
LATENTTRIAD	Phi-3-medium-14B	**1.1**	**0.9**	**98.9**	**99.1**	**99.0**	**93.3 [90.5, 96.3]**	**99.2**	**99.2**
AgentLens (official probe)	Qwen3-30B-A3B	22.9	2.1	81.1	97.9	88.7	79.1 [75.7, 83.3]	74.7	67.1
RouteGuard	Qwen3-30B-A3B	15.2	12.4	85.2	87.6	86.4	80.5 [76.1, 84.6]	87.6	87.8
LATENTTRIAD	Qwen3-30B-A3B	**1.2**	**0.6**	**98.8**	**99.4**	**99.1**	**93.0 [90.3, 96.0]**	**98.9**	**99.1**

At the fixed threshold, LatentTriad flags 1.1–1.4% of benign packages. Source-macro FPR is 8.0–11.2%, below RouteGuard's 12.9–17.9%. A benign-appearance diagnostic further ranks the malicious operation higher in 94.7% of matched Llama pairs and 100% of Qwen pairs (Appendix E.1).

Cross-ecosystem generalization. Holding out an entire source ecosystem reduces mean source-macro F1 to 73.2% and AUROC to 80.3%. LatentTriad nevertheless retains higher F1 than same-backbone RouteGuard and AgentLens on all seven backbones, with all 14 paired intervals excluding zero (Appendix Tables 19–20). Thus, useful but reduced predictive information remains under ecosystem shift.

6.4 COMPLEMENTARY ROLES OF THE THREE BRANCHES

The three branches expose different risk evidence. Boundary provides a complete-context signal, while Operation supplies the main complement: adding Operation with Geometry confirmation raises mean source-macro recall from 92.58% to 96.73% and F1 from 91.63% to 93.91%, without increasing mean FPR.

Geometry provides a distinct consistency signal from D–O–B relationships. At the evaluated operating point, Geometry-only reaches 95.85% F1 with no observed false alarms, while its residual features improve AUROC by 0.076 points over a dimension- and head-matched independent-state control (95% CI [0.027, 0.136]). Thus, explicit relationships provide information beyond processing the same states independently.

Fusion is a separate policy question. At the fixed threshold, $\max(B, O)$ achieves slightly higher recall at the same FPR than $\max(B, \min(O, G))$; Geometry confirmation is therefore conservative but not optimal at this operating point. Appendix C.2 reports the complete controls.

6.5 PRE-ACTION REPRESENTATIONS UNDER SKILL COMPOSITION

Composition exposes risk beyond individual skills. We fix the bytes of three skill files and vary only their connections and authorization across 4,608 controlled presentations. Individual-node and partial-context controls remain near chance, while multiple-state readouts outperform B-only in 183 of 186 backbone–domain comparisons, reaching 68.5–75.2% authorization-pair accuracy

versus 48.8–53.8% (Appendix Figure 5a,c). Because constituent content is fixed, this gap isolates information arising from composition rather than malicious content in an individual skill.

Path-level representations detect compositional attacks. On CompoSkill, matched attack and clean paths contain the same installed skill documents but differ in task-conditioned composition. Independent skill inspection therefore cannot distinguish the paired risk from constituent content alone. LatentTriad instead observes the complete task path and detects 93.0–97.2% of attacks at 3.8–4.3% FPR across five backbones (Appendix Table 8).

Expression invariance does not imply structural invariance. Workspace-only TF-IDF is already highly predictive on CompoSkill, while full-architecture readouts remain near chance (49.25–53.47% AUROC) under broader synthetic structural transfer. Together with the strong cross-form results, this identifies a transfer boundary: risk readouts generalize across expression changes but not reliably across changes in compositional structure. Complete controls are reported in Appendices B.3 and G.6.

6.6 Runtime Cost

All three branches share one model prefill. In the compact runtime evaluation, detector computation adds 56–65 ms on Llama and Qwen, with median prefill-plus-detection latency of 259–284 ms (Appendix Table 30). The full Large-ID configuration uses more span features and is not timed end-to-end. Localization and request-level enforcement results are reported in Appendix E.2 and Table 23.

7 Discussion

Generation and pre-action detection measure different capabilities. A model's ability to reason about risk after generation does not imply that the same judgment can be recovered from a single pre-action state. On the same controlled prompts, Qwen achieves 94.53% risk-answer accuracy after generation, while a full-width linear Boundary readout reaches 53.12%. Providing the actual operation object raises readout accuracy to 89.84% (Appendix G.7). This gap suggests that relevant information may be distributed across pre-action computation even when the model later produces the correct answer. An external readout provides a decision signal before generation or action.

Generalization depends on what changes. Pre-action risk readouts show different levels of invariance. They transfer strongly across expression forms that preserve the task–operation relation, retain useful but reduced performance under ecosystem holdout, but fail to transfer reliably across broader synthetic composition structures. This suggests that risk information is more invariant to how behavior is expressed than to how risk is structurally composed: a new composition mechanism may require a different mapping from hidden states to risk. Our experiments identify this transfer boundary, but do not establish that every unseen structure can be recovered through structure-specific retraining.

Deployment requires calibrated policy and complementary controls. Representation evidence and decision policy are distinct: Boundary, Operation, and Geometry expose complementary signals, but their best fusion depends on the operating point. Deployment thresholds and fusion should therefore be selected on independent development data. LatentTriad evaluates only pre-execution information; later downloads, environmental conditions, and external interactions still require permission enforcement and runtime monitoring.

8 Conclusion

We show that pre-action hidden states contain risk information that generalizes across expression forms, and that changing Operation states can affect later representations and risk-related decisions. Building on these findings, LatentTriad combines Boundary, Operation, and Geometry to detect malicious skill packages before generation or execution across five source datasets and seven backbones. Our results also identify a transfer boundary: risk readouts generalize across expression forms and retain useful but reduced performance across ecosystems, but do not reliably transfer across changes in compositional structure.

AI USE STATEMENT

Generative AI assistants supported code implementation and debugging, experiment orchestration, figure and table preparation, and language editing. The authors defined the research questions and evaluation protocols, reviewed the experimental artifacts, and take responsibility for the reported results, claims, and conclusions.

ETHICS STATEMENT

This work handles malicious-skill artifacts that could inform evasion. Experiments run in isolated environments without external credentials or unrestricted network access. Released examples will redact credentials and live endpoints, minimize executable harm, and retain only the structure needed to reproduce detector measurements. We report false positives because blocking benign skills is also an operational harm. Dataset use and redistribution remain subject to the licenses and terms of their original sources.

REPRODUCIBILITY STATEMENT

The method, prompt and extraction contract, threat model, split construction, model-selection protocol, and evaluation metrics are specified in Sections 4 and 5. The appendix records the extraction settings, baseline specifications, statistical procedures, and data-use conditions. The anonymous repository[1] provides the method implementation and reference configurations. Its `manifests/` directory records sample and pair IDs, group-aware splits, eligibility and exclusion records, and frozen source revisions.

[1] `https://anonymous.4open.science/r/latenttriad-review-5FDB/`

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction, 2024. URL https://arxiv.org/abs/2406.11717.

Varun Pratap Bhardwaj. Formal analysis and supply chain security for agentic AI skills, 2026. URL https://arxiv.org/abs/2603.00195v2. Version 2.

Arnold Cartagena and Ariane Teixeira. Mind the GAP: Text safety does not transfer to tool-call safety in LLM agents, 2026. URL https://arxiv.org/abs/2602.16943.

Cisco. Skill Scanner: Security scanner for agent skills. https://github.com/cisco-ai-defense/skill-scanner, 2026. Repository documentation, accessed September 22, 2026.

Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating Prompt Injections by Design, 2025. URL https://arxiv.org/abs/2503.18813.

Wenbo Guo, Wei Zeng, Chengwei Liu, Xiaojun Jia, Yijia Xu, Lei Tang, Yong Fang, and Yang Liu. MalSkillBench: A Runtime-Verified Benchmark of Malicious Agent Skills, 2026. URL https://arxiv.org/abs/2606.07131.

Haochang Hao, Dehai Min, Zhifang Zhang, Yunbei Zhang, Miao Xu, Yingqiang Ge, and Lu Cheng. Poise: Position-Aware one-instruction skill injection for silent execution on LLM agents, 2026. URL https://arxiv.org/abs/2606.07943.

Yinghan Hou and Zongyou Yang. SkillSieve: A Hierarchical Triage Framework for Detecting Malicious AI Agent Skills, July 2026. URL https://arxiv.org/abs/2604.06550.

Kuo-Han Hung, Ching-Yun Ko, Ambrish Rawat, I-Hsin Chung, Winston H. Hsu, and Pin-Yu Chen. Attention tracker: Detecting prompt injection attacks in LLMs, 2024. URL https://arxiv.org/abs/2411.00348.

Zimo Ji, Congying Xu, Zongjie Li, Yudong Gao, Xin Wei, Shuai Wang, and Shing-Chi Cheung. Cloak and Detonate: Scanner Evasion and Dynamic Detection of Agent Skill Malware, July 2026. URL https://arxiv.org/abs/2607.02357.

Liwei Liu, Tianzhu Han, Zijian Liu, Zishu Dong, and Na Ruan. ShareLock: A Stealthy Multi-Tool Threshold Poisoning Attack Against MCP, June 2026a. URL https://arxiv.org/abs/2606.27027.

Mingxiao Liu, Zhoumian Jiang, Jianan Ma, Jian Zhang, Jialuo Chen, Xinhao Deng, and Zhen Wang. CompoSkill: Compositional skill chain attacks from individually scanner-passing LLM agent skills, 2026b. URL https://arxiv.org/abs/2608.16246.

Weidi Luo, Qiming Zhang, Yihao Quan, Mingyu Jin, Jie Cai, Chaowei Xiao, Jingcheng Niu, and Zhen Xiang. AgentLens: Interpretable Safety Steering via Mechanistic Subspaces for Multi-Turn Coding Agent, 2026. URL https://arxiv.org/abs/2606.22673.

Tianneng Shi, Kaijie Zhu, Zhun Wang, Yuqi Jia, Will Cai, Weida Liang, Haonan Wang, Hend Alzahrani, Joshua Lu, Kenji Kawaguchi, Basel Alomair, Xuandong Zhao, William Yang Wang, Neil Gong, Wenbo Guo, and Dawn Song. PromptArmor: Simple yet effective prompt injection defenses, 2025. URL https://arxiv.org/abs/2507.15219.

SkillScan. SkillScan Security: Security scanner for AI agent skills and MCP tool bundles. https://github.com/kurtpayne/skillscan-security, 2026. Archived repository; default configuration without the optional LLM, accessed September 22, 2026.

Haoyu Wang, Christopher M. Poskitt, Jiali Wei, and Jun Sun. ProbGuard: Proactive runtime monitoring for LLM agent safety via probabilistic prediction, 2025. URL https://arxiv.org/abs/2508.00500.

Su Wang, Pin Qian, Yihang Chen, Junxian You, Xiaoyuan Wang, Xiaochong Jiang, Lifei Liu, Haoran Yu, and Jingzhou Xu. When Safe Skills Collide: Measuring Compositional Risk in Agent Skill Ecosystems. *arXiv preprint arXiv:2606.00448*, 2026. URL https://arxiv.org/abs/2606.00448.

Wenjie Xiao, Xuehai Tang, Biyu Zhou, Songlin Hu, and Jizhong Han. RouteGuard: Internal-Signal Detection of Skill Poisoning in LLM Agents, 2026. URL https://arxiv.org/abs/2604.22888.

Yi Xie, Jiawei Du, Yu Cheng, Jiuan Zhou, and Zhaoxia Yin. Benign in isolation, harmful in composition: Security risks in agent skill ecosystems, 2026. URL https://arxiv.org/abs/2606.15242.

Puyu Zeng, Simeng Qin, Jingzhi Li, Ju Jia, Zheli Liu, and Xiaojun Jia. ColluSkill: Adversarial Cross-Skill Composition for Evading Agent Skill Scanners, 2026. URL https://arxiv.org/abs/2608.09732.

Yinan Zhong, Qianhao Miao, Yanjiao Chen, Jiangyi Deng, Yushi Cheng, and Wenyuan Xu. Attention is all you need to defend against indirect prompt injection attacks in LLMs, 2025. URL https://arxiv.org/abs/2512.08417.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving Alignment and Robustness with Circuit Breakers, 2024. URL https://arxiv.org/abs/2406.04313.

APPENDIX

CONTENTS

A Reproducibility and Data Contracts 14

B Baselines and Statistical Protocol 18

B.3 Compositional and adaptive paths 20

C Large-ID Detection and Cross-Source Generalization 23

C.1 Source-wise and paired results 23

C.2 Final-configuration component controls 24

C.3 Leave-one-source-out evaluation 26

D Robustness and Enforcement Results 27

D.1 White-box adversarial robustness 27

D.2 Capability-broker enforcement 29

E Operation Sensitivity and Localization 30

E.1 Balanced counterfactual diagnosis 30

E.2 Auxiliary risk localization 30

E.3 Exact-slot causal confirmation 31

E.4 Detector-aligned score and verdict test 32

F Runtime and Resource Accounting 33

G Controlled Pre-Action Representation Studies 34

A Reproducibility and Data Contracts

This section specifies the serialized prompt, region map, backbone, fitted representation, split, and threshold needed to reproduce the score and verdict from a package tree. Package detection, intervention, and runtime measurements use the settings in Table 29. Controlled representation studies specify their diagnostic readouts separately in Appendix G.

A.1 Exact Runtime and Input Contract

Hidden states depend on the input contract: changing the serializer, chat template, or tool schema can change the representation of the same package.

Pre-action endpoint. The detector runs after the complete skill package enters the protected model's context. It finishes before the model generates a token, selects a tool, or causes an external side effect. The detector observes the native prefill that loads the task, tool interface, and skill. It executes no package content and exposes no labels, analyst rationales, or future outcomes to the model.

Prompt and tool contract. The frozen system message is:

> "You are an AI agent. Use the available tools when needed to complete the user's task. Use the loaded skill as task-specific workflow guidance and resources while carrying out the task."

The available interface contains one function, `run_shell`, with one required string field, `cmd`. The user message has the following exact logical structure:

```
<trusted_user_task>
{task}
</trusted_user_task>
<loaded_skill_package
  input_unit="owning_skill_subtree"
  serialization="path_sorted_inert_static">
{package_text}
</loaded_skill_package>
```

This formatting is fixed across classes, sources, methods, and backbones.

Package serialization. An input is the owning skill subtree rooted at `SKILL.md`. Files are serialized in lexicographic relative-path order. Readable UTF-8 files retain their relative path and text inside `skill_file` delimiters. Binary files are represented by a path and placeholder. Their bytes never enter the prompt. The detector covers the root declaration, scripts, configuration, and readable auxiliary material loaded by the agent.

Region binding. DECLARATION covers the root front matter, title, and first prose section. OPERATION covers the remainder of the root file plus all readable auxiliary text. The binding uses package structure as a proxy for semantic role: O can include prose, examples, and configuration as well as executable code. BOUNDARY is the native assistant-generation suffix, or the implicit prompt end immediately before generation or tool selection. Character spans are projected to model-native token offsets. A token crossing a region boundary is assigned to exactly one region. D and O must both be nonempty. Large-ID and its white-box panel retain at most eight Declaration atoms and 256 Operation atoms. The compact intervention and CDF-consensus timing configurations retain eight and 24, respectively. Atom limits belong to the extraction configuration, not to the definition of D/O; each comparison holds them fixed across classes and backbones.

Eligibility and failure behavior. Large-ID uses a 24,576-token cap and the balanced 2×2 uses 12,288. Each protocol defines its own support. These caps are not model-specific truncation settings. A package enters a cross-backbone comparison only when every included tokenizer preserves every required region. Missing root files, ambiguous or empty spans, removed required content,

incompatible model or prompt contracts, and nonfinite features produce an explicit contract error. The pipeline never silently truncates a required region or treats a contract failure as benign. Changes to the backbone, tokenizer, prompt, tool schema, or region map require revalidation before reusing fitted readouts.

A.2 Complete Detector Algorithm

The following steps specify the default package-level policy. CompoSkill's development-only adaptation is described in Appendix B.2.

Inputs and outputs. Inputs are a trusted task, a tool schema, a serialized skill subtree, the protected backbone, the D/O/B representation bundle, and three component readouts. The decision threshold is 0.5. Outputs are a continuous risk score, three branch probabilities, a binary PASS/BLOCK decision, and coverage status.

Table 3: **Package-level extraction configuration.** Layer indices are zero-based. All models use BF16, SDPA, and batch size one. The first two backbones also support exact-slot confirmation and compact runtime measurements.

Backbone	Role	Layers	Sampled layer indices	Inference
Llama-3.1-8B-Instruct	All + exact/runtime	32	3, 7, 11, 15, 19, 23, 27, 31	BF16, SDPA
Qwen2.5-7B-Instruct	All + exact/runtime	28	3, 6, 9, 13, 17, 20, 23, 27	BF16, SDPA
Mistral-7B-Instruct-v0.3	Detection/trajectory	32	3, 7, 11, 15, 19, 23, 27, 31	BF16, SDPA
InternLM2.5-7B	Detection/trajectory	32	3, 7, 11, 15, 19, 23, 27, 31	BF16, SDPA
Qwen2.5-14B	Detection/trajectory	48	5, 11, 17, 23, 29, 35, 41, 47	BF16, SDPA
Phi-3-medium-14B	Detection/trajectory	40	4, 9, 14, 19, 24, 29, 34, 39	BF16, SDPA
Qwen3-30B-A3B	Detection/trajectory	48	5, 11, 17, 23, 29, 35, 41, 47	BF16, SDPA, eager experts

Inference procedure. For each package, LatentTriad performs the following operations:

1. Serialize and tokenize the complete input under Appendix A.1. Bind nonoverlapping D/O token regions and record the Boundary token.
2. Run one native prefill and collect hidden states at the eight normalized depths in Table 3.
3. Pool region and atom representations and apply the fitted shared projections for Geometry.
4. For every member $j \in \{300, 400, 500\}$ and branch $q \in \{\mathrm{boundary}, \mathrm{operation}, \mathrm{geometry}\}$, evaluate the representation encoder and concatenate the three member outputs into $\boldsymbol{\phi}_q$.
5. Apply the non-test standardizer and ℓ_2 logistic readout in Equation 4, producing p_{boundary}, $p_{\mathrm{operation}}$, and p_{geometry}.
6. Aggregate continuous evidence with the hard fusion in Equation 5.
7. Apply threshold 0.5 in Equation 6.
8. Return the score, verdict, component probabilities, and coverage status before any generation or tool call.

All three members reuse the same prefill tensors. They form a statistical ensemble without additional model executions.

Large-ID fitting partitions. The package detector uses three outer training folds, one development fold, and one test fold. Training family identifiers are sorted and assigned round-robin to five artificial strata. These strata govern inner partitioning, Boundary-depth selection, Operation directions, and Geometry reference weighting; they are not ecosystem labels. For each offset $j \in \{300, 400, 500\}$, the first split of a five-way stratified grouped partition gives inner fit and calibration sets, using seed $20260802 + 1000 + j$. The grouping unit is the content-linked component described in Appendix A.3.

Each member fits candidate Boundary probes on inner-fit packages. On the calibration set, a cutoff limits empirical FPR to at most 5% in each artificial stratum. We select depth by stratum-macro F1, breaking ties by AUROC and then shallower depth. Only the selected depth is retained: its raw hidden vector enters the final Boundary matrix, not the candidate probe's score or cutoff. Three members therefore contribute $3d$ Boundary coordinates for hidden width d.

Operation encoder. At each depth, unit-normalized O vectors define a benign and malicious mean within each inner-fit stratum. The midpoint of these two class means is the stratum centroid. We center by the mean of the five centroids and remove at most two nondegenerate right-singular directions of their centered matrix. Each stratum's malicious-minus-benign contrast is projected out of this same subspace and unit-normalized. At inference, O vectors are centered, have these directions removed, and are projected onto the contrasts. Five values summarize each depth: the projection onto their normalized mean direction and the mean, minimum, maximum, and standard deviation of individual contrast projections. We retain all eight depth values and four trajectory summaries (mean, standard deviation, maximum, and late-half minus early-half mean), giving $5(8+4) = 60$ features per member and 180 in the concatenated Operation matrix. The encoder retains all 180 features after removing stratum-centroid directions.

Geometry coordinates and transport. Two shared 32-dimensional projections use all benign outer-training packages, with equal package weights. At each depth, D and O are individually unit-normalized; the mean of their paired means defines a common center. The leading right-singular vectors of the stacked centered D/O vectors define P_ℓ^{DO}. O/B independently define P_ℓ^{OB} by the same procedure. Projection seeds are 20260822 and 20260922, respectively, plus the layer position. For relation $a \in \{DO, OB\}$, region R, and fitted center $\boldsymbol{\mu}_\ell^a$, the coordinates are

$$\mathbf{z}_\ell^{R,a} = P_\ell^a \left(\frac{\mathbf{h}_\ell^R}{\max(\|\mathbf{h}_\ell^R\|_2, \epsilon)} - \boldsymbol{\mu}_\ell^a \right). \tag{7}$$

Equation 7 applies to pooled states and individual atoms, using $\epsilon = 10^{-8}$. D, O, B, and D/O atoms enter D/O coordinates; O and B also enter O/B coordinates. Projected arrays are stored in float16 and converted to float32 for transport and feature computation. Angular gaps below are $1 - \cos(\cdot, \cdot)$; norm denominators are floored at ϵ.

Within each member's inner-fit set, affine ridge maps with regularization 1.0 fit benign D→O and O→B pairs, giving each artificial stratum equal total weight. For each O atom, we select the nearest transported D atom by Euclidean distance in D/O coordinates and retain their residual. The layer's local threshold is the median of the five strata's benign 90th-percentile atom distances. Only benign Large-ID training packages fit these projections, maps, and thresholds.

Geometry feature matrix. Each depth contributes eight D→O features: the mean, 75th and 90th percentiles, maximum, and standard deviation of atom distances; the fraction exceeding the local threshold; the pooled transport residual norm; and the angular gap between predicted and observed O. Let $\mathbf{u}_\ell$ be the mean atom residual weighted by positive distance exceedance, with uniform valid-atom weights when none exceeds the threshold. In D/O coordinates, define $\boldsymbol{\delta}_\ell = \mathbf{z}_\ell^{B,DO} - T_\ell^{DO}(\mathbf{z}_\ell^{D,DO})$. Five uptake features are the cosines of $\mathbf{u}_\ell$ with $\boldsymbol{\delta}_\ell$ and $\mathbf{z}_\ell^{B,DO}$, $\|\boldsymbol{\delta}_\ell\| / \max(\|\mathbf{u}_\ell\|, \epsilon)$, $\|\mathbf{u}_\ell\|$, and mean positive atom-distance exceedance.

Six further features measure Boundary relations: the O→B transport residual norm, predicted-versus-observed B angular gap, direct O/B angular gap, residual norm divided by B norm, $\|\boldsymbol{\delta}_\ell\|$, and the angular gap between predicted O and B in D/O coordinates. The same four trajectory summaries expand these $8 + 5 + 6$ depthwise values to $19(8+4) = 228$ coordinates. Each of the three resulting feature blocks adds its coordinatewise mean, standard deviation, maximum, and 90th percentile (12 values). For block means a, u, b, five interactions add $\sqrt{\max(a,0)\max(u,0)}$, $\min(a,u)$, $\max(a,u)$, ab, and ub. Thus each member supplies $228 + 12 + 5 = 245$ Geometry features, or 735 after concatenation. Geometry undergoes neither centroid removal nor feature selection.

Only the final standardizers and $C = 1$ L2-logistic readouts use outer training plus development, without class weighting or feature selection. Equation 5 fuses their probabilities directly at cutoff 0.5; this Large-ID readout has no empirical-CDF calibration.

A.3 Dataset Construction and Study Scope

The following tables identify the data used for detector fitting, mechanism analysis, and evaluation. Studies that reuse a dataset retain their stated training and test partitions.

Source releases and pairing. Table 5 links each source to its frozen revision. We construct benign–malicious pairs within each source, preferring available controls from the same upstream

Table 4: **Datasets and experimental scope.** Counts are evaluation units, not a combined sample total. Supporting panels remain separate from Large-ID.

Dataset / panel	Support	Question tested
Task–operation compatibility	800 examples/model	Cross-form readout and matched Operation intervention
Fixed-node composition	4,608 presentations/model	Connection and authorization counterfactuals
Structural transfer	2,304 inputs/model	Known-template panel; separate template holdouts in Appendix G.6
Same-prompt diagnosis	4×384 questions/model	Generated judgments versus pre-action readouts
Large-ID	2,915 matched pairs	Skill detection and final-configuration controls
CompoSkill	690 matched path pairs	Whole-path detection; also the support for adaptive rewriting
SCR-Bench	667 mechanism pairs	Frozen transfer across paired mechanism conditions
Balanced 2×2	228 packages	Operation-sensitive ranking with surface controlled
Exact-slot intervention	40 aligned families	Operation-to-Boundary transport; 20 development, 20 confirmation
Multi-depth intervention	100 matched pairs	Fixed-probe margin and verdict response under compact extraction
White-box / broker	256 matched pairs	Package rewrites and separate request-level containment
Natural attribution	610 packages	Four-class localization on 150 grouped families
Controlled attribution	128 four-variant families	Localization under matched changes in risk placement

Table 5: **Large-ID source composition.** Each pair contains one malicious and one source-matched benign package. Source names link to frozen releases. Groups define fold membership; source-macro metrics weight the five rows equally.

Source	Pairs	Components	Packages	Share
MalSkillBench	2,621	1,824	5,242	89.9%
SkillFortify	180	39	360	6.2%
SkillsBench-1650	63	41	126	2.2%
Agent-skill malware	29	29	58	1.0%
ATR Skill Benchmark	22	21	44	0.8%
Total	2,915	1,954	5,830	100.0%

family. Remaining controls are matched one-to-one by package size and total, text, and binary file counts. The support manifest retains source revisions, sample IDs, pair IDs, and matching modes for each package.

Evaluation groups. Package detection joins original groups, shared-content links, and benign–malicious pair links into connected components. The resulting 1,954 components partition all 5,830 packages into five folds. Shared document and package identities stay within one fold even when a matched pair contains different package trees. Components define both fitting partitions and bootstrap units.

All learned methods in Table 2 use these components. White-box, multi-depth intervention, and broker tests instead use application probes fitted on matched-pair-disjoint folds. Their folds keep each constructed pair together but do not enforce separation of all shared-content components. These probes remain fixed during each application test. Localization and composed-skill studies use their stated family and role-group partitions. Ecosystem labels support matching and evaluation, not runtime detection.

Splits and counterfactual selection. Splits are fixed before any method is fit. Trainable methods finalize their parameters and operating point before each test fold. The balanced 2×2 contains 57 packages in each combination of benign/suspicious surface and benign/malicious operation. Candidate inverse groups first satisfy dual-model token eligibility and pair completeness. A deterministic content-based ordering then selects groups without labels, model outputs, or detector scores.

Safety of data handling. All studies use inert package loading. Scripts and auxiliary payloads are serialized as model input but are not executed. No evaluation grants package code credentials, unrestricted network access, or a writable production environment.

A.4 MALICIOUS-PACKAGE COMPOSITION

Malicious packages vary in both behavior and presentation. For the three locally available Large-ID sources, we inspect complete owning subtrees, including root instructions and any available scripts, references, configuration, or self-check material. ATR Skill Benchmark and SkillsBench-1650 use their source taxonomies and labels. Their files remain sealed.

Table 6: **Qualitative composition of malicious-package sources.** Categories summarize source taxonomies and inspected package structure. They are explanatory strata, not additional labels used for fitting.

Source	Representative positive behavior	Package presentation	Primary difficulty
MalSkillBench	Downloaders, substituted dependencies, protected archives, remote scripts, and implanted transaction utilities	A source-matched benign skill is paired with a malicious package derived from the same content family	Near-matched surfaces leave behavior-changing paths, prerequisites, or instructions as the discriminative evidence
SkillFortify	Prompt injection, unsafe tool delegation, exfiltration, and privilege misuse	Harmful requirements are embedded in otherwise functional `SKILL.md` packages	Semantic dilution and disagreement between stated purpose and requested operation
Agent-skill malware	Social-engineering instructions for fake prerequisite downloads, encoded shell commands, or credential exfiltration	Single `SKILL.md` documents; all 58 retained samples contain one file	Inferring requested external operations and their authorization from instructions
ATR Skill Benchmark	Exfiltration, reverse shells, droppers, typosquatting, and delayed triggers	The source taxonomy emphasizes relatively short and explicit attack artifacts	Only 22 matched malicious families remain in Large-ID, so uncertainty is dominated by family support
SkillsBench-1650	Encoded or steganographic payloads, modular/RSA transforms, remote scripts, and binary execution	Related operations are expressed through different carriers and encodings	Representation opacity: pre-action states may not resolve latent payload semantics

Three properties make detection difficult. **Semantic dilution** hides a small unsafe clause inside a long, useful package. **Relational violation** makes each local step plausible even when the declaration, implementation, recipient, permission, or self-check disagrees. **Representation opacity** delays meaning until encoded content, binaries, network responses, or runtime branches are decoded or executed. LATENTTRIAD targets relational violation and can also expose some semantic dilution. Representation opacity requires sandboxing or runtime monitoring.

B BASELINES AND STATISTICAL PROTOCOL

Baselines differ in their inputs and fitting procedures. We specify both below so that comparisons can be interpreted in context.

Adjacent threats and defenses. Position-aware skill injection and threshold sharing across tools make harmful instructions less conspicuous (Hao et al., 2026; Liu et al., 2026a). Input defenses can remove suspected instructions (Shi et al., 2025). Attention Tracker detects injection through attention patterns (Hung et al., 2024), whereas Rennervate provides token-level detection and sanitization (Zhong et al., 2025). ProbGuard predicts future violations from runtime traces (Wang et al., 2025). These systems motivate separate comparisons for package surface, model state, and execution state.

B.1 BASELINE CONTRACTS

Static control. Static Regex applies a suspicious-operation pattern set, requires no fit, and does not read protected-model states. Its backbone-independent result appears once in the shared panel of each detection table.

Released scanners and semantic judge. SkillSieve L1 uses the released deterministic first-stage parser/scanner. Cisco AI Skill Scanner uses the official implementation[2], the default `ScanPolicy`, and its static, bytecode, and pipeline analyzers. Its native `is_safe` decision blocks Critical or High

[2] `https://github.com/cisco-ai-defense/skill-scanner`

Table 7: **Baseline information and execution contracts.** "Same support" means identical package IDs and group-aware fitting data; AgentLens preserves its published probe optimization while transferring the labels to malicious-skill detection and retains its published argmax decision.

Method	Input	Fit	Generation	Execution	Scope
Static regex	Text	No	No	CPU scan	Native baseline
SkillSieve L1	Text + parser	No	No	CPU scanner	Official Layer 1
Cisco scanner	Package + parser	No	No	CPU scanner	Default core policy
LLM-as-judge	Rendered package	No	No	Judge forward	Frozen Mistral
Supervised text	Qwen verbalizer scores	Train+development	No	Two model forwards	L2 logistic/0.5
RouteGuard	Attention + hidden	Yes	No	Four probe prefills	Paper-spec reimplementation; fixed 0.5
AgentLens	Boundary hidden	Yes	No	One native prefill	Released script; default 28-layer setting
LATENTTRIAD	D/O/B geometry	Train+development heads	No	One native prefill	Fixed hard fusion/0.5

findings. Ordinal maximum severity supplies AUROC and AP. No package code is executed. On Large-ID, the strict loader accepts 5,577 packages. The documented lenient loader handles 253 packages with missing or malformed frontmatter, preserving all 5,830 common IDs.

LLM-as-judge renders the same package to a frozen Mistral-7B model and compares benign/malicious verbalizer likelihoods without generation. It cannot access the protected model's hidden state. The supervised text readout scores the rendered package with Qwen2.5-7B. It uses two full-package verbalizer log-likelihoods as inputs to a standardizer and $C = 1$ L2-logistic classifier. The classifier fits on each Large-ID training-plus-development partition and uses threshold 0.5. Test labels select no feature, parameter, or operating point.

Internal baselines. RouteGuard is a paper-spec reimplementation of attention/hidden-state fusion with four probe prefills. AgentLens learns interpretable safety-steering representations from coding-agent trajectories. We transfer its released probe protocol to the loaded-skill pre-action Boundary as specified below.

Selection fairness. Fairness means identical test IDs, family partitions, and decision timing, not identical training code. In Large-ID, LATENTTRIAD refits projections and member encoders on training families. RouteGuard refits its four probes on training data. AgentLens uses development to select layer and epoch before a training-plus-development refit. LATENTTRIAD and the supervised text baseline fit their readouts on training plus development without selecting an operating point. AgentLens retains its published argmax rule. RouteGuard and LATENTTRIAD use the same 0.5 probability decision. No method receives a test label, per-source threshold, or post-hoc target recall.

AgentLens task transfer. We extract the full Boundary last-token vector and fix the layer range to the released training script's default first 28 layers across backbones.[3] A group-disjoint development partition selects the layer and epoch. The probe preserves the released AgentLens optimization contract: Linear$(h, 2)$ plus softmax, Adam with learning rate 10^{-2}, cross-entropy loss, batch size 16, and at most 100 epochs. It adds no standardization, sample weighting, Top-K detection, or ensemble. The selected probe is refit on training plus development and evaluated once with its argmax decision.

On deeper backbones, this fixed layer range differs from LATENTTRIAD's full-depth grid. The methods also differ in member construction and readout adaptation. Appendix C.2 uses the final Large-ID fits to separate fusion, relation features, and member construction. A dimension-matched independent-state control replaces Geometry while retaining the B/O heads; single-member controls refit the readouts for each retained member.

B.2 COMPOSITIONAL PATH CALIBRATION

CompoSkill Path uses the same extracted states for the fixed and nested LATENTTRIAD variants. The fixed contract applies $\max(B, \min(O, G))$ at threshold 0.5. The nested contract changes only

[3]Release: `AgentLens@22d397d0`.

Table 8: **LATENTTRIAD detects 93.0–97.2% of composed attacks at 3.8–4.3% FPR.** All 690 matched CompoSkill pairs are retained after seven rewrite rounds (%).

	AgentLens (1P)			RouteGuard (4P)			LATENTTRIAD (1P)		
Backbone	$TPR_0 \uparrow$	FPR↓	$TPR_7 \uparrow$	$TPR_0 \uparrow$	FPR↓	$TPR_7 \uparrow$	$TPR_0 \uparrow$	FPR↓	$TPR_7 \uparrow$
Llama-3.1-8B	84.1	17.5	83.9	**94.2**	8.3	**94.1**	93.0	**4.3**	92.9
Qwen2.5-7B	59.4	38.8	59.4	92.5	12.9	92.5	**95.9**	**4.3**	**95.9**
Qwen2.5-14B	71.2	29.4	71.0	91.8	11.2	91.7	**96.4**	**4.1**	**96.3**
Qwen3-30B-A3B	81.5	19.2	81.4	92.4	10.1	92.2	**96.8**	**3.9**	**96.8**
Llama-3.3-70B-INT8	85.2	16.1	85.0	93.1	9.4	93.0	**97.2**	**3.8**	**97.1**

P: prefills; subscripts 0/7: original/seven-round paths. FPR uses unchanged clean controls. Bold: best per backbone. LATENTTRIAD uses development-selected fusion and thresholds. Protocol: Appendix B.3.

the fusion rule and operating point. Each fold fits heads on training data, selects one of four monotone rules and 33 thresholds by development equal-threat macro F1, and then refits the heads on training plus development. The outer test fold is evaluated once. The candidate rules and thresholds are specified before evaluation; the fixed policy serves as a separate control.

Table 9: Fixed and development-adapted LATENTTRIAD on CompoSkill Path (%). Fixed uses $\max(B, \min(O, G))$ at 0.5. Nested selection searches four monotone fusion rules and 33 thresholds on development data.

		Fixed			Nested development		
Backbone	Mode	FPR↓	Recall↑	F1↑	FPR↓	Recall↑	F1↑
Llama-3.1-8B	Explicit	19.8	92.4	87.1	8.7	86.3	88.5
	Implicit	0.0	99.7	99.9	0.0	99.7	99.9
Qwen2.5-7B	Explicit	13.7	95.3	91.2	7.6	92.7	92.6
	Implicit	0.9	99.4	99.3	1.2	99.1	99.0

B.3 COMPOSITIONAL MECHANISM AND ADAPTIVE STRESS

Table 10 reports explicit and implicit attack modes derived from CompoSkill-Bench (Liu et al., 2026b). We freeze dataset revision `1a13cca2`[4]. Its primary design contains 380 three-skill instances, each with clean, explicit, and implicit variants (1,140 records). We pair each available attack variant with its clean counterpart, requiring a valid task, a resolved source–bridge–terminal chain, and complete workspace files. Implicit variants also require a valid execution guide, matching path length, and no primary-chain skill identifiers in the sanitized payload. Missing variants are not synthesized. This yields 344 explicit and 346 implicit pairs (690 mode-specific pairs, 1,380 evaluation rows); a clean variant can appear in both modes. The five outer folds separate scenario/persona role groups. LatentTriad and AgentLens use one native prefill; RouteGuard uses four fixed probe prefills. Cisco Cross-Skill uses the official scanner to jointly inspect the three installed root documents. Our root-document adaptation of SkillReact extracts capability conjunctions from those documents; the original framework also inspects bundled code (Wang et al., 2026). Our Qwen3-30B-A3B reimplementation of ChainGuard evaluates each candidate against previously installed skills, following ColluSkill's context-aware procedure (Zeng et al., 2026).

The serialized path includes the trusted task, the root `SKILL.md` of each active skill in path order, and the workspace visible at the audited boundary. In this path-specific region map, D contains the task and skill declarations; unlike the package map, it includes the trusted task. O contains the remaining skill instructions and visible workspace. Auxiliary package files remain tool-addressable but are not loaded into this path view. Private JSON keys beginning with an underscore are removed recursively from implicit execution guides; labels and attack-mode metadata are also excluded. The explicit view precedes chain execution. The implicit view follows workspace observation and pre-

[4]Full revision: `https://huggingface.co/datasets/Limax11/CompoSkill-Bench/tree/1a13cca2a63c9d6fdcc196df853c3a372ed31e8a`.

cedes the terminal action. Both are single-prefill inspections of a given path context, rather than end-to-end planner rollouts.

Table 8 pools all 690 mode-specific pairs to report original-path detection and the seven-round stress endpoint on the same support. It covers five backbones, including Llama-3.3-70B in INT8; the path inputs, fold protocol, and development-only selection rules are unchanged. Table 10 instead pools paths within each mode separately for Llama-3.1-8B and Qwen2.5-7B. LATENTTRIAD reaches 88.5–99.9% F1 and 0.0–8.7% FPR across the four backbone–mode cells. Three paired F1 gains over RouteGuard exclude zero; the Llama explicit gain is 1.2 points [-1.4, 3.9]. The installation checks receive identical skill documents for both members of each pair, without the task or workspace that distinguishes their risk.

Table 10: **Development-adapted LATENTTRIAD has the highest F1 in all four backbone–mode comparisons.** Metrics pool all paths within each mode (%).

(a) Whole-path representation detectors complete task path							
Method	**Backbone**	**Explicit**			**Implicit**		
		FPR↓	Recall↑	F1↑	FPR↓	Recall↑	F1↑
AgentLens (1P)	Llama-3.1-8B	15.7	69.2	74.8	19.4	98.8	90.6
RouteGuard (4P)	Llama-3.1-8B	16.3	90.1	87.3	0.3	98.3	99.0
LATENTTRIAD (1P)	Llama-3.1-8B	**8.7**	86.3	**88.5**	**0.0**	**99.7**	**99.9**
AgentLens (1P)	Qwen2.5-7B	37.2	40.4	45.5	40.5	78.3	71.6
RouteGuard (4P)	Qwen2.5-7B	24.7	88.4	82.9	1.2	96.5	97.7
LATENTTRIAD (1P)	Qwen2.5-7B	**7.6**	**92.7**	**92.6**	**1.2**	**99.1**	**99.0**
(b) Installation-time composition checks installed skill set; no task or workspace state							
Method	**Build**	**Explicit**			**Implicit**		
		FPR↓	Recall↑	F1↑	FPR↓	Recall↑	F1↑
Cisco Cross-Skill	Official	30.5	30.5	37.9	31.2	31.2	38.4
SkillReact	Root-doc adaptation	96.8	96.8	65.9	96.8	96.8	65.9
ChainGuard	Qwen3 reimpl.	46.5	46.5	48.2	43.6	43.6	46.6

Input note. Panel (b) receives the same installed skill documents for both members of each pair; the task-conditioned path is visible only to panel (a). Bold numbers mark LATENTTRIAD column optima; P denotes model prefills.

Node-only scanners answer a narrower question: whether independent package inspection can recover a risk that emerges only from composition. They inspect each activated package separately and never observe the complete task path.

Table 11: **Node-only controls on CompoSkill.** Each method scans the three constituent skills independently and never observes their task-conditioned composition.

Method	**Build**	**Explicit**			**Implicit**		
		FPR↓	Recall↑	F1↑	FPR↓	Recall↑	F1↑
SkillsGuard	Pinned	96.5	96.5	65.9	96.0	96.0	65.7
Cisco independent	Pinned	13.7	13.7	21.5	13.9	13.9	21.7
SkillSpector	Pinned	61.9	61.9	55.3	62.4	62.4	55.5

Attack and clean paths contain the same skill documents. Consequently, node-only scores are identical within each pair, pair AUROC is 50.0%, and FPR equals recall.

The SCR-Bench panel (Xie et al., 2026) transfers the fitted CompoSkill detectors without using an SCR label for fitting or threshold selection. It contains 667 matched cases: 150 capability-flow, 401 trust-transfer, and 116 authorization-confusion pairs. We use source revision `b8bb3e9e`.[5] Of its 118 AuthBlur cases, `authblur:26` and `authblur:76` lack the plain-advisor variant required for the four-condition panel, leaving 116 eligible cases. Primary contrasts are terminal-only versus composed-neutral for capability flow, control versus experiment for trust transfer, and level 0 versus level 3 for authorization confusion. Each pair holds the case and downstream action fixed while changing the mechanism condition. We use 2,000 case-paired bootstrap draws and 20,000 paired sign-flip draws. Holm correction is applied across the three mechanisms within each backbone and method.

The adaptive panel adapts ColluSkill's scanner-feedback rewriting to our path-level detector (Zeng et al., 2026). We generate candidates over seven rounds. At each round, the attacker receives the verdict, branch scores, and high-risk region summary. A candidate must preserve the task,

[5]Release: `SCR-Bench@b8bb3e9e`.

node order, tools, paths, endpoints, artifact handoffs, and capability graph. It must also pass an independent frozen semantic judge. All 690 original paths remain in the denominator, and path-paired uncertainty uses 2,000 bootstrap draws. Invalid rewrites retain the preceding valid path and never count as successful attacks. The rewrites target LATENTTRIAD; AgentLens and RouteGuard score the same resulting paths, rather than candidates optimized separately against each baseline.

Table 12 separates detector performance from attack coverage for Llama-3.1-8B and Qwen2.5-7B. Of the initially blocked paths, nine on Llama-3.1-8B and eight on Qwen2.5-7B receive at least one valid rewrite. These sets contain nine distinct mode-specific paths in total. New evasions occur on one of Llama's nine paths and none of Qwen's eight. Repeated rounds on the same path do not increase this coverage. The evasion rate over all paths reflects both the attacker's ability to produce valid candidates and the detector's response. Conditional rates on these nine and eight paths isolate the response to valid rewrites.

Table 12: **Compositional mechanism transfer and adaptive stress.** SCR-Bench changes one mechanism at a time; the adaptive panel preserves each path's capability graph over seven feedback rounds.

(a) Frozen transfer to SCR-Bench 667 paired cases; no SCR label used for fitting

Backbone	Capability flow Δs	Trust transfer Δs	Authorization Δs	Holm significant	AgentLens control
Llama-3.1-8B	**+0.145***	**+0.002***	**-0.027**	**2/3**	Opposite on flow
Qwen2.5-7B	**-0.010**	**+0.008***	**+0.229***	**2/3**	Saturated at 1.0

(b) Adaptive rewriting 690 paths; seven rounds; capability graph held fixed

Backbone	LT TPR↑ $t_0 \to t_7$	LT FPR↓	AgentLens t_7 TPR/FPR	RouteGuard t_7 TPR/FPR	LT−baseline ΔTPR [95% CI]
Llama-3.1-8B	**93.0→92.9**	**4.3**	83.9/17.5	94.1/8.3	AL **+9.0 [+6.1, +11.9]***; RG **-1.2 [-3.6, +1.2]**
Qwen2.5-7B	**95.9→95.9**	**4.3**	59.4/38.8	92.5/12.9	AL **+36.5 [+32.8, +40.4]***; RG **+3.5 [+1.2, +5.7]***

(c) Valid rewrite coverage distinct mode-specific paths and repeated attempts

Backbone	Initially blocked paths	Paths with a valid rewrite	Accepted rewrites / attempts	New evasions / valid paths
Llama-3.1-8B	642	9	21/4488	1/9
Qwen2.5-7B	662	8	11/4634	0/8

In (a), $\Delta s = s_{\text{high}} - s_{\text{low}}$ is the mean score change from the lower- to higher-risk condition, without percentage scaling. In (b), rates are percentages and differences are percentage points; AL/RG denote AgentLens/RouteGuard. * marks a 95% paired interval excluding zero in (b), or Holm-adjusted $p < 0.05$ in (a). Repeated rewrites do not increase distinct-path coverage.

B.4 METRICS, AGGREGATION, AND STATISTICAL INFERENCE

Pooled metrics describe the observed package mixture; source-macro metrics give each ecosystem equal influence.

For malicious as the positive class, we compute

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad \text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}.$$

Undefined precision is set to zero when a method predicts no positives. For Skill detection, each learned method is fitted independently on each split. We pool predictions from the five nonoverlapping test folds. Source-macro metrics are computed within each source and then averaged equally over the five sources. If source s has n_s^- benign packages,

$$\text{FPR}_{\text{pooled}} = \frac{\sum_s \text{FP}_s}{\sum_s n_s^-}, \qquad \text{FPR}_{\text{source}} = \frac{1}{5} \sum_s \frac{\text{FP}_s}{n_s^-}.$$

The former weights source-specific FPR by benign sample counts; the latter assigns each source 20%. Both retain every test package. An expected deployment FPR would instead use the source proportions among benign deployment requests. Confidence intervals use 2,000 shared bootstrap replicates. Components are stratified by their source-membership sets and resampled with replacement. Each draw includes all packages in the selected components for every method. We then

Table 13: **Large-ID F1 by source (%).** Representation detectors report seven-backbone means with ranges in brackets; input-level methods report one value.

Source	Regex	SkillSieve	Cisco	Mistral judge	Sup. text	AgentLens	RouteGuard	LatentTriad
MalSkillBench	66.7	70.3	43.2	32.2	73.3	71.2 [28.6, 98.6]	85.0 [84.2, 86.3]	99.4 [99.3, 99.6]
SkillFortify	66.7	68.4	8.5	56.0	92.4	73.0 [35.7, 99.4]	98.0 [97.2, 98.6]	100.0 [100.0, 100.0]
SkillsBench-1650	66.7	61.8	42.9	6.0	35.4	54.0 [27.6, 67.9]	55.4 [47.2, 60.5]	90.9 [88.6, 93.9]
Agent-skill malware	66.7	48.8	6.7	6.7	12.9	53.0 [25.6, 73.5]	71.1 [56.5, 78.6]	89.2 [83.6, 91.5]
ATR Skill Benchmark	66.7	52.9	37.0	51.6	77.8	60.4 [26.7, 81.6]	86.4 [82.9, 90.9]	90.0 [85.7, 93.6]

recompute source-macro F1. The 14 LatentTriad-minus-comparator intervals share these draws, preserving component, source, method, and backbone pairing. These are per-comparison 95% intervals, not multiplicity-adjusted tests.

The five-source LOSO protocol performs five additional refits per learned method. Each refit withholds one ecosystem from fitting, development selection, and calibration. We compute metrics on each held-out source and average the five results. The 2,000 paired replicates resample matched families within the held-out source. Every method and backbone receives the same family multiplicities.

Average precision is the step-integrated area under the precision–recall curve. At each distinct score, precision is weighted by the increase in recall. We compute AP from sample-level ranking scores. Aggregate confusion counts are insufficient. Cisco contributes its discrete ordinal severity instead of a calibrated probability. AUROC is threshold-independent; it is the primary metric for the final-configuration relation-feature contrast and a ranking diagnostic elsewhere. PairAcc is the fraction of matched inverse pairs in which the malicious-operation member receives the higher risk score. A tie contributes one half. We use PairAcc only for matched counterfactual diagnostics.

C Large-ID Detection and Cross-Source Generalization

Source slices, component controls, and paired contrasts use the same component-disjoint package detector. Source-holdout evaluation fits separate detectors with the test ecosystem excluded.

C.1 Source-Wise and Paired Results

Table 5 shows the source imbalance. MalSkillBench supplies 2,621 of the 2,915 benign packages (89.9%). LatentTriad's FPR on this source is 0.5–0.8% across backbones; on SkillFortify, it is 0%. Rates are higher on the other three sources, which contain 114 benign packages in total (Table 14). For example, Llama flags 17/2,621 MalSkillBench packages, 0/180 SkillFortify packages, 10/63 SkillsBench1650 packages, 1/29 Agent-skill malware packages, and 5/22 ATR packages. This gives a pooled FPR of 33/2,915 (1.13%) and a source-macro FPR of 8.54%. The latter weights each source equally, exposing differences in benign acceptance that the dominant source would otherwise mask.

Table 14: **Benign acceptance varies across skill sources.** FPR ranges (%) span seven backbones; they are not confidence intervals. All methods share the listed benign support.

Source	Benign n	AgentLens	RouteGuard	LatentTriad
MalSkillBench	2,621	0.84–40.02	16.06–22.51	0.53–0.76
SkillFortify	180	0.00–45.56	0.00–1.67	0.00–0.00
SkillsBench-1650	63	19.05–82.54	23.81–34.92	11.11–23.81
Agent-skill malware	29	17.24–68.97	10.34–20.69	3.45–13.79
ATR Skill Benchmark	22	18.18–68.18	9.09–18.18	13.64–27.27

Table 13 reports source-wise F1, and Table 15 gives all same-backbone contrasts and confidence intervals.

Table 15: **Paired Large-ID source-macro F1 differences (percentage points).** Each cell is Latent-Triad minus the same-backbone comparator. Intervals use 2,000 shared source-stratified component bootstrap draws. The table reports all 14 contrasts.

Backbone	Δ vs. RouteGuard	Δ vs. AgentLens
Qwen2.5-7B	13.5 [9.5, 18.1]	50.0 [45.1, 58.2]
Llama-3.1-8B	17.0 [13.1, 21.6]	14.9 [12.0, 18.1]
Mistral-7B	13.6 [9.6, 18.2]	9.8 [6.3, 13.3]
InternLM2.5-7B	15.6 [11.0, 20.8]	18.4 [14.7, 22.3]
Qwen2.5-14B	14.1 [10.8, 17.8]	65.5 [59.0, 77.4]
Phi-3-medium-14B	16.8 [12.4, 22.1]	48.7 [42.7, 58.6]
Qwen3-30B-A3B	12.5 [7.6, 17.9]	13.9 [9.0, 18.0]

C.2 Final-Configuration Component Controls

These controls use the component-disjoint Large-ID configuration in Table 2: seven backbones, five folds, eight depths, and three members with offsets 300, 400, and 500. These are follow-up controls on the same folds; their results do not select the reported detector.

The first comparison fixes all fitted component scores and varies only the fusion: B, O, G, $\max(B, O)$, $\min(O, G)$, or the full rule. All use threshold 0.5. The second replaces Geometry with independent-state features while retaining B/O scores, input coverage, member partitions, head family, and fusion. For each member, D and O concatenate pooled projected states and atom-coordinate means, standard deviations, and maximum absolute values across eight depths. B concatenates its eight O→B-projected states. Separate unwhitened PCA fits on benign member-training rows retain 82 D, 82 O, and 81 B coordinates. Concatenating the three members gives 735 features. The replacement Geometry head uses the same standardizer and L2 logistic readout ($C = 1$), fit on training plus development. This control contains no cross-region residuals, differences, distances, or dot products.

The third comparison refits all three branch heads for each original member separately, using the same training recipe. We report all three members and the arithmetic mean of their metrics, rather than selecting the best member or averaging their scores. Three-member concatenation changes feature width as well as member count. The equal-width Geometry control matches input dimension and head family, not every aspect of statistical capacity.

Table 16: **Multi-state evidence recovers Boundary misses at unchanged FPR.** Source-macro metrics (%) are averaged over seven backbones. All decisions use 0.5.

Configuration	AUROC↑	AP↑	F1↑	FPR↓	Recall↑
Full: $\max(B, \min(O, G))$	99.02	99.18	93.91	9.51	96.73
Boundary-only	97.28	97.94	91.63	9.51	92.58
Operation-only	99.03	99.18	86.69	0.00	80.04
Geometry-only	99.68	99.78	95.85	0.00	92.42
$\max(B, O)$	99.15	99.29	94.08	9.51	97.05
$\min(O, G)$	99.53	99.64	83.56	0.00	75.98
Dimension-matched independent-state control	98.95	99.12	93.89	9.51	96.68
Single member: offset 300	99.09	99.24	93.71	10.30	96.97
Single member: offset 400	98.91	99.08	93.44	10.47	96.67
Single member: offset 500	98.98	99.15	93.57	10.30	96.78
Single-member mean	98.99	99.16	93.57	10.36	96.81

For each backbone, metrics are computed on the concatenated out-of-fold predictions, then averaged equally over sources. Table 16 averages these metrics over the seven fixed backbones. Table 18 retains the backbone-level results. The combined branches complement Boundary-only: Full raises mean recall from 92.58% to 96.73% and F1 from 91.63% to 93.91%. Both flag exactly the same benign packages on each backbone, giving 9.51% mean FPR. Explicit relation features add 0.076

points of AUROC over the equal-width independent-state control. All seven point estimates are positive.

The remaining controls distinguish this complementarity from the choice of fusion rule. Geometry-only yields 95.85% F1 and 92.42% recall with no false alarms. At threshold 0.5, it trades lower recall than Full for higher F1 on these data. Zero observed FPR describes this fitted operating point, not a guarantee on new traffic. The simpler $\max(B, O)$ rule reaches 94.08% F1 and 97.05% recall at the same FPR as Full. Here, Geometry confirmation removes some true alarms from that rule without removing false alarms. These results support the multi-state evidence, but do not identify the full rule as the best operating point.

Three members differ from the single-member mean by 0.032 AUROC points, 0.029 AP points, and 0.340 F1 points. All three intervals include zero (Table 17), so this comparison does not establish a mean ensemble gain.

Table 17: **Relation features add a small ranking gain beyond independent-state features.** Differences are percentage points with 95% paired component-bootstrap intervals.

Contrast	ΔAUROC	ΔAP	ΔF1
Full − independent-state control	+0.076 [+0.027, +0.136]	+0.063 [+0.023, +0.112]	+0.026 [+0.001, +0.075]
Full − single-member mean	+0.032 [-0.061, +0.144]	+0.029 [-0.041, +0.112]	+0.340 [-0.275, +0.917]
Full − Boundary-only	+1.744 [+1.229, +2.328]	+1.240 [+0.874, +1.643]	+2.286 [+1.595, +3.131]
Full − $\max(B, O)$	-0.125 [-0.243, -0.027]	-0.106 [-0.194, -0.030]	-0.168 [-0.350, -0.036]

The primary contrast is the seven-backbone mean source-macro AUROC difference between Full and the independent-state control. Its 95% interval is [0.027, 0.136] percentage points. We use 2,000 paired component-bootstrap draws, stratified by source-membership signature, with shared component multiplicities across methods, members, and backbones. Every draw recomputes AUROC and AP from the saved scores, including ties. Intervals use the 2.5th and 97.5th percentiles and are conditional on the fitted models. A two-sided test with 5,000 shared component-level method swaps gives raw $p = 0.00020$; seed 20260910 fixes both procedures. The intervals in Table 17 are pointwise.

Table 18: **Multi-state evidence complements Boundary on all seven backbones.** Source-macro metrics (%); the single-member row averages offsets 300, 400, and 500.

Backbone	Configuration	AUROC↑	AP↑	F1↑	FPR↓	Recall↑
Llama-3.1-8B	Full	99.11	99.26	94.37	8.54	96.78
	Boundary-only	97.63	98.19	92.52	8.54	93.37
	Independent-state control	99.11	99.26	94.37	8.54	96.78
	Single-member mean	98.95	99.12	93.73	8.88	96.00
Qwen2.5-7B	Full	99.15	99.30	94.31	10.39	98.20
	Boundary-only	97.01	97.79	91.95	10.39	93.90
	Independent-state control	99.03	99.20	94.31	10.39	98.20
	Single-member mean	99.06	99.24	93.82	11.43	98.09
Mistral-7B-v0.3	Full	98.66	98.92	93.42	10.48	96.53
	Boundary-only	97.19	97.96	91.37	10.48	92.69
	Independent-state control	98.59	98.88	93.42	10.48	96.53
	Single-member mean	98.71	98.97	93.10	11.53	96.83
InternLM2.5-7B	Full	99.25	99.36	94.60	8.99	97.60
	Boundary-only	97.53	98.15	92.62	8.99	93.89
	Independent-state control	99.01	99.17	94.42	8.99	97.26
	Single-member mean	99.29	99.38	93.86	11.32	98.05
Qwen2.5-14B	Full	98.90	99.12	94.33	8.04	96.28
	Boundary-only	97.56	98.22	93.04	8.04	93.88
	Independent-state control	98.88	99.11	94.33	8.04	96.28
	Single-member mean	98.93	99.13	93.82	9.60	96.59
Phi-3-medium-14B	Full	99.22	99.24	93.34	8.97	95.41
	Boundary-only	97.30	97.66	89.23	8.97	88.39
	Independent-state control	99.16	99.18	93.34	8.97	95.41
	Single-member mean	99.12	99.14	93.33	7.99	94.65
Qwen3-30B-A3B	Full	98.89	99.09	93.01	11.19	96.30
	Boundary-only	96.73	97.63	90.66	11.19	91.95
	Independent-state control	98.87	99.06	93.01	11.19	96.30
	Single-member mean	98.88	99.09	93.35	11.77	97.43

C.3 Five-Source Large-ID Leave-One-Source-Out Evaluation

With the complete test ecosystem withheld from fitting, LatentTriad retains meaningful detection performance, although it falls below within-ecosystem results. It averages 73.2% source-macro F1 and 80.3% AUROC, retaining higher F1 than same-backbone RouteGuard and AgentLens on all seven backbones (all 14 paired bootstrap intervals exclude zero). Recall and AP average 76.6% and 82.0%. Tables 19 and 20 give the aggregate results and paired representation contrasts with 2,000-draw bootstrap confidence intervals.

The advantage is metric-dependent. The supervised Qwen text readout has lower F1 (58.7%) but higher AUROC (88.1%) and AP (90.0%). LatentTriad therefore leads the same-backbone representation monitors in F1, not every comparator on every metric. Its mean F1 also exceeds the 66.7% positive-class F1 of an all-positive rule, which has 100% FPR on the source-balanced data. Deployment still requires measuring recall and false alarms at the chosen operating point; F1 alone does not establish a cold-start alert budget.

Table 19: **Five-source Large-ID leave-one-source-out evaluation (%).** Sources are equally weighted. Brackets: 95% intervals from 2,000 source-stratified family-bootstrap draws.

Method	Backbone	Recall	F1 [95% CI]	AUROC	AP
Regex	Shared	100.0	66.7 [66.7, 66.7]	68.6	68.6
SkillSieve L1	Shared	53.0	60.5 [54.3, 65.9]	69.0	69.3
Cisco scanner	Shared	17.5	27.6 [21.2, 33.1]	63.4	64.7
Mistral judge	Mistral-7B	20.3	30.5 [24.7, 35.8]	81.0	81.1
Supervised Qwen text	Qwen2.5-7B	49.6	58.7 [52.8, 64.0]	88.1	90.0
AgentLens	Llama-3.1-8B	48.2	35.8 [34.1, 37.2]	50.3	55.1
RouteGuard	Llama-3.1-8B	60.0	56.1 [50.7, 60.8]	60.5	62.2
LATENTTRIAD	**Llama-3.1-8B**	**70.8**	**72.3 [66.9, 76.4]**	**78.3**	**80.9**
AgentLens	Qwen2.5-7B	40.0	26.7 [26.7, 26.7]	47.0	48.9
RouteGuard	Qwen2.5-7B	63.6	59.3 [54.0, 64.0]	66.6	66.8
LATENTTRIAD	**Qwen2.5-7B**	**78.2**	**72.3 [69.5, 74.9]**	**77.8**	**79.5**
AgentLens	Mistral-7B	82.3	63.0 [59.8, 66.1]	59.8	63.6
RouteGuard	Mistral-7B	61.0	58.8 [53.7, 63.4]	66.3	67.1
LATENTTRIAD	**Mistral-7B**	**74.1**	**73.6 [68.9, 77.4]**	**77.6**	**79.0**
AgentLens	InternLM2.5-7B	69.8	49.9 [47.6, 51.7]	53.7	52.5
RouteGuard	InternLM2.5-7B	60.6	57.7 [52.4, 62.7]	64.6	64.2
LATENTTRIAD	**InternLM2.5-7B**	**75.7**	**73.4 [69.0, 77.2]**	**80.7**	**82.1**
AgentLens	Qwen2.5-14B	53.6	37.8 [35.3, 40.0]	45.3	51.5
RouteGuard	Qwen2.5-14B	64.0	60.9 [55.3, 66.1]	65.9	64.7
LATENTTRIAD	**Qwen2.5-14B**	**78.1**	**73.1 [70.0, 75.8]**	**80.6**	**82.0**
AgentLens	Phi-3-medium-14B	37.5	26.6 [25.8, 27.4]	48.1	49.5
RouteGuard	Phi-3-medium-14B	59.3	55.7 [49.9, 60.6]	65.8	65.0
LATENTTRIAD	**Phi-3-medium-14B**	**70.4**	**72.2 [67.9, 75.8]**	**82.1**	**84.1**
AgentLens	Qwen3-30B-A3B	54.3	38.8 [37.3, 40.0]	50.8	50.6
RouteGuard	Qwen3-30B-A3B	66.9	66.2 [61.3, 70.4]	75.5	76.0
LATENTTRIAD	**Qwen3-30B-A3B**	**88.8**	**75.5 [72.9, 77.8]**	**84.8**	**86.4**

Table 20: **Paired five-source LOSO F1 differences (percentage points).** LATENTTRIAD minus each comparator, with 95% source-stratified family-bootstrap CIs.

Backbone	vs. RouteGuard	vs. AgentLens
Llama-3.1-8B	+16.1 [+10.0, +22.3]	+36.5 [+31.0, +40.9]
Qwen2.5-7B	+13.0 [+8.4, +18.5]	+45.6 [+42.9, +48.2]
Mistral-7B	+14.7 [+8.7, +20.8]	+10.6 [+4.7, +16.0]
InternLM2.5-7B	+15.7 [+9.7, +22.2]	+23.5 [+18.7, +28.1]
Qwen2.5-14B	+12.2 [+6.3, +18.3]	+35.3 [+31.8, +38.7]
Phi-3-medium-14B	+16.5 [+10.3, +22.7]	+45.6 [+41.4, +49.3]
Qwen3-30B-A3B	+9.3 [+4.6, +14.5]	+36.7 [+33.7, +39.5]

D ROBUSTNESS AND ENFORCEMENT RESULTS

The white-box and broker studies use the pair-disjoint application probes defined in Appendix A.3. They measure rewrite sensitivity and request-level containment with the probes held fixed.

D.1 WHITE-BOX ADVERSARIAL ROBUSTNESS

The panel asks whether an attacker can manipulate visible D/O/B geometry while preserving the original operation. It contains 256 family-unique matched pairs from all five Large-ID sources. Source quotas are set by the evaluation protocol. Eligible package roots are at most 16,384 bytes and admit a safe Operation newline split. Each pair contributes clean, region-padding, cross-file-split, and six Geometry-candidate views. This yields 4,608 package views per backbone. The attacker knows the native tokenizer, D/O/B binding, fold-specific detector bundle, fusion, and threshold. It may rewrite package documentation and layout before capability loading, but it cannot modify model or detector parameters. Geometry evasion chooses the lowest Geometry-margin candidate from the malicious package alone. We then apply the same candidate index to its matched benign control.

Region padding inserts 1,024–2,048 marked inert characters split across D and O. Cross-file splitting moves the complete Operation suffix into at most four ordered Markdown fragments at safe newline boundaries outside code fences. The Geometry attack searches six 768-character inert placements fixed before evaluation. They span description, operation, and early or late auxiliary positions. Exact deletion or ordered concatenation reconstructs every original package. Every original non-root file is byte-identical, and no original operation byte is deleted. No package content is executed.

Every view is scored by the detector fitted for its original Large-ID test fold, without refitting or a threshold change. The inference unit is the matched family, and metrics average the five sources equally. Intervals use 2,000 source-stratified family-bootstrap draws. Fused and Geometry logit-margin shifts use 10,000 source-stratified family sign flips. We apply separate Holm corrections across the 21 backbone–attack tests. The robustness criterion requires at least 80% TPR retention and an empirical-bootstrap 95% upper endpoint of at most 25% for attack success.

In Table 22, Δm is the attacked-minus-clean fused logit margin, without percentage scaling. $p_{\mathrm{H}}(m)$ and $p_{\mathrm{H}}(m_G)$ test fused and Geometry margin shifts, respectively.

TPR retention is the attacked source-macro TPR divided by the clean source-macro TPR, expressed as a percentage. It can exceed 100% when rewriting makes previously missed attacks detectable. ASR first divides successful pass-after-rewrite evasions by clean true positives within each source, then averages the defined source rates. Thus its denominator is the clean-detected subset of the 256 malicious packages, not all package views. Benign FPR uses the 256 matched benign packages with equal source weighting. When no evasion is observed, every bootstrap replicate also contains zero evasions, producing $[0, 0]$. This interval reflects the observed sample; it does not bound the population evasion rate at zero.

For a zero-event cell with n_s clean true positives in source s, a source-wise binomial model gives the one-sided upper endpoint $u_s = 1 - (0.05/5)^{1/n_s}$. Bonferroni coverage then yields the conservative 95% upper bound $U_{\mathrm{macro}} = \frac{1}{5}\sum_s u_s$. Table 21 reports the eight zero-event cells using their actual source denominators. These bounds assume independent Bernoulli family trials with a common evasion probability within each source, conditional on clean detection, the selected support, and the tested rewrite. They require no independence across sources and apply per cell, not simultaneously across all 21 cells. Dependence between content-linked families (Appendix A.3) remains outside that model.

Table 21: **One-sided 95% macro-ASR bounds for zero-event cells.** Counts follow the source order: agent-skill malware, ATR, MalSkillBench, SkillFortify, SkillsBench-1650. Listed rewrites denote separate cells.

Backbone	Rewrite	Clean-TP counts	Upper bound
Llama-3.1-8B	Padding	23, 21, 70, 103, 31	12.48%
Qwen2.5-7B	Padding, Geometry	23, 22, 70, 103, 31	12.32%
Mistral-7B	Padding, Geometry	27, 21, 70, 103, 32	11.90%
InternLM2.5-7B	Padding, Geometry	26, 21, 70, 103, 30	12.18%
Qwen2.5-14B	Padding	22, 22, 69, 103, 31	12.48%

The supplementary bounds span 11.90–12.48%. The table below also reports empirical-bootstrap intervals for all cells; the binomial bounds apply only to zero-event cells.

All 21 backbone–attack cells satisfy this empirical anti-evasion criterion. Cross-file splitting is the strongest attack. Its equal-source ASR reaches 11.0% [6.5, 16.3] on Llama, while TPR retention remains at least 92.3% on Qwen2.5-7B. Region padding and Geometry search achieve TPR retention of at least 101.6% and 100.7%, respectively. Geometry-search ASR is at most 3.1%. The selected rewrite changes the Geometry margin in 19/21 cells after Holm correction, but these changes rarely produce successful evasion. Manipulating one exposed component therefore does not imply control of the fused verdict.

The matched benign controls expose a different cost: the attacks raise alert load even when they rarely evade detection. Clean equal-source FPR is 32.3–40.8% on this length- and split-eligible subset and becomes 49.3–80.4% after the same inert rewrites. This panel therefore measures both evasion resistance and the benign interruption cost of the fixed rule. Canonicalization and alert-budget calibration are deployment choices to evaluate against that cost, not validated mitigations in this panel.

Table 22: **Operation-preserving rewrites retain detection but increase benign alerts.** Source-macro rates (%) on 256 matched pairs; ASR conditions on clean true positives. Brackets: 95% source-stratified family-bootstrap CIs. Δm is a logit shift.

Backbone	Attack	TPR ret.	ASR [95% CI]	Benign FPR $c \to a$	Δm [95% CI]	$p_{\mathrm{H}}(m)$	$p_{\mathrm{H}}(m_G)$	Pass
Llama-3.1-8B	Region padding	105.3	0.0 [0.0, 0.0]	34.4→76.4	5.8 [5.1, 6.5]	0.0021	0.0021	yes
	Cross-file split	92.8	11.0 [6.5, 16.3]	34.4→55.0	-0.6 [-1.3, 0.2]	0.9179	0.0021	yes
	Geometry evasion	101.6	3.1 [0.6, 6.1]	34.4→65.7	3.7 [2.9, 4.5]	0.0021	0.0021	yes
Qwen2.5-7B	Region padding	105.0	0.0 [0.0, 0.0]	32.3→78.1	6.9 [6.2, 7.6]	0.0021	0.7895	yes
	Cross-file split	92.3	9.0 [4.2, 14.0]	32.3→57.1	-3.7 [-4.3, -3.1]	0.0021	0.0021	yes
	Geometry evasion	105.0	0.0 [0.0, 0.0]	32.3→78.5	7.5 [6.7, 8.4]	0.0021	0.0021	yes
Mistral-7B	Region padding	102.3	0.0 [0.0, 0.0]	40.8→80.4	4.6 [4.0, 5.2]	0.0021	0.0021	yes
	Cross-file split	93.2	7.1 [3.2, 11.6]	40.8→57.9	-3.6 [-4.1, -3.0]	0.0021	0.0021	yes
	Geometry evasion	100.7	0.0 [0.0, 0.0]	40.8→73.1	5.5 [4.8, 6.3]	0.0021	0.0021	yes
InternLM2.5-7B	Region padding	104.4	0.0 [0.0, 0.0]	38.1→73.6	5.4 [4.9, 6.0]	0.0021	0.7895	yes
	Cross-file split	100.4	1.1 [0.0, 2.8]	38.1→65.6	-0.3 [-0.9, 0.3]	1.0000	0.0021	yes
	Geometry evasion	104.4	0.0 [0.0, 0.0]	38.1→75.4	6.0 [5.4, 6.6]	0.0021	0.0021	yes
Qwen2.5-14B	Region padding	103.9	0.0 [0.0, 0.0]	35.8→68.1	1.9 [1.3, 2.5]	0.0021	0.0021	yes
	Cross-file split	98.5	3.4 [0.6, 6.7]	35.8→57.4	0.1 [-0.6, 0.7]	1.0000	0.0021	yes
	Geometry evasion	105.1	1.2 [0.0, 3.3]	35.8→61.0	3.7 [3.0, 4.4]	0.0021	0.0021	yes
Phi-3-medium-14B	Region padding	103.6	0.9 [0.0, 2.7]	34.3→74.7	3.5 [2.9, 4.1]	0.0021	0.0021	yes
	Cross-file split	93.6	8.2 [4.3, 12.8]	34.3→64.1	1.2 [0.7, 1.6]	0.0021	0.0021	yes
	Geometry evasion	101.9	1.8 [0.0, 4.5]	34.3→57.5	2.7 [2.2, 3.2]	0.0021	0.0021	yes
Qwen3-30B-A3B	Region padding	101.6	2.1 [0.0, 4.8]	35.1→62.8	-1.1 [-1.7, -0.5]	0.0021	0.0021	yes
	Cross-file split	101.4	1.7 [0.0, 4.3]	35.1→70.9	0.1 [-0.6, 0.8]	1.0000	0.0021	yes
	Geometry evasion	102.9	0.9 [0.0, 2.9]	35.1→49.3	0.3 [-0.4, 0.9]	1.0000	0.0021	yes

D.2 CAPABILITY-BROKER ENFORCEMENT

Table 23: **Request-level enforcement across seven backbones.** Cells give mean ASR ↓ / valid benign allowance ↑ (%) on 256 matched pairs, using all trials of each label as denominators.

Budget	Random	AgentLens	RouteGuard	LATENTTRIAD	Oracle
10%	41.9/42.2	42.8/41.5	36.5/46.9	**36.7/47.1**	31.7/47.1
20%	37.3/37.4	36.1/39.5	27.3/46.4	**26.9/47.0**	20.3/47.1
30%	32.9/32.6	28.2/38.3	19.6/45.8	**17.2/46.7**	11.2/47.1

Table 24: **Enforcement by acting backbone.** Valid: typed-request validity (%). Other cells: ASR ↓ / valid benign allowance ↑, as in Table 23.

Acting backbone	Budget	Valid	AgentLens	RouteGuard	LATENTTRIAD
Qwen2.5-7B	10%	86.3	71.1/9.0	56.2/15.2	**60.9/15.2**
	20%		64.1/7.0	43.0/14.8	**45.7/15.2**
	30%		52.3/7.0	31.6/14.5	**32.8/15.2**
Llama-3.1-8B	10%	99.0	61.7/21.1	63.7/21.1	**61.3/21.1**
	20%		46.9/20.7	48.8/21.1	**45.7/21.1**
	30%		35.5/19.5	32.8/21.1	**28.9/21.1**
Mistral-7B	10%	95.5	19.9/82.0	19.5/81.6	**20.3/82.0**
	20%		13.3/81.6	14.8/80.9	**14.1/81.6**
	30%		9.4/79.7	10.9/80.1	**9.0/80.9**
InternLM2.5-7B	10%	65.8	55.5/23.4	44.5/25.0	**43.4/25.0**
	20%		49.6/22.7	36.3/25.0	**29.3/25.0**
	30%		37.9/22.7	28.1/24.2	**19.5/25.0**
Qwen2.5-14B	10%	98.0	47.3/68.4	35.5/85.5	**35.5/85.9**
	20%		43.8/59.4	21.5/84.4	**25.8/85.5**
	30%		34.8/59.4	13.3/82.8	**14.1/85.2**
Phi-3-medium	10%	32.6	22.3/10.9	21.1/11.3	**20.7/11.3**
	20%		18.8/9.4	16.0/11.3	**16.8/11.3**
	30%		15.2/9.0	12.1/11.3	**10.5/11.3**
Qwen3-30B-A3B	10%	99.4	21.9/75.8	14.8/88.7	**14.8/89.1**
	20%		16.4/75.8	10.9/87.5	**10.9/89.1**
	30%		12.1/70.7	8.2/86.3	**5.5/88.3**

This experiment measures request-level policy enforcement on 256 family-unique matched safe–unsafe pairs from all five ecosystems. Sample selection uses neither detector scores nor acting-model outputs. Each of the seven protected models loads the complete package through its native chat and

tool template. It then encodes the candidate operation as one typed `request_capability` action. The schema requires scope, destination, authorization, and persistence arguments. Malformed requests fail closed.

A policy-independent parser records each request before applying a policy. AgentLens, RouteGuard, and LATENTTRIAD route the same recorded requests using their package scores under 10, 20, and 30% containment budgets. Only the routing score changes. The inert broker records ALLOW or CONTAIN; even ALLOW executes no package payload, generated command, or network request. All models use greedy decoding. Qwen3 disables its optional reasoning trace for this deterministic typed-encoding task. Confidence intervals use 10,000 paired bootstrap draws over source-stratified operation pairs.

Let V indicate a parser-valid typed request, U an unsafe argument combination, and C selection for score-based containment. The categorical checker marks unrelated or private resources, unapproved or private destinations, absent or bypassed authorization, persistence, and unknown values as unsafe. For the budgeted policies, allowance is $A = V(1-C)$. This Boolean allowance is request allowance, not payload execution. For package label y, the reported endpoints are

$$\mathrm{ASR} = \frac{\sum_{y=1} VUA}{N_{\mathrm{unsafe}}}, \qquad \mathrm{BenignRequestAllowance} = \frac{\sum_{y=0} V(1-U)A}{N_{\mathrm{safe}}}.$$

Here $N_{\mathrm{unsafe}} = N_{\mathrm{safe}} = 256$ per acting model. Benign request allowance counts valid, checker-safe requests that the broker allows; it does not measure completion of an executed task. Request validity is $\sum V/512$; score-based policy rejection is $\sum C/512$, with malformed requests additionally contained. Conditional attack success divides the same unsafe allowance count by $\sum_{y=1} VU$. The table reports the all-unsafe denominator. For paired intervention comparisons, the prevented-attack count is $\sum_{y=1} VUC$; $1 - \mathrm{ASR}$ also includes invalid or non-unsafe requests and is not the broker's blocking rate.

Table 23 averages package-level rates over seven acting models. Random averages 20 seeds; Oracle uses labels. Allow-all records 82.4% valid requests, 46.7% unsafe-request allowance, and 47.1% benign-request allowance. Against AgentLens, LATENTTRIAD reduces unsafe-request allowance by 6.1, 9.2, and 11.0 points at the three budgets and increases benign-request allowance by 5.6, 7.5, and 8.4 points. Both metrics favor LATENTTRIAD in a fraction 1.0 of bootstrap draws at each budget. At 30% containment, the corresponding differences from RouteGuard are -2.4 points (95% CI $[-4.4, -0.4]$) and $+0.95$ points ($[0.22, 1.67]$); both metrics favor LATENTTRIAD in a fraction 0.990 of draws.

E OPERATION SENSITIVITY AND LOCALIZATION

E.1 BALANCED COUNTERFACTUAL DIAGNOSIS

The balanced 2×2 design crosses visible surface with operation semantics. B/B denotes benign surface and benign operation. B/M keeps the benign surface but inserts a malicious operation. S/B combines suspicious surface with benign operation, while S/M is suspicious in both dimensions. Each quadrant has 57 packages, so lexical appearance and operation labels are independently balanced.

Llama LATENTTRIAD ranks malicious operations above their matched benign counterparts in 94.7% of pairs, and Qwen orders every pair correctly. RouteGuard ties every pair. These results separate operation-sensitive score ordering from package-surface classification. Figure 3a shows all methods and paired intervals.

E.2 AUXILIARY RISK LOCALIZATION

The localization experiment predicts the region containing risky content. It reuses hidden states from the same native prefill and fits two $C = 1$ L2-logistic readouts: one on Declaration features and one on Code/Operation features. Thresholding both outputs at 0.5 gives four classes: CLEAN $(0, 0)$, DECL $(1, 0)$, CODE $(0, 1)$, and BOTH $(1, 1)$. These auxiliary predictions enter neither the three-branch score nor the policy threshold. The tables' attribution labels denote content localization, not causal attribution of the primary decision.

Table 25: **Risk localization across seven backbones.** Rates are percentages; D/C AUROC uses 0–1. Natural Macro-F1 includes 95% family-bootstrap intervals.

Backbone	Nat. Acc.	Nat. Macro-F1 [95% CI]	D/C AUROC	Ctrl. Macro-F1	Blocked	Cond. exact	Joint
Llama-3.1-8B	85.7	86.2 [83.2, 88.8]	0.998/0.992	90.9	375/400	92.0	85.7
Qwen2.5-7B	82.1	82.4 [79.6, 85.3]	0.998/0.984	89.2	370/400	89.5	82.1
Mistral-7B-v0.3	82.1	82.7 [79.7, 85.5]	0.998/0.985	86.2	369/400	88.6	82.1
InternLM2.5-7B	79.7	80.7 [77.7, 83.5]	0.996/0.981	86.2	363/400	89.5	79.7
Qwen2.5-14B	87.7	88.1 [85.5, 90.6]	0.999/0.994	90.4	373/400	93.6	87.7
Phi-3-medium-14B	83.1	84.0 [81.0, 86.9]	0.997/0.988	90.9	373/400	89.0	83.1
Qwen3-30B-A3B	87.5	88.0 [85.4, 90.5]	0.999/0.993	92.3	371/400	94.3	87.5

Table 26: **Auxiliary attribution cost.** Times measure only the two diagnostic heads after the native prefill. The heads reuse the detector's hidden states and add no model forward.

Backbone	Med. ms	P95 ms	Prefill %	VRAM MB	Fwds.
Llama-3.1-8B	1.8	3.1	1.26	14.2	1
Qwen2.5-7B	1.6	2.8	1.24	12.8	1
Mistral-7B-v0.3	1.7	2.9	1.30	13.5	1
InternLM2.5-7B	1.7	3.0	1.26	13.1	1
Qwen2.5-14B	2.4	4.1	1.07	18.6	1
Phi-3-medium-14B	2.3	3.9	1.05	17.9	1
Qwen3-30B-A3B	3.1	5.2	0.99	24.5	1

The natural panel contains 610 packages from 150 grouped families: 210 CLEAN, 115 DECL, 175 CODE, and 110 BOTH. It spans five source ecosystems and six malicious-risk categories. The controlled panel contains 128 base families, each represented by clean, Declaration-only, Code-only, and dual-region variants. All members of a family remain in one of five outer folds. The reported intervals resample families within source 2,000 times.

In Table 25, D/C denotes the Declaration and Code heads. "Blocked" counts primary-detector blocks among 400 malicious packages; conditional exact attribution measures correctness within those blocks. Joint success requires both the correct primary verdict and exact attribution among all 610 natural packages.

Because the two heads reuse cached D/O states, their median cost is 1.6–3.1 ms and their peak additional memory is below 25 MB on every backbone.

E.3 EXACT-SLOT O→B CAUSAL CONFIRMATION

We test whether Operation states affect the pre-action Boundary, beyond supporting a diagnostic readout. The test contains 40 source–target families. Within each family, operation slots have equal token counts and identical absolute Operation and Boundary positions under both native tokenizers. A score-blind odd/even split assigns 20 families to development and 20 to confirmation. Development selects the best of seven normalized intervention layers by its smallest mean margin over the controls (Llama layer 3 and Qwen layer 9). No confirmation family informs layer selection, controls, or gates.

At the selected layer, the exact source Operation token states replace the aligned target states. Let $\mathbf{b}_t$, $\mathbf{b}_s$, and $\mathbf{b}_p$ denote the target, donor, and patched Boundary vectors. Recovery is the normalized projection of the patch-induced displacement onto the donor direction:

$$\text{Recovery} = \frac{(\mathbf{b}_p - \mathbf{b}_t)^\top (\mathbf{b}_s - \mathbf{b}_t)}{\|\mathbf{b}_s - \mathbf{b}_t\|_2^2 + 10^{-8}}.$$

It measures directional displacement, not fractional distance reduction. Four controls isolate specificity: eight same-energy random directions at the same slot, the semantic delta at an equal-length neutral slot, a permutation of the source token order, and a per-token rescaled delta from another family. Confidence intervals use 10,000 paired family-bootstrap replicates.

Exact replacement produces positive Boundary recovery in all 20 confirmation families on each backbone. Llama's median recovery is 0.592. Its mean recovery gains over the random, neutral-position, token-permutation, and mismatched-semantic controls are 0.420 [0.379, 0.457], 0.439

[0.407, 0.474], 0.323 [0.277, 0.373], and 0.219 [0.186, 0.251], respectively. Qwen's median recovery is 0.944. Its corresponding gains are 0.743 [0.672, 0.804], 0.876 [0.823, 0.929], 0.540 [0.464, 0.613], and 0.366 [0.277, 0.451]. Every confirmation gate passes on both backbones. Replacing Operation states therefore moves the downstream Boundary toward the donor state on held-out families. This measures causal O→B influence; the learned transport residual is a separate statistical proxy. The following experiment measures application-probe response to multi-depth Operation interventions.

E.4 Detector-Aligned Score and Verdict Test

This experiment uses 100 distinct Large-ID matched families, with 20 pairs drawn from each source by a deterministic, score-blind rule. Every pair uses the fixed application probe for which its matched family was outer-test data. It recomputes the eight-depth input with a compact limit of eight D and 24 O atoms, reusing the Large-ID projections, member encoders, and final readouts. No head is refit and no additional ensemble is formed. The intervention starts from the benign package and inserts the matched malicious Operation trajectory at doses 0.25, 0.50, and 1.00. It patches the seven causally upstream depths of the detector's eight-depth trajectory; the final depth is excluded because it cannot alter the same-layer Boundary state.

At each depth, Operation atoms are aligned by normalized order. Their malicious-minus-benign mean difference is added to all target-atom tokens. Two controls use independent Gaussian directions at the same atoms and depths, with every atom-level norm matched to the real intervention. We then recompute the Boundary state, the three component margins, and the fixed-threshold verdict. For component $k \in \{B, O, G\}$, the logit margin is $m_k = \log[p_k/(1-p_k)]$. We report the fused margin

$$m(x) = \max\{m_B(x), \min(m_O(x), m_G(x))\} = \operatorname{logit} s(x). \tag{8}$$

Thus $m(x) \geq 0$ gives the same verdict as $s(x) \geq 0.5$ in Equation 6. The intervention generates no tokens and executes no payload.

Point estimates weight the five sources equally. Confidence intervals use 10,000 source-stratified family-bootstrap replicates. One-sided family sign-flip tests compare the real intervention with its random controls; Holm correction is applied across seven backbones separately for each endpoint. Eligibility for the verdict analysis is fixed before intervention: the benign score must be below the operating threshold and its matched malicious score at or above it. For each pair, dose AUC is the trapezoidal integral of the real-minus-control response over doses $\{0, 0.25, 0.50, 1.00\}$.

Branch activity is measured on unpatched benign targets. Verdict Δ is the full-dose real-minus-random pass-to-block rate among eligible pairs.

Table 27: **Operation interventions shift Boundary and fused margins.** Dose AUC is relative to norm-matched random controls. Brackets: 95% source-stratified family-bootstrap CIs; *: Holm-corrected positive sign-flip test.

Backbone	Natural PairAcc (%)	Boundary recovery	Boundary margin	Operation margin	Geometry margin	Fused margin $m(x)$
Llama-3.1-8B	89.0* [83.0, 94.0]	0.096* [0.077, 0.116]	2.880* [2.281, 3.473]	1.496* [1.149, 1.839]	0.639 [0.082, 1.197]	2.465* [2.035, 2.920]
Qwen2.5-7B	90.0* [84.0, 95.0]	0.174* [0.151, 0.196]	4.436* [3.836, 5.028]	1.632* [1.245, 2.013]	2.703* [2.175, 3.230]	3.336* [2.826, 3.841]
Mistral-7B	86.0* [79.0, 92.0]	0.107* [0.091, 0.123]	2.554* [2.077, 3.043]	2.429* [2.032, 2.842]	1.833* [1.392, 2.272]	2.520* [2.134, 2.902]
InternLM2.5-7B	88.0* [82.0, 94.0]	0.125* [0.111, 0.138]	3.219* [2.593, 3.829]	2.117* [1.718, 2.527]	1.300* [0.794, 1.794]	2.978* [2.470, 3.479]
Qwen2.5-14B	88.0* [82.0, 94.0]	0.111* [0.093, 0.130]	4.452* [3.868, 5.040]	2.221* [1.782, 2.677]	1.959* [1.408, 2.515]	3.621* [3.089, 4.159]
Phi-3-medium-14B	89.0* [83.0, 94.0]	0.125* [0.107, 0.143]	4.500* [3.827, 5.187]	2.188* [1.745, 2.646]	0.349 [−0.137, 0.830]	3.316* [2.877, 3.755]
Qwen3-30B-A3B	92.0* [87.0, 97.0]	0.119* [0.105, 0.133]	4.930* [4.298, 5.567]	2.100* [1.667, 2.532]	1.606* [1.162, 2.048]	3.230* [2.731, 3.755]

Table 28: **Operation interventions change fixed-probe verdicts.** Verdict Δ: full-dose real-minus-random pass-to-block rate on threshold-eligible pairs (pp). Brackets: 95% source-stratified family-bootstrap CIs.

Backbone	B active (%)	O active (%)	G active (%)	Eligible pairs	Verdict Δ pp [95% CI]	p_{H}
Llama-3.1-8B	33.0	51.0	16.0	54/100	53.4 [42.4, 64.3]	$< 10^{-3}$
Qwen2.5-7B	22.0	48.0	30.0	58/100	39.9 [25.9, 53.1]	$< 10^{-3}$
Mistral-7B	29.0	53.0	18.0	54/100	46.0 [34.5, 57.5]	$< 10^{-3}$
InternLM2.5-7B	40.0	23.0	37.0	61/100	66.1 [55.1, 77.2]	$< 10^{-3}$
Qwen2.5-14B	26.0	36.0	38.0	51/100	34.5 [29.3, 39.6]	$< 10^{-3}$
Phi-3-medium-14B	30.0	38.0	32.0	60/100	47.5 [33.6, 62.9]	$< 10^{-3}$
Qwen3-30B-A3B	24.0	44.0	32.0	58/100	39.8 [28.5, 51.2]	$< 10^{-3}$

Natural malicious-over-benign PairAcc is 86–92%. The real trajectory patch produces positive Boundary-recovery, Boundary-head, Operation-head, and fused logit-margin dose responses on all seven backbones after Holm correction ($p_{\mathrm{H}} < 10^{-3}$ throughout). Geometry also responds on five. The final logit-margin dose AUC ranges from 2.465 to 3.621. Depending on the backbone, 51–61 of the 100 pairs are threshold-eligible. Within this set, the real patch raises pass-to-block flips by 34.5–66.1 percentage points over the norm-matched controls; every interval excludes zero. The single-depth experiment tests O→B influence. The multi-depth experiment measures joint three-branch response and verdict sensitivity under compact extraction; neither isolates Boundary as the necessary mediator of the fused decision.

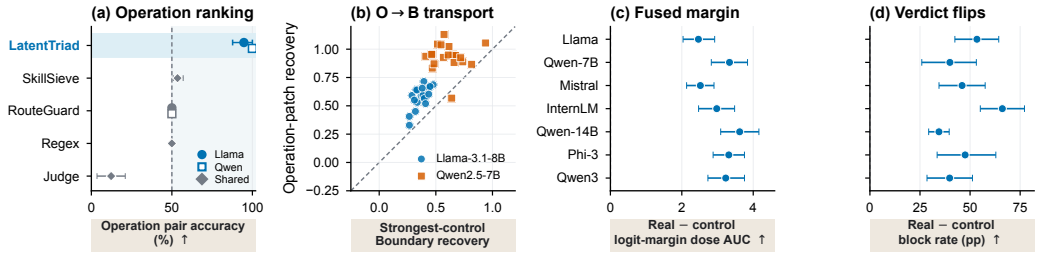


Figure 3: **Operation interventions change Boundary states and frozen policy verdicts.** (a) Operation ranking. (b) One-layer Boundary recovery. (c,d) Seven-depth intervention effects on margins and eligible verdicts. Bars are 95% intervals; (c,d) use source-stratified family bootstrap.

F Runtime and Resource Accounting

F.1 Measurement Protocol and Costs

Runtime measurements use the compact CDF-consensus configuration, rather than the package detector. Table 29 lists the extraction and scoring settings for each study.

Table 29: **Extraction and readout configurations by study.** Atom limits are D/O maxima.

Study	Atoms	Readout	Evidence scope
Package detection	8/256	Concatenated logistic	Content-component-disjoint folds
White-box / broker	8/256	Fixed application probes	Matched-pair-disjoint folds; rewrites and request policies
Multi-depth intervention	8/24	Fixed application probes	Same fitted heads; compact extraction for intervention
Compact runtime	8/24	Member CDF scores	Llama-3.1-8B and Qwen2.5-7B timing

The compact measurements use one RTX 4090 host with Python 3.12.13, PyTorch 2.6.0, CUDA 12.4, Transformers 5.13.0, BF16, SDPA, and batch size one. After 20 warmups, we measure 120 MalSkillBench packages per backbone three times with GPU synchronization. This yields 360

Table 30: **Compact 8/24-atom CDF-consensus timing.** Each backbone has 360 paired measurements. Incremental time includes hook callbacks, D/O/B pooling, feature construction, CDF calibration, and scoring; hook setup is excluded.

Backbone	Stage	Mean ms	Median ms	P95 ms	Std. ms
Llama-3.1-8B	Native prefill	334.1	219.4	905.2	365.8
	LATENTTRIAD incremental	64.6	64.7	71.1	9.1
	Prefill + detector	398.7	283.9	1028.4	367.9
Qwen2.5-7B	Native prefill	310.8	202.7	838.5	333.8
	LATENTTRIAD incremental	56.4	56.1	62.0	9.6
	Prefill + detector	367.2	258.7	896.5	334.5

paired measurements per backbone. Six empirical token-length bins contribute 20 packages each. Labels and detector scores do not affect selection. Peak process GPU memory is 17,574 MiB for Llama and 17,242 MiB for Qwen.

The frozen timing benchmark subtracts a paired no-hook prefill from the instrumented forward and scoring interval. The increment includes hook callbacks, D/O/B pooling, projections, feature construction, member-level CDF calibration, and consensus scoring. Model loading, tokenization and sample preparation, bundle/projection loading, and hook registration/removal occur outside the timed interval. It aggregates Boundary and Operation member margins by their medians and Geometry by its maximum. Its 64.7 ms Llama and 56.1 ms Qwen median increments measure this compact pipeline, not the current 256-atom, concatenated-logistic Large-ID pipeline.

The capability-broker's greedy typed-request generation is not included in detector increments. Table 30 separates native prefill from detector overhead; neither overhead alone nor the compact configuration establishes total agent latency or the full Large-ID configuration's runtime.

G CONTROLLED PRE-ACTION REPRESENTATION STUDIES

These studies distinguish information accessible to a diagnostic readout from the performance of the deployed detector. Unless specified otherwise, the three backbones are Llama-3.1-8B, Qwen2.5-7B, and Mistral-7B-v0.3. B32 and O32 read 32-dimensional Boundary and Operation representations, respectively. The independent-state (Ind.192) and residual-feature (Rel.192) readouts each use 192 features. They differ from the full-width Boundary and 735-dimensional Geometry components in Large-ID. These studies do not change the reported package detector or its threshold.

G.1 RISK PREDICTION ACROSS EXPRESSION FORMS

The controlled dataset has 800 examples from ten domain blocks, five task–operation compatibility rules, and four expression forms: natural language, Python plans, JSON configuration, and mixed text. Matched operations preserve the trusted task and vary whether the selected object satisfies its compatibility rule. Same-group variants stay together. Five folds hold out domain blocks; the all-forms condition includes every form in training and test. Each of the four leave-one-form-out conditions excludes its test form from both training and development and evaluates it on held-out domains.

All projections and benign-reference maps fit on training examples. Development selects layer, logistic regularization, and the operating threshold. Ind.192 constructs features separately from each state; Rel.192 includes benign transport residuals. B32, O32, and D32 read one region, while TF-IDF and length provide text controls. The seven methods, three models, and five settings yield 105 aggregate metric rows. Table 31 reports all settings. Each all-forms evaluation uses 800 examples; each held-out-form evaluation uses 200.

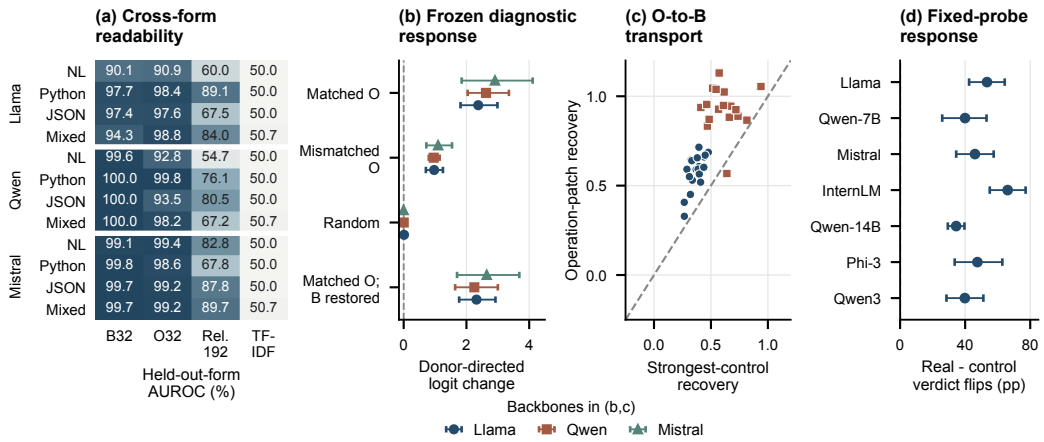


Figure 4: **Pre-action operation information is readable across forms and functionally affects frozen judgments.** (a) Cross-form AUROC. (b) Diagnostic logit shifts. (c) One-layer Boundary recovery by family. (d) Compact-probe verdict flips relative to controls. Intervals are 95% bootstrap CIs; intervention protocols differ across (b–d).

Table 31: Cross-form risk AUROC (%). All seven readouts and all form holdouts; five domain-disjoint folds.

Model	Test form	B32	O32	D32	Ind.192	Rel.192	TF-IDF	Length
Llama	All	99.75	99.98	50.00	99.28	99.77	50.23	50.00
Llama	NL	90.09	90.92	50.00	89.26	59.98	50.00	50.00
Llama	Python	97.66	98.36	50.00	92.12	89.09	50.02	50.00
Llama	JSON	97.44	97.63	50.00	87.83	67.54	50.00	50.00
Llama	Mixed	94.33	98.81	50.00	89.33	84.02	50.70	50.00
Qwen	All	100.00	99.86	50.00	99.89	99.93	50.23	50.00
Qwen	NL	99.59	92.81	50.00	78.59	54.66	50.00	50.00
Qwen	Python	100.00	99.80	50.00	95.00	76.08	50.02	50.00
Qwen	JSON	100.00	93.47	50.00	88.52	80.52	50.00	50.00
Qwen	Mixed	99.95	98.21	50.00	98.20	67.23	50.70	50.00
Mistral	All	99.88	98.86	50.00	99.99	99.95	50.23	50.00
Mistral	NL	99.14	99.41	50.00	90.91	82.81	50.00	50.00
Mistral	Python	99.75	98.59	50.00	96.15	67.79	50.02	50.00
Mistral	JSON	99.66	99.20	50.00	75.42	87.85	50.00	50.00
Mistral	Mixed	99.67	99.22	50.00	93.67	89.69	50.70	50.00

AUROC pools out-of-fold predictions from frozen heads. Intervals resample the ten domain blocks 5,000 times, keeping related variants together. Paired AUROC tests enumerate the 1,024 domain-level method-score swaps and apply Holm correction to three prespecified comparisons per model and setting. In all-forms evaluation, Rel.192 exceeds TF-IDF on all three backbones (corrected $p = 0.00586$), but does not significantly exceed B32 or Ind.192. This comparison concerns the fitted lexical baseline, not every possible static analyzer.

Pairwise ordering and pooled ranking measure different properties. On many holdouts, matched-pair accuracy remains high even when pooled AUROC decreases. Score scales can differ across domains and independently fitted fold heads. For example, Mistral's held-out Python Rel.192 readout has 67.79% pooled AUROC but 100% matched-pair accuracy. Qwen's natural-language holdout also loses pairwise ordering, so score-scale changes do not explain every failure. These exploratory comparisons do not select a new calibration or test threshold.

Layerwise distances use training-fitted coordinate standardization. We compare a semantic-preserving form change with a risk-changing operation under fixed form. Form changes dominate whole-vector distances at most depths, despite strong risk readouts. Qwen's late Boundary

states show the opposite ordering; Llama and Mistral do not share this reversal. Pointwise domain-bootstrap intervals describe individual depths, without selecting test-set peaks or claiming simultaneous layerwise significance.

G.2 Matched Operation Interventions

For each of the 800 recipients per model, the all-forms Rel.192 diagnostic and its development threshold remain frozen. At the depth preceding the selected readout, we intervene on Operation states using a donor matched in domain, compatibility rule, expression form, and trusted task. Only the operation object changes. Both benign-to-malicious and malicious-to-benign directions are included. We report the logit-margin change directed toward the donor label, not a probability change or an attack execution rate.

Controls comprise an unchanged forward pass, a sham replacement, same-position norm-matched random perturbations, unrelated Declaration perturbations with matched total energy, norm-matched semantically mismatched donors, and the matched Operation donor followed by restoration of Boundary to its unpatched state. Five random seeds are averaged within each recipient before domain aggregation. They do not increase the independent sample count. Table 32 includes all conditions.

Table 32: Donor-directed diagnostic logit changes. Intervals resample ten domains.

Condition	Llama	Qwen	Mistral
Matched O	2.376 [1.812, 2.988]	2.622 [2.038, 3.353]	2.909 [1.849, 4.110]
Matched O, B restored	2.320 [1.768, 2.924]	2.251 [1.640, 3.001]	2.646 [1.701, 3.685]
Norm-matched random	0.011 [0.007, 0.015]	0.002 [-0.006, 0.008]	-0.004 [-0.011, 0.002]
Mismatched O	0.971 [0.708, 1.251]	0.957 [0.801, 1.146]	1.097 [0.722, 1.540]
Unrelated D	0.115 [-0.011, 0.260]	0.158 [0.060, 0.243]	-0.019 [-0.053, 0.021]
Unpatched	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Sham	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

Matched donors flip the frozen diagnostic decision on 46.75%, 34.13%, and 28.75% of recipients on Llama, Qwen, and Mistral, respectively. All four prespecified matched-control contrasts per model pass the domain-level sign-swap test with Holm-corrected $p = 0.0078125$. Equal corrected values arise from the finite set of domain sign swaps, not equal effect sizes.

Sham effects and verdict flips are zero. Most of the matched-donor effect remains after Boundary restoration: the diagnostic still directly reads O and cross-region residuals. This supports functional sensitivity to the operation intervention without establishing Boundary as a necessary mediator. The separate exact-slot and compact detector-aligned experiments in Appendices E.3 and E.4 retain their own inputs, heads, and statistical units.

G.3 Region Masks on Identical Hidden States

D/O are structural proxies, not a natural-language/code partition. The heading parser can treat a #-prefixed line within a code fence as a heading. Across 5,830 packages, this moves the declaration boundary inside a fence in 27 cases (0.46%). We test all 27 affected cases and 27 matched controls on Llama and Qwen. The serialized input, tokens, hidden states, and Boundary state are unchanged; only the D/O pooling masks change.

Neither model changes a binary verdict on the affected cases. Maximum absolute probability changes are 0.30932 on Llama and 0.02283 on Qwen. Thus, this concrete parsing issue does not change decisions in the tested set, but some scores are sensitive to the mask. The experiment is neither a semantic segmentation accuracy estimate nor evidence that all possible region definitions are equivalent. All reported detector results use the same structural parsing rule; this test changes masks only for diagnosis.

G.4 CompoSkill Input Dependence

The monitoring set comprises 344 explicit and 346 implicit malicious–benign pairs: 690 mode-specific pairs, or 1,380 records. These are not 690 independent attack families. Their family union contains 361 identities, 329 shared between modes, grouped into 76 scenario/persona role groups. The five threat categories contribute 150 data-exfiltration, 128 memory-tampering, 140 multi-agent-hijack, 123 privilege/dangerous-command, and 149 resource-exhaustion pairs.

Input ablations refit diagnostic readouts under the same five role-group folds. They compare the task, workspace, installed nodes, local contexts, and full path, using the original labels for the full paths. A workspace-only TF-IDF readout reaches 98.38% explicit and 100.00% implicit AUROC. Pure-node views are at 50% because each matched pair has identical skill content. On Llama, a full-path diagnostic reaches 84.97/99.62% AUROC (explicit/implicit), versus 93.64/100.00% for workspace only. On Qwen, the corresponding values are 95.75/99.99% and 95.24/100.00%. These are diagnostic refits, not the final path detector.

Workspace cues therefore suffice for strong prediction on this benchmark. This does not mean that the final detector relies only on workspace cues. It means that benchmark accuracy alone cannot identify how the model represents skill dependencies. The controlled fixed-node study below tests that question without directly exposing the final target in the workspace.

G.5 Fixed-Node Composition Counterfactuals

We fix the bytes of three skill files and cross their connection and authorization settings. The workspace supplies routing and a starting key; the composed finite map determines the target. A deterministic, no-I/O oracle checks whether that target is authorized. The 144 adaptable cases and 31 domain labels are drawn from SCR-CapFlow's organization, yielding 1,152 bundles and 4,608 four-cell presentations. This is one controlled finite-map mechanism, not 31 independent mechanisms or the original SCR-Bench attack evaluation.

Five domain-grouped folds keep related variants together. Layers and hyperparameters are selected on development. Input-deletion controls retain the original full-path labels: their chance performance indicates missing information for that label, not safety of a newly defined partial task. Table 33 includes all 19 methods.

Figure 5a contains 186 comparisons: three backbones, 31 domains, and two multi-state readouts. Panel b compares the two 192-dimensional readouts on the same 93 backbone–domain pairs; B32 uses 32 dimensions. For panel c, each backbone contributes 2,304 authorization pairs. The empirical distributions weight backbones and domains equally and retain all held-out pairs. A positive score difference ranks the unauthorized input higher; magnitudes are not calibrated across readouts. Panels share inputs, not independent replicates.

Figure 6 separates the paired responses by backbone. Each panel retains all 2,304 held-out authorization pairs for each multi-state readout, without jitter or filtering. The two readouts share pairs, yielding 13,824 plotted points rather than independent experimental replicates. The diagonal indicates equal numerical score differences, not equivalent calibration across heads.

Table 33: Finite-map composition AUROC (%), equally weighted across 31 domain labels. All 19 fixed readouts.

Readout	Llama	Qwen	Mistral
Full-path B32	50.83	54.69	52.77
Task only	50.00	50.00	50.00
Workspace only	50.00	50.00	50.00
Task + workspace	50.00	50.00	50.00
Skills only	50.00	50.00	50.00
Node 1	50.00	50.00	50.00
Node 2	50.00	50.00	50.00
Node 3	50.00	50.00	50.00
Local context 1	50.00	50.00	50.00
Local context 2	50.00	50.00	50.00
Local context 3	50.00	50.00	50.00

Readout	Llama	Qwen	Mistral
Node aggregate (32)	50.00	50.00	50.00
Local aggregate (32)	49.97	49.74	49.99
Local concatenation (96)	49.70	50.03	50.36
Local squared (64)	65.07	59.20	65.42
Independent-state squared (192)	80.41	70.77	70.17
Relational residual squared (192)	74.55	72.68	78.99
TF-IDF	50.00	50.00	50.00
Length	50.00	50.00	50.00

Independent-state and residual features both use squared transformations and 192 dimensions. Relative to the independent construction, residual AUROC changes by -5.86 [-7.77, -3.92] points on Llama, $+1.92$ [-1.45, $+5.23$] on Qwen, and $+8.82$ [$+7.08$, $+10.51$] on Mistral. Holm-corrected paired p values are 0.00080, 0.27275, and 0.00080. The residual contribution therefore depends on the backbone. Comparisons with B32 change feature width and construction as well as the regions read. Using squared features in a readout does not identify a particular nonlinear computation inside the model.

The corresponding composition intervention uses all 4,608 recipients per model, half with changed risk and half with unchanged risk. Matched workspace donors keep skill files and the trusted task fixed. Donor-directed diagnostic margin changes are 0.172/0.292/0.154 for Llama/Qwen/Mistral, versus 0.002/-0.0001/-0.0005 for norm-matched random controls and 0.069/0.084/0.044 for mismatched donors. Restoring B leaves 0.172/0.288/0.158. However, unchanged-risk recipients still flip decisions at 14.92/7.06/8.47%, compared with 14.31/6.85/2.02% on changed-risk recipients. The intervention changes the diagnostic but does not isolate a risk-specific causal effect. It therefore does not establish a universal composition mechanism.

G.6 Full-Architecture Structural Transfer

The broader synthetic test uses six templates: dictionary binding, role requirements, credential validation, delegation chains, scope intersection, and conditional fallback. Each model has 2,304 inputs: 576 training, 192 development, and 1,536 test presentations. The test uses new opaque symbol assignments, with both single-skill and composed packaging and two neutral-edit variants. The known-template panel trains on all six templates. Six leave-one-template-out panels train on 480 and develop on 160 examples, testing 256 presentations of the omitted template. Template reuse and symbol changes define the generalization axis; these are not independent natural attack mechanisms.

Unlike the pooled diagnostic readouts, this test fits the complete LatentTriad architecture from genuine local atoms at eight depths. It retains 32-dimensional latent coordinates, offsets 300/400/500, the unmodified component-head family, and fixed max–min fusion. The models run in BF16 without quantization, CPU offload, or truncation. Four state-feature controls, Boundary-only, and LatentTriad are evaluated on every panel. Independent-linear here is fitted on synthetic training data, whereas the independent-state control in Appendix C.2 is fitted on natural packages.

Table 34: Full-architecture transfer AUROC (%). Both neutral-edit variants and every control. Held-out values average all six template panels.

Model	Templates	Edit	Package	B	LT	Ind.L	Ind.S	Orth.S	Quad.
Llama	Known	0	Single	51.01	52.31	53.91	54.02	53.47	52.93
Llama	Known	0	Composed	51.60	53.47	54.46	54.41	54.38	54.18
Llama	Known	1	Single	49.87	51.38	51.45	51.24	51.22	51.20
Llama	Known	1	Composed	49.98	50.05	50.93	50.90	50.88	51.03
Llama	Held out	0	Single	49.76	49.25	50.08	49.85	49.72	49.45
Llama	Held out	0	Composed	50.34	51.06	50.85	50.86	50.62	50.15
Llama	Held out	1	Single	49.72	50.05	49.95	49.84	50.23	49.32
Llama	Held out	1	Composed	50.54	50.91	50.42	50.29	50.78	50.55
Qwen	Known	0	Single	49.12	51.30	52.85	53.19	52.72	51.50
Qwen	Known	0	Composed	50.29	51.12	52.08	52.10	51.99	52.03

Model	Templates	Edit	Package	B	LT	Ind.L	Ind.S	Orth.S	Quad.
Qwen	Known	1	Single	50.75	50.16	52.52	52.13	51.35	52.13
Qwen	Known	1	Composed	47.79	49.64	50.02	49.80	49.30	49.64
Qwen	Held out	0	Single	53.69	52.96	53.52	53.92	53.81	52.62
Qwen	Held out	0	Composed	51.29	52.86	53.52	53.53	53.52	53.24
Qwen	Held out	1	Single	51.63	52.64	53.32	53.39	53.27	52.72
Qwen	Held out	1	Composed	50.65	53.09	52.72	52.90	53.01	52.23

Table 34 reports all 96 aggregate AUROC cells. Ind.L and Ind.S use independent linear and squared features; Orth.S uses orthogonal squared features; Quad. uses explicit quadratic features. Held-out entries are means of all six template panels, not pooled scores from different fitted heads. We provide descriptive results without treating six constructed templates as a population of independent mechanisms.

For the main neutral variant, LATENTTRIAD AUROC ranges from 49.25% to 53.47%. At both development-calibrated 5% and 10% FPR budgets, LATENTTRIAD and Boundary-only make identical test decisions on every panel and packaging. This identity also holds for the second neutral variant. Development-budget calibration does not ensure the same test FPR: at the 10% budget, test FPR is 3.12–9.38% in known-template panels but 28.65–34.38% under template holdout. This test therefore supports neither stable structural transfer nor a low-FPR advantage over Boundary-only. It refits the architecture on synthetic training data; it does not reevaluate the natural-package checkpoints.

The finite-map result and this broader transfer test answer different questions. Retaining local atoms at eight depths does not recover transfer here, so missing local features alone cannot explain the result.

G.7 SAME-PROMPT READOUTS AND GENERATED JUDGMENTS

This exploratory program test uses three operators (object flow, delegation, and conditional selection), six lexical families, four expression forms, two packaging styles, and both answer orders. Each model has four condition–target panels of 384 questions: the actual object must either be inferred or is supplied, and the target is either the object or its risk. Three operator folds rotate training, development, and test roles. Actual-object and risk heads are trained separately. All three fixed seeds reuse the same questions, so they do not triple the independent support.

For each question, the Boundary readout and generation start from the exact same prompt. The readout uses the first prefill; generation is evaluated at 96, 256, and 512 tokens. Bfull uses the model's full hidden width rather than a compressed B32/B128 projection. The table reports all fixed methods, including random-training-label and text controls. Development chooses regularization from $\{0.01, 0.1, 1\}$, with ties resolved toward smaller values.

Table 35: Same-prompt readout accuracy (%). All three seed means; 384 questions per model, condition, and target.

Model	Object	Target	B32	B128	Bfull	Random	Text	Options
Llama	Inferred	actual	55.47	60.16	61.46	47.92	50.00	50.00
Llama	Inferred	risk	49.83	51.82	53.39	50.69	50.00	50.00
Llama	Supplied	actual	67.19	78.12	75.52	47.31	50.00	50.00
Llama	Supplied	risk	59.72	60.94	59.90	49.74	50.00	50.00
Qwen	Inferred	actual	52.00	52.78	51.30	49.39	50.00	50.00
Qwen	Inferred	risk	51.65	54.51	53.12	48.96	50.00	50.00
Qwen	Supplied	actual	81.08	89.24	89.84	48.09	50.00	50.00
Qwen	Supplied	risk	75.09	79.86	89.84	51.22	50.00	50.00

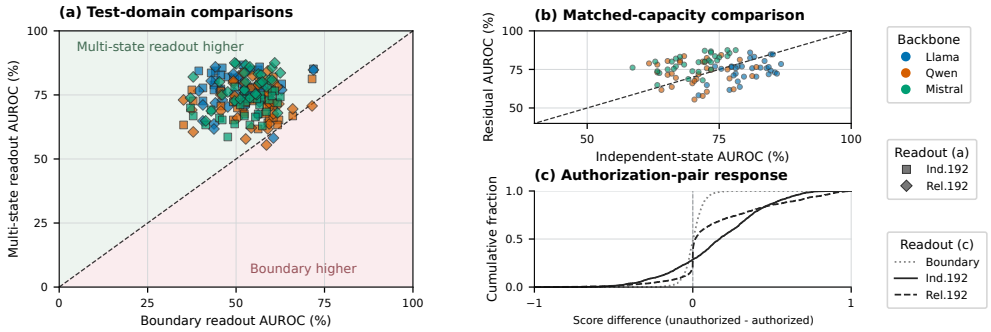


Figure 5: **Multi-state readouts expose controlled composition risk; residual gains vary across domains.** (a) Multi-state versus Boundary AUROC by domain. (b) Matched-capacity independent-state versus residual-feature AUROC. (c) Cumulative score differences (unauthorized minus authorized). Diagonals mark equal AUROC; positive differences indicate correct risk ordering.

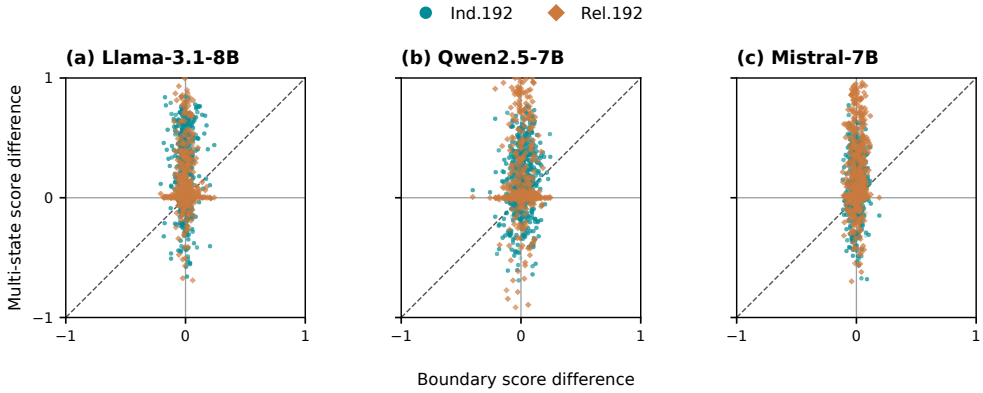


Figure 6: **Authorization-pair responses by backbone.** Each point compares Boundary and multi-state score differences (unauthorized minus authorized). Positive values indicate correct risk ordering. Colors distinguish the two multi-state readouts; axes use a common scale.

Table 36: Same-prompt generation accuracy (%). Parentheses count format failures, retained as errors; each denominator is 384.

Model	Object	Target	96 tokens	256 tokens	512 tokens
Llama	Inferred	actual	55.47 (166)	76.30 (42)	77.60 (32)
Llama	Inferred	risk	61.72 (77)	76.56 (3)	76.56 (3)
Llama	Supplied	actual	59.90 (139)	80.21 (15)	80.21 (13)
Llama	Supplied	risk	69.27 (87)	84.64 (8)	84.90 (4)
Qwen	Inferred	actual	52.34 (176)	93.23 (4)	94.01 (1)
Qwen	Inferred	risk	23.18 (286)	92.71 (10)	94.53 (2)
Qwen	Supplied	actual	51.82 (185)	95.05 (18)	99.74 (0)
Qwen	Supplied	risk	44.27 (211)	89.58 (14)	91.15 (6)

Random denotes the B128 random-training-label control; Text uses the full question and Options removes the program. Format failures are counted as errors rather than discarded. In particular, Qwen's inferred-risk accuracy rises from 23.18% at 96 tokens to 94.53% at 512 tokens, with format failures falling from 286 to 2. A short-budget failure is not evidence that the model cannot solve the task.

Bfull inferred-risk accuracy is 53.39% on Llama and 53.12% on Qwen, compared with 76.56% and 94.53% after 512-token generation. Supplying the actual object raises Bfull to 59.90% and 89.84%. This separates successful generated reasoning from accessibility to the tested fixed linear head. It does not prove that the first prefill lacks the information, that generation creates new knowledge, or that model scale determines detectability.