SECRESPOND: BENCHMARKING AND TRAINING AI AGENTS FOR REAL-WORLD POST-COMPROMISE INCIDENT RESPONSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Model (LLM) agents are increasingly adopted in real-world security operations with access to host artifacts and command-line interfaces (CLIs), making it critical to assess their security capabilities. However, existing cybersecurity benchmarks mostly focus on pre-compromise settings where agents operate in a clean environment before an attack occurs. This leaves the post-compromise setting underexplored. To address this gap, we introduce SECRESPOND, the first benchmark for evaluating LLM agents to autonomously conduct the post-compromise incident-response workflow in an end-to-end manner. Given a forensic disk snapshot of a compromised host together with the alerts, vulnerability scans, and baseline checks reported by a host security product, agents are required to produce forensic reports on intrusions, baseline risks, and vulnerability risks, together with a remediation plan. We instantiate this task across 10 cyber ranges, each constructed from a distinct compromised cloud host, spanning 4 entry-point types, 21 ATT&CK techniques, and 5 operating systems. We evaluate 23 frontier LLMs on the OpenCode agent harness. Experimental results show that although current agents can reliably uncover the problems exposed by alerts, they struggle to proactively investigate the disk for silent intrusions and to produce comprehensive, verified remediation plans. We further propose a training recipe for post-compromise agents with skill internalization and RETrace, a report-anchored credit assignment, which assigns each rubric checkpoint's score to the corresponding output span and the investigation step. On out-of-domain ranges, RETrace improves over outcome-reward RL by 8.8% in detection and 7.2% in planning.

1 INTRODUCTION

Large Language Model (LLM) agents, equipped with iterative reasoning and autonomous tool-use abilities (OpenAI, 2026; Qwen, 2026; Z.ai, 2026), demonstrate remarkable potential in tackling complex, professional real-world tasks through command-line interfaces (CLIs) (Merrill et al., 2026; Ye et al., 2026; Li et al., 2026), among which cybersecurity is a significant deployment target (Wang et al., 2025; Deason et al., 2025). Among security operations, post-compromise incident response is particularly demanding. When an intrusion triggers the alerts of a host security product, the security team must investigate the compromised host, reconstruct the attack chain, and remediate the system. In practice, this process is carried out manually by experts, which is labor-intensive and hard to scale. The traditional agentless tools deployed in the current workflow only match static patterns against a fixed knowledge base and cannot reconstruct the attack chain and plan the remediation. Therefore, automating this process with LLM agents is promising, but how reliably they can fulfill such a pipeline without human intervention remains unclear.

Existing cybersecurity benchmarks mostly target ***pre-compromise*** settings, where agents find, exploit, or patch weaknesses in a clean environment (Zhang et al., 2025; Wang et al., 2025; Wiz, 2026; Meta, 2025). Other works evaluate whether agents can analyze attacks that have already happened by reasoning over isolated alerts (Deason et al., 2025) and system logs (Wu et al., 2025), or predefined incident response sub-tasks with a human executing the model's commands (Lin et al., 2025). This leaves the ***autonomous post-compromise workflow*** underexplored, where the agent must actively examine the host for hidden evidence and plan the remediation, shown in Table 1.

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.

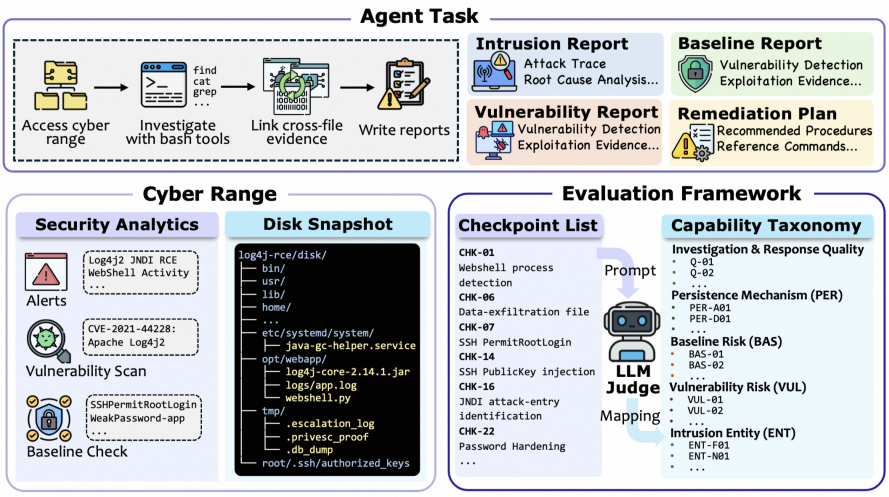


Figure 1: Overview of SECRESPOND. In our benchmark task, the agent is expected to investigate a cyber range and produce incident-response reports. Each range includes a forensic disk snapshot of a real cloud host compromised through an end-to-end attack, together with alerts and scan findings from a security product. The agent is evaluated by LLM judges against range-specific checkpoints, each mapped to an item of the capability taxonomy.

To address this gap, we introduce **SECRESPOND**, the first benchmark for evaluating LLM agents to autonomously conduct post-compromise incident-response workflow in an end-to-end manner, as shown in Figure 1. Given a forensic disk snapshot of a compromised host together with the alerts, vulnerability scans, and baseline checks of a host security product, the agent is required to investigate the host and report the intrusion, the vulnerability and baseline risks, and a remediation plan. We instantiate this task on 10 cyber ranges from hosts compromised through end-to-end attacks over real network protocols, spanning 4 entry-point types, 21 ATT&CK techniques, and 5 operating systems. To make such open-ended reports rigorously judgeable, we propose a hierarchical rubric by decomposing the ranges into 280 checkpoints. We further map them to 52 items of a five-dimensional capability taxonomy, which makes results comparable across heterogeneous ranges and reflects model capability in the full incident response workflow from investigation to remediation.

We conduct extensive evaluation of 23 different large language models (LLMs) on a representative agent harness, OpenCode (OpenCode, 2026). Our experiments reveal that agents can surface the problems exposed by alerts and surpass the agentless tool used in the current workflow, but they struggle to investigate the hidden or erased traces and remediate the risks thoroughly. Even the strongest model, Claude Opus 4.7, achieves only 72.4% averaged over detection and planning, and performance drops further as the attack chain grows longer and broader. The results indicate that the fundamental bottleneck lies in the investigation depth rather than in interpreting the evidence. To explore how to address this bottleneck, we first *internalize* expert procedural priors by fine-tuning on skill-guided trajectories, which transfers the investigation routine. In order to reach deeper evidence on unseen hosts, we further apply *agentic reinforcement learning* to let the agent learn from its own investigative rollouts. However, optimizing with a single outcome reward spread over multiple investigative steps and reports cannot indicate the responsible step or paragraph, and thus degrades the performance. We therefore propose **RETrace** (**R**eport-anchored credit assignment with **E**vidence **Trace**back), which assigns each fine-grained rubric item to the corresponding output span and back to the step that produced its evidence, outperforming GRPO by 8.8% in detection and 7.2% in planning on out-of-domain ranges, with the largest gain on the checkpoints without an explicit alert.

In summary, our contributions are as follows: **❶ A benchmark for end-to-end post-compromise incident response with a hierarchical capability evaluation framework.** We establish SECRESPOND, the first benchmark for evaluating LLM agents to autonomously conduct the post-compromise incident-response workflow end to end, with 10 cyber ranges built from real compromised hosts and a hierarchical rubric mapped to a five-dimensional capability taxonomy. **❷ An empirical study of frontier agents on real-world post-compromise tasks.** We evaluate 23 LLMs and show that current

Table 1: Comparison with existing cybersecurity benchmarks. **Post-Compromise**: target the real-world scenario after a host has already been attacked; **Real Filesystem**: grounded in a host disk; **Multi-Step Analysis**: require multiple investigative steps; **Cross-File Analysis**: the evidence is correlated across multiple files; **Rubric-Based Evaluation**: the evaluation metrics apply fine-grained rubrics; **CLI-Compatible**: the benchmark supports evaluating CLI agent harnesses.

Benchmark	Scope	Post-Compromise	Real Filesystem	Multi-Step Analysis	Cross-File Analysis	Rubric-Based Evaluation	CLI-Compatible
SecBench (Jing et al., 2024)	Knowledge	✗	✗	✗	✗	✗	✗
NYU CTF Bench (Shao et al., 2024)	CTF	✗	✗	✓	✓	✗	✗
CyberSecEval (Wan et al., 2024)	Offensive	✗	✗	✓	✗	✓	✗
CyBench (Zhang et al., 2025)	CTF	✗	✗	✓	✓	✗	✓
CVE-Bench (Zhu et al., 2025)	CVE	✗	✗	✓	✗	✗	✗
AutoPatchBench (Meta, 2025)	Patch	✗	✓	✓	✓	✗	✗
CyberGym (Wang et al., 2025)	Vuln	✗	✓	✓	✓	✗	✓
CyberSOCEval (Deason et al., 2025)	SOC	✗	✗	✗	✗	✗	✗
ExCyTIn-Bench (Wu et al., 2025)	Threat Detection	✓	✗	✓	✓	✗	✗
CyberModelArena (Wiz, 2026)	Offensive	✗	✓	✓	✓	✗	✓
IRBench (Lin et al., 2025)	Incident Response	✓	✓	✓	✓	✗	✗
SECRESPOND (Ours)	Incident Response	✓	✓	✓	✓	✓	✓

agents reliably discover alerted problems but struggle with proactive investigation and thorough remediation. ❸ **A training recipe for post-compromise agents from skill internalization to credit assignment.** We internalize expert procedural priors by fine-tuning on skill-guided trajectories and propose RETrace, a Report-anchored credit assignment method with Evidence Traceback for agentic reinforcement learning, which deepens investigation on unseen ranges.

2 RELATED WORK

Agent Benchmarks. To comprehensively evaluate the capabilities of LLM agents from multiple aspects, many benchmarks have been developed. Jimenez et al. (2024); Deng et al. (2025); Merrill et al. (2026) validate agent models' coding abilities. Barres et al. (2025); He et al. (2025) benchmark agents in multi-turn conversational interaction. Recently, agentic harnesses (OpenClaw, 2026; OpenCode, 2026) have encouraged evaluation of agents operating within these scaffolds (Team et al., 2026; Zhang et al., 2026b; Ye et al., 2026). Nevertheless, these evaluations target software engineering and general productivity tasks, overlooking agents' potential on security operations.

Benchmarking LLMs in Cybersecurity. Existing benchmarks generally evaluate the cybersecurity capabilities of LLM agents from two aspects, offensive and defensive. Offensive benchmarks progress from assessing security knowledge (Li et al., 2024) to interactive exploitation, where agents are required to solve CTF challenges (Zhang et al., 2025; Shao et al., 2024), exploit real-world vulnerabilities (Wang et al., 2025; Zhu et al., 2025), or execute a full attack lifecycle (Wiz, 2026). Defensive benchmarks cover threat report understanding (Deason et al., 2025) and patch generation (Meta, 2025). Closest to our setting, ExCyTIn-Bench (Wu et al., 2025) focuses on intrusion investigation using alerts and security logs, and IRBench (Lin et al., 2025) evaluates incident response on machines from public practice platforms against predefined sub-tasks, with a human executing its commands. None of the existing work exposes autonomous agents to actual end-to-end host intrusion with forensic artifacts, including persistence mechanisms and deliberately erased traces.

Training LLM Agents. Current methods can be categorized into two directions: *(i) Injecting behavior with teacher guidance* by imitating teacher trajectories (Qin et al., 2024; Gao et al., 2026) or on-policy distillation that matches the teacher's token-level distribution on the student's own rollouts (Lu & Lab, 2025; Fang et al., 2026). *(ii) Optimizing the policy through reinforcement learning* with verifiable rewards (Guo et al., 2025; Zheng et al., 2025) or rubric-based reward (Zhang et al., 2026a) to improve agentic capabilities. However, aggregating the reward into a single scalar for each trajectory carries less information about responsible steps or tokens. Credit assignment methods introduce finer-grained signals by step-level evaluators (Lightman et al., 2024; Zhou et al., 2024) or comparing alternative continuations at intermediate states (Kazemnejad et al., 2024; Feng et al., 2026). These methods are difficult to adapt to long-horizon open-ended tasks, where rollouts rarely share a state and a step cannot be judged before its finding reaches the output. We instead anchor the credit on the corresponding output spans, and trace it back to the actions that produce the evidence.

3 SECRESPOND BENCHMARK

SECRESPOND evaluates LLM agents' ability to investigate a post-compromise host snapshot and produce comprehensive forensic reports on intrusion analysis, baseline assessment, and vulnerability verification, together with a structured remediation plan. An overview is shown in Figure 1. In this section, we first describe the benchmark task (§3.1). Then, we define a capability taxonomy that specifies what the agent should be able to accomplish in the incident-response process (§3.2). Finally, we introduce the evaluation framework, which decomposes each range into fine-grained checkpoints that link concrete forensic evidence to capability items, enabling hierarchical evaluation (§3.3).

3.1 OVERVIEW OF TASKS

To comprehensively evaluate the agents' incident response abilities, we construct 10 cyber ranges covering multiple attack scenarios. A cyber range includes a frozen and reproducible forensic disk snapshot of a host that has been compromised by an end-to-end attack delivered over real network protocols, along with a set of security analytics from a security product. These ranges cover different operating systems, entry points, ATT&CK techniques, and attack stages, which are selected to mimic the distribution in actual incidents based on a correlation analysis of alerts, vulnerabilities, and baselines from 372 real compromised cloud hosts. The details of ranges are summarized in Table 4, and Appendix A.2 explains the range environment construction.

Specifically, each range exposes two sources of evidence to the agent. The forensic disk snapshot typically covers all the directories of the compromised host, and contains the intrusion artifacts naturally generated by the attack, including logs, configuration changes, and residual files, which serve as the primary forensic evidence. Beyond the disk snapshot, the agent also has access to three types of security analytics from the host security product on the victim machine: (i) real-time alerts from the anomaly detection engine monitoring suspicious processes, network connections, and login events, (ii) vulnerability findings from scans of installed packages and applications, and (iii) baseline check results from configuration checks. Only 30–40% of the attack actions in each range trigger a product alert, while the remaining 60–70% leave only silent file and configuration artifacts on disk. In such an environment, the agent is required to produce one progress file and four reports: (i) an ***intrusion report*** reconstructing the attack chain covering entry point, lateral movements, persistence mechanisms, and final impacts, (ii) a ***vulnerability report*** verifying exploited weaknesses present on the host, (iii) a ***baseline report*** assessing host configuration against security baselines, and (iv) a ***remediation plan*** with recommended steps and reference commands for executing the fixes.

3.2 CAPABILITY TAXONOMY

To benchmark agent models' incident response capability, we organize the required capabilities to investigate and remediate a compromised host into a five-dimensional Capability (CAP) taxonomy. **Intrusion Entity (ENT)** denotes the ability to locate and handle the artifacts left by an intrusion, spanning malicious processes, tampered files, and malicious network entities. **Persistence Mechanism (PER)** indicates the ability to recover the mechanisms an attacker installs to survive reboot and cleanup, including scheduled tasks, backdoors, kernel-level techniques, etc. **Baseline Risk (BAS)** is the ability to examine host configuration against security baselines, covering SSH hardening, credential safety, permission auditing, etc. **Vulnerability Risk (VUL)** represents the ability to confirm the weaknesses on the host, spanning vulnerabilities of Java components, web application code, container escape, etc. **Investigation & Response Quality (Q)** reflects the ability to produce a thorough, well-evidenced, and trustworthy investigation, evaluating honesty, comprehensiveness, business-impact assessment, etc. In total, the taxonomy comprises 52 items across five dimensions as shown in Table 5, including 12 in ENT, 21 in PER, 7 in BAS, 4 in VUL, and 8 in Q. These capability items represent the benchmark target and form the foundation to build compromised ranges.

3.3 CHECKPOINT DECOMPOSITION

To evaluate the capabilities above, we decompose them into fine-grained checkpoints within each range to measure whether the corresponding capability is fulfilled in the specific scenario. A checkpoint (CHK) is the smallest scoring unit, specifying the evidence required to discover a finding in a specific range. We design 280 checkpoints in total. Each checkpoint matches one or more

items in the CAP taxonomy, and multiple checkpoints may be assigned to the same CAP, as detailed in Table 6. We first define **CHK-score**, which grades each CHK along two axes, detection and planning. *Detection* measures whether the agent correctly uncovers problems on the compromised host, including intrusion entities, baseline risks, and vulnerability risks. *Planning* validates whether the agent proposes a correct and complete remediation plan that resolves the existing problems. The normalized score of a single checkpoint on axis a is defined as the CHK-score, which represents the agent model's performance at the checkpoint level: $\text{CHK-score}_c^a = \frac{s_c^a}{M^a} \in [0,1], (a \in \{\text{det}, \text{plan}\})$, where $M^{\text{det}} = 3$ and $M^{\text{plan}} = 2$ are the maximum scores of the two axes (see details in Appendix D.1), and $s_c^a \in \{0, 1, \ldots, M^a\}$ is the score assigned to checkpoint c on axis a.

Since each checkpoint is anchored to a capability item, we can aggregate the checkpoints from different ranges mapped to the same capability item into a capability-level score, which we define as the **CAP-score**, which is the ratio of the achieved scores to the maximum scores over all checkpoints mapped to the corresponding capability, computed as: $\text{CAP-score}^a = \frac{1}{|\mathcal{C}_{\text{CAP}}^a|} \sum_{c \in \mathcal{C}_{\text{CAP}}^a} \text{CHK-score}_c^a$, where $\mathcal{C}_{\text{CAP}}^a$ is the set of checkpoints mapped to the capability item.

4 Experiments

4.1 Experimental Settings

We evaluate 23 large language models spanning 8 model series and their successive versions, as reported in Table 2. All models are driven by the same agent harness, OpenCode (OpenCode, 2026), and exposed to the same task environment, instruction prompt (see Appendix C.1), and tool set to ensure a fair comparison. We use default parameters for each model with reasoning enabled.

Since the outputs are open-ended reports that cannot be verified against a fixed ground truth such as flags or execution outputs, we follow the LLM-as-a-Judge paradigm (Gu et al., 2026). Specifically, we score each generated report against the range checklist defined in §3.3. To reduce single-judge bias, we adopt three strong proprietary models, Claude Opus 4.7 (Anthropic, 2026), Gemini 3.1 Pro (The Gemini Team, 2026), and GPT-5.4 Pro (OpenAI, 2026), as independent judges. Each judge runs in the same OpenCode harness, reads the reports produced by the assessed model, and assigns a score for every checkpoint against its criteria (see Appendix C.3). The CHK-score_c^a of each checkpoint is averaged over the three judges. We report the range-level CHK-score, the mean of CHK-score_c^a over the checkpoints of a range on axis $a \in \{\text{det}, \text{plan}\}$, for range-specific analysis in Table 2. Following the capability taxonomy described in § 3.2, we aggregate CHK-score into CAP-score using the mapping between range checkpoints and capability items (see Table 6), and use CAP-scores for range-aggregated analysis in Figure 2. We validate this protocol against human judgment on 60 checkpoints covering all 10 ranges, graded blindly by a security expert. The judges reach a quadratic-weighted κ of 0.94 and a mean absolute error of 0.15 against the expert (Appendix D.5).

4.2 Benchmark Results

Finding 1: Detection consistently outperforms planning across all models. As shown in Table 2, Claude Opus 4.7 achieves the highest average. All the models achieve a higher range-level CHK-score on Detection than on Planning. For example, GPT-5.5 scores 70.7% on detection but only 36% on planning, while Claude Sonnet 4.5 presents the smallest gap but still reaches 8.5%. This implies that even when the agent models can successfully detect the problems from the disk snapshot, they fail to provide a comprehensive approach to fixing them. The remediation plans often stop at removing the visible artifact and leave secondary risks or undetected persistence mechanisms untouched.

Finding 2: Range difficulty varies, and model performance declines as the attack chain grows longer and broader. The overall CHK-score across models on each range reveals a clear difficulty stratification that aligns with the structural complexity of each scenario. Agent models generally show strong performance on Log4j-RCE, Docker-Escape, and Redis-RCE ranges with a single entry point, a linear attack chain, and common persistence exposed in alerts. Ranges like SSH-Miner, Next.js-RCE, and Jenkins-RCE include multi-step privilege escalation, cross-service correlation and more diverse persistence, leading to a decrease in both detection and planning scores. The hardest ranges are Shiro-Fastjson, RDP-Service-Abuse, NPM-Worm, and ASP.NET-ViewState, which combine a broad attack surface, cross-service or cross-runtime correlation, and deliberately disguised persistence.

Table 2: The range-level CHK-score$_r^a$ of Detection and Planning for each range. "Det" denotes CHK-score$_r^{det}$, the percentage of achieved detection scores over all its checkpoints, averaged across three LLM judges; "Plan" denotes the corresponding CHK-score$_r^{plan}$ on the planning dimension.

Model	SSH-Miner		Redis-RCE		Docker-Escape		Jenkins-RCE		Shiro-Fastjson		Log4j-RCE		Next.js-RCE		NPM-Worm		ASP.NET-ViewState		RDP-Service-Abuse		Overall	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7†	**77.8**	**83.3**	82.4	69.6	90.8	75.5	**78.9**	**70.2**	71.6	52.5	88.4	78.7	77.8	60.1	—	—	83.3	37.9	59.9	63.9	**79.0**	**65.7**
Claude Opus 4.6	75.3	48.0	79.6	60.8	**92.1**	62.3	66.7	60.6	66.7	43.2	88.4	68.5	76.5	64.9	81.5	60.9	77.8	**49.3**	77.8	61.1	78.2	58.0
Claude Opus 4.5	72.5	61.7	70.4	55.9	79.2	57.0	71.9	62.3	69.6	55.2	85.9	71.3	66.7	60.8	69.8	65.4	56.1	27.3	53.1	47.2	69.5	56.4
Claude Sonnet 4.6	75.8	44.1	68.5	61.8	83.3	63.2	67.2	59.7	76.8	60.9	87.9	**81.5**	71.6	**70.2**	68.7	53.8	58.1	29.5	**89.5**	**66.7**	74.7	59.1
Claude Sonnet 4.5	**77.8**	76.5	86.1	75.6	84.2	78.1	62.6	55.3	73.8	**71.9**	78.3	39.8	65.7	69.1	41.0	49.4	73.8	41.7	53.7	54.7	69.7	61.2
GPT-5.5	67.6	27.4	78.2	49.0	68.0	35.1	73.7	33.4	63.9	27.0	**96.5**	33.3	58.5	40.4	55.8	41.0	81.3	19.7	63.6	53.7	70.7	36.0
GPT-5.4 Pro	65.2	27.4	75.9	47.1	66.7	41.2	69.0	19.4	80.5	63.0	91.9	42.6	65.7	44.7	65.5	37.2	**85.9**	31.9	58.6	52.8	72.5	40.7
GPT-5.4	57.5	25.5	57.4	34.3	69.4	20.2	55.0	29.0	56.3	38.5	85.8	41.7	51.0	36.9	57.8	32.7	67.2	26.5	49.4	47.2	60.7	33.2
GPT-5.2 Pro	61.3	21.5	59.7	43.1	54.2	34.2	73.1	39.4	65.9	36.4	88.4	44.5	71.9	44.0	51.3	34.6	56.6	22.0	79.6	50.9	66.2	37.1
Gemini 3.1 Pro	66.7	38.3	55.1	38.2	60.7	43.0	60.2	39.5	43.0	28.7	72.2	32.4	46.4	32.1	43.0	11.5	47.5	22.7	53.1	28.7	54.8	31.5
Gemini 3 Flash	58.5	25.5	62.5	34.3	57.4	39.5	56.1	32.5	50.1	27.6	80.8	51.9	57.5	31.5	41.9	26.3	49.0	17.4	60.5	42.6	57.4	32.9
GLM-5.1	75.4	58.8	80.1	62.8	91.2	**86.0**	71.4	53.5	**82.5**	63.5	90.4	64.8	**93.1**	61.9	75.8	60.9	61.6	27.3	42.0	52.8	76.3	59.2
GLM-5	75.8	56.9	60.2	49.0	75.5	51.8	50.3	34.3	49.4	42.7	88.4	64.8	76.5	62.4	50.4	31.4	67.2	26.5	54.9	38.9	64.9	45.9
DeepSeek V4 Pro	76.3	61.8	67.1	56.9	75.9	50.0	63.2	59.7	73.1	55.2	87.9	49.1	75.2	57.1	80.1	64.1	71.2	32.6	62.3	47.2	73.2	53.4
DeepSeek V3.2	60.9	53.0	65.7	48.1	65.3	37.8	50.9	47.4	68.9	47.9	80.3	68.6	58.5	40.5	36.2	21.1	41.9	16.7	63.6	54.6	59.2	43.6
Qwen3.7 Max	75.8	50.0	**87.5**	**76.5**	79.2	64.0	72.5	51.7	70.4	46.4	90.4	60.1	82.4	63.1	**84.9**	48.7	84.9	40.9	51.8	41.7	78.0	54.3
Qwen3.7 Plus	70.5	36.2	86.6	71.6	75.9	57.0	71.9	69.3	82.0	70.8	88.9	75.0	70.9	61.9	73.5	**65.5**	81.8	35.6	53.7	45.3	75.6	58.8
Qwen3.6 Plus	70.5	54.9	66.2	52.9	77.3	64.1	64.3	32.5	60.0	44.3	84.8	54.6	73.2	67.9	65.0	38.4	68.2	31.8	54.9	41.7	68.4	48.3
Qwen3.5 Plus	65.7	51.0	66.2	58.8	61.6	46.5	67.8	49.2	43.5	32.3	82.8	57.4	64.7	55.4	61.8	37.2	70.2	21.2	51.2	31.5	63.5	44.0
Kimi K2.6	60.4	34.3	63.0	27.4	88.4	79.0	76.6	53.5	53.1	40.6	84.9	55.6	63.4	31.5	52.4	36.5	64.7	33.3	51.3	38.0	65.8	43.0
Kimi K2.5	71.5	61.8	72.7	64.8	75.9	59.7	55.6	43.0	51.6	46.4	88.9	69.4	63.1	53.6	43.3	35.9	46.0	23.5	52.5	49.1	62.1	50.7
MiniMax M2.7	69.1	36.3	62.0	51.9	74.5	56.2	63.7	49.1	55.1	41.6	76.8	38.0	66.4	45.2	47.6	32.0	42.9	14.4	42.6	35.2	60.1	40.0
MiniMax M2.5	63.3	38.3	56.5	30.4	61.6	48.2	47.4	32.5	38.5	22.9	76.3	45.4	57.8	42.3	31.9	23.1	46.5	12.9	43.8	33.3	52.4	32.9

† Claude Opus 4.7 returned a safety refusal on every NPM-Worm attempt and produced no report, so its overall score is therefore computed as the average over the remaining nine ranges.

Finding 3: Agents substantially outperform the traditional agentless detection baseline on average, and the gap is largest on detecting persistence mechanism. We also compare the agent models against an agentless baseline used in the traditional incident-response workflow, a production scanner that inspects a host's disk snapshot purely by static pattern matching against a predefined knowledge base. The scanner reaches 43.6% on ENT and 50.0% on VUL, but only 20.7% on BAS and 2.1% on PER, because it reports only isolated static artifacts and cannot analyze persistence mechanisms, reconstruct attack chains, or attribute entry points. In contrast, agents actively connect scattered artifacts across files into an intrusion chain, uncovering persistence mechanisms hidden in the system, and further provide root-cause analyses and remediation plans.

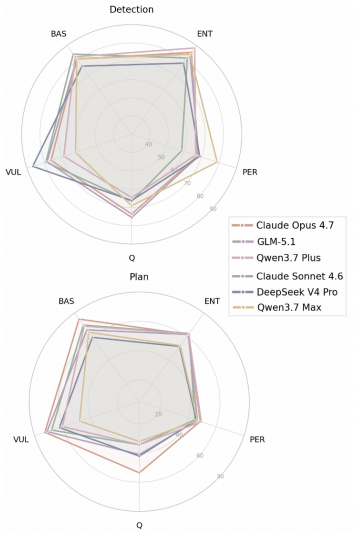
Finding 4: Capabilities are imbalanced on different dimensions. As shown in Figure 2, none of the models dominates across all five dimensions. On the detection side, models reliably identify intrusion entities but struggle to uncover persistence mechanisms. Most models exceed a 70% detection CAP-score on Intrusion Entity (ENT), whereas Persistence Mechanism (PER) has the lowest detection CAP-score, since persistence hides in locations with no active signal. For example, even the strongest model, Claude Opus 4.7, reaches 93% on the detection checkpoints that an alert reveals but only 67% on the silent intrusions (Figure 6). On the planning side, models remediate risks effectively but perform weakly on intrusion entity and persistence mechanism cleanup. Baseline Risk (BAS) and Vulnerability Risk (VUL) receive the highest planning CAP-scores. Lastly, Investigation & Response Quality is consistently among the weakest capabilities across models, reflecting both inadequate detection and unverified remediation.

5 Training Recipe for SecRespond

The analysis in §4.2 shows that agents uncover the alerted problems but leave silent persistence undetected and the remediation incomplete. A typical failure mode is procedural: the agent investigates only the exposed evidence, and lacks a systematic response procedure. We therefore distill the operational experience of security incident-response teams into a skill that contains general investigation routines and response procedures with no range-specific knowledge leakage such as evidence locations. Prompting the agent with the skill raises planning but barely changes detection. On Qwen3.5-9B the gains are 7.2% and 0.9% respectively (Table 3), and a similar trend is observed on frontier models as in Appendix E.2. This is because the response procedure can be followed directly, whereas the prompting does not make the model execute the investigation routine over multiple steps. This raises two questions: (1) *Can the procedural prior be internalized into the model weights*, and (2) *Can post-training teach the model to investigate beyond the procedure and uncover deeper evidence*. We study both questions on a 9B model and find that the difficulty of SecRespond

Model	ENT		PER		BAS		VUL		Q	
	Detect	Plan	Detect	Plan	Detect	Plan	Detect	Plan	Detect	Plan
Agentless	43.6	—	2.1	—	20.7	—	50.0	—	—	—
Claude Opus 4.7	85.4	61.9	67.8	48.7	80.8	**74.8**	76.2	**72.6**	73.6	53.6
Claude Opus 4.6	86.0	58.1	72.1	45.8	78.6	67.6	79.2	63.3	68.8	34.4
Claude Opus 4.5	76.5	60.8	56.9	44.0	75.8	63.5	72.6	54.4	62.4	33.3
Claude Sonnet 4.6	81.6	**62.7**	58.7	44.9	**84.2**	69.8	79.2	67.8	66.7	34.4
Claude Sonnet 4.5	72.8	58.4	56.1	44.1	66.2	70.8	68.1	66.7	71.6	**66.7**
GPT-5.5	78.4	31.1	71.3	28.8	63.6	46.8	61.5	40.0	61.4	24.4
GPT-5.4 Pro	78.3	38.3	76.4	35.2	73.5	51.9	61.5	45.6	64.0	35.5
GPT-5.4	69.9	29.2	47.6	24.4	61.8	46.4	53.3	37.8	55.2	15.5
GPT-5.2 Pro	73.8	36.0	67.6	39.0	62.4	44.2	42.2	17.8	62.9	27.8
Gemini 3.1 Pro	65.2	33.8	48.7	24.1	52.1	34.5	37.0	31.1	43.8	12.2
Gemini 3 Flash	66.5	35.3	47.8	23.5	58.5	36.2	62.2	43.3	45.4	12.2
GLM-5.1	84.1	61.8	68.4	46.7	81.1	65.7	78.5	71.1	**75.5**	41.1
GLM-5	67.8	42.1	54.2	36.6	67.8	54.1	68.9	58.8	56.5	25.5
DeepSeek V4 Pro	78.1	51.7	69.0	45.8	76.1	59.4	**86.7**	62.2	66.2	42.2
DeepSeek V3.2	69.1	42.9	47.2	30.1	57.7	50.0	69.6	54.5	54.8	36.7
Qwen3.7 Max	84.1	52.4	**79.0**	46.7	80.9	63.7	62.2	47.8	69.1	32.2
Qwen3.7 Plus	**88.4**	62.3	66.7	**49.4**	82.4	68.9	68.9	60.0	64.6	34.4
Qwen3.6 Plus	77.1	47.1	57.5	32.2	74.7	62.6	54.8	52.3	58.0	24.5
Qwen3.5 Plus	80.2	44.6	49.5	31.5	61.9	49.0	47.4	44.4	52.0	30.0
Kimi K2.6	72.6	43.9	57.9	33.9	66.7	48.6	55.5	41.1	55.7	26.7
Kimi K2.5	72.0	52.4	49.5	36.0	57.7	52.9	65.9	60.0	54.9	42.3
MiniMax M2.7	66.0	34.1	43.7	30.3	68.6	51.2	63.7	38.9	52.8	32.3
MiniMax M2.5	61.5	33.3	38.4	22.0	55.2	38.7	40.0	35.5	42.8	14.4

(a)



(b)

Figure 2: (a) The CAP-score^{a} of agent models and the traditional agentless scanning method across the five capability dimensions. For each dimension, "Detect" denotes CAP-score^{det} averaged across three LLM judges; "Plan" denotes CAP-score^{plan}. (b) The CAP-scores of representative models across the five capability dimensions.

lies in improving investigation depth, since the training signal of standard post-training methods is diluted over the investigation steps and report sections of a trajectory. To tackle this problem, we propose **RETrace** (**R**eport-anchored credit assignment with **E**vidence **Trace**back) for finer-grained credit assignment over long-horizon, multi-report agent trajectories.

Training Setup. To measure generalization to unseen attack scenarios, we split the 10 ranges into two folds: each range is in-domain (ID) for one fold and out-of-domain (OOD) for the other. Evaluation is conducted on both folds separately with all scores averaged over 3 inference runs and 3 judges. Appendix F.1 gives the fold allocation, data statistics, and other implementation details.

5.1 Skill Internalization

We first explore if the model itself can be enhanced by internalizing the skills into the model weights. We attempt two types of methods that can distill external knowledge, Supervised Fine-Tuning (SFT) and On-Policy Distillation (OPD) (Lu & Lab, 2025). SFT is the most effective, since training on skill-guided teacher trajectories improves the base model by 12.9% and 17.5% on detection and planning on ID ranges and outperforms SFT with no-skill trajectories. On-policy distillation (OPD), which lets the teacher supervise the student's own rollouts at a token level, falls below the base model under both reverse and forward KL. This is because the token-level guidance is conditioned on the student's own context. It aligns the student with the teacher on the student's predicted actions, but rarely provides signal on the teacher's investigation trajectory, so the student still skips many of the tool calls that the teacher performs. Additional analysis is discussed in Appendix F.1. Nevertheless, internalization has a limitation. On OOD ranges, the detection score gains only 3.5% over the base model, far below the 12.9% gain on ID ranges. Detection ability on unseen ranges is thus what the procedural prior fails to transfer. SFT can transfer the rules of investigation and remediation across ranges, but the evidence beyond these rules can only be reached with exploratory actions. We therefore let the model learn from the outcome of its own investigation with reinforcement learning.

5.2 Report-Anchored Credit Assignment with Evidence Traceback

In this section, we explore whether the model can learn from the outcome of its own rollouts to achieve a more thorough investigation beyond the internalized skills. We conduct RL with the range-level

Table 3: Training results on SECRESPOND: average range-level CHK-scores (%) on the in-domain (ID) and out-of-domain (OOD) ranges, averaged over 2 folds, 3 inference runs, and 3 judges. "Det" and "Plan" denote the detection and planning scores.

Method	Supervision	ID		OOD	
		Det	Plan	Det	Plan
Skill internalization					
Qwen3.5-9B	–	56.4±1.5	33.0±1.4	56.4±1.5	33.0±1.4
+ skill prompt	–	57.3±0.6	40.2±0.9	57.3±0.6	40.2±0.9
SFT	Qwen3.5-397B	63.1±2.3	42.9±2.3	59.2±1.7	36.3±1.5
SFT	Qwen3.5-397B + skill	69.3±1.3	50.5±1.6	59.9±1.3	45.0±1.8
OPD	Qwen3.5-397B	51.7±1.2	22.4±1.2	48.6±1.3	20.8±1.1
Reinforcement Learning from SFT checkpoint					
+GRPO	outcome reward	65.5±1.5	48.5±1.6	56.5±2.1	42.8±2.0
+GSPO	outcome reward	65.8±2.0	47.1±1.5	57.9±1.6	42.6±1.7
+RETrace (Ours)	outcome reward with credit assign	**73.5**±0.8	**55.2**±0.9	**65.3**±1.5	**50.0**±1.2
w/o traceback	outcome reward with credit assign	66.6±1.3	51.3±1.5	60.1±1.2	45.2±1.8
w/o report anchor	outcome reward with credit assign	68.5±1.6	47.8±1.9	62.7±1.1	41.8±1.4
Scaling to Larger Base Model					
Qwen3.5-27B	–	55.7±0.8	39.7±1.1	55.7±0.8	39.7±1.1
SFT	Qwen3.5-397B + skill	66.3±1.3	50.7±1.3	60.4±1.1	47.8±1.5
+GRPO	outcome reward	61.3±1.5	46.6±1.3	60.2±1.0	49.3±1.4
+RETrace (Ours)	outcome reward with credit assign	**70.5**±0.8	**54.8**±1.2	**66.8**±1.1	**54.3**±1.2

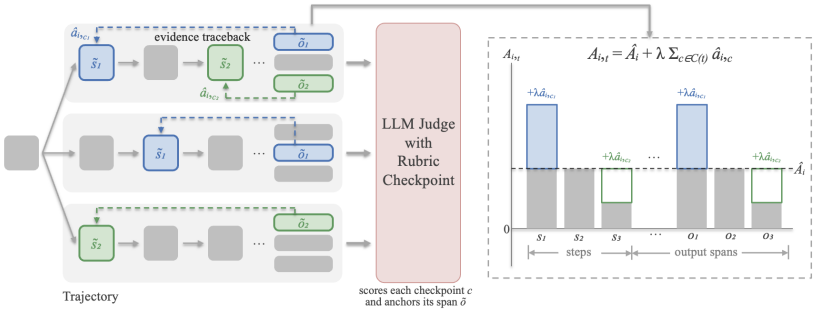
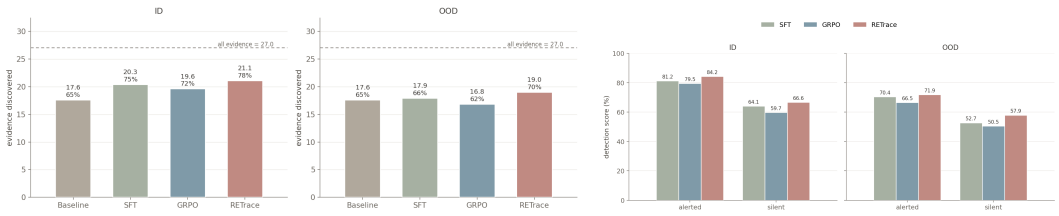


Figure 3: Overview of RETrace. For each rollout, the judge scores every rubric checkpoint c and anchors it to the related output span $\tilde{o}$, and the evidence in that span is traced back to the step $\tilde{s}$ with matched tool result. The advantage of each token is the addition of trajectory-level advantage $\hat{A}_i$ and the credit $\hat{a}_{i,c}$ of every corresponding checkpoint.

score as the reward. However, we find that further training with GRPO (Guo et al., 2025) drops below SFT by 3.8%/2.0% on ID ranges and 3.4%/2.2% on OOD ranges. Moreover, GSPO (Zheng et al., 2025), which is designed for long sequences, shows similar degradation. The training reward of GRPO stays flat during training, and the model uncovers fewer evidence items than SFT on OOD ranges, with the largest drop on the checkpoints that no alert reveals (see Appendix F.4). We attribute the failure to the granularity of the reward. A rollout contains multiple investigation steps followed by four reports, each contributing to different rubric checkpoints. Aggregating the checkpoint scores into a scalar reward and spreading the advantage uniformly credits every step and paragraph equally, regardless of its contribution. The policy receives the same signal for different steps and all the paragraphs in the generated reports. With such signals, the model cannot distinguish which step or paragraph produced silent findings that are not reflected in the alerts, and the exploratory steps that uncover them are never credited explicitly.

A natural method is to evaluate each investigation step and use the score as the credit for the step. However, the reward is based on the reports rather than the steps, so step-level credit does not ensure that the evidence found by a step is reflected in the report, and is thus not aligned with the final reward, which is illustrated in Appendix F.3. We therefore anchor credit on the report and trace it back to the investigation, called RETrace (Report-anchored credit assignment with evidence Traceback), as shown in Figure 3. In detail, when the judge scores a rubric item, it also localizes the output span that satisfies it, and the credit of that item is assigned to the tokens of that span rather than to the whole output. For an item that depends on evidence found in the trajectory, the credit additionally adds to the tokens of the action that introduced it by tracing the evidence mentioned in that span. In



(a) (b)

Figure 4: (a) The number of required evidence items that appear in the tool outputs of a trajectory. (b) The detection score on the checkpoints that an alert explicitly reveals and the rest silent intrusion for SFT, GRPO, and RETrace.

SECRESPOND, the rubric items are checkpoints, the spans are report paragraphs, and the checkpoint score is traced further to the investigative step. For each checkpoint, the judge additionally returns the paragraph that satisfies it and the corresponding evidence, such as a file path, a command, or a log line. Each evidence item is matched to the earliest step whose tool output contains it. For token t of rollout i, the advantage is $A_{i,t} = \hat{A}_i + \lambda \sum_{c \in \mathcal{C}(t)} \hat{a}_{i,c}$, where $\hat{A}_i$ is the normalized advantage as in GRPO, $\mathcal{C}(t)$ is the set of checkpoints whose paragraph or traced step covers t, and $\hat{a}_{i,c}$ is the normalized score of checkpoint c. An action is thus rewarded only when its evidence is written into the output and accepted by the judge.

As shown in Table 3, RETrace increases by 5.4% in detection and 5.0% in planning over SFT in OOD ranges. Compared to GRPO with the same outcome reward, RETrace gains 8.8% in detection and 7.2% in planning on OOD ranges. We ablate RETrace by keeping the checkpoint credit on the report paragraphs only (w/o traceback) or on the investigation steps only (w/o report anchor), with the remaining tokens receiving the trajectory-level advantage. Removing the traceback lowers OOD detection by 5.2% and planning by 4.8%, so tracing credit back to the investigation steps is what drives detection. Removing the report anchor lowers planning by 7.4% on ID and 8.2% on OOD ranges, below SFT, since crediting steps alone does not reward writing the evidence into the report. Scaling to a larger base model, RETrace also improves over SFT by 4.2%/4.1% on ID ranges and 6.4%/6.5% on OOD ones. The gain on unseen ranges indicates that the model learns a general investigation behavior rather than overfitting the training ranges.

Analysis. To examine the behavior after training with RETrace, we report the required evidence items that appear in the tool outputs of a trajectory, which measures the depth of the investigation. As shown in Figure 4 (a), SFT and GRPO hardly uncover more evidence than the base model on OOD ranges, while RETrace discovers the most on both ID and OOD ranges. Figure 4 (b) further splits the detection checkpoints by whether an alert reveals them as defined in Appendix D.4. GRPO falls below SFT on both alerted and silent checkpoints, while RETrace shows the largest improvement on the silent checkpoints. This demonstrates that crediting investigative steps through the output findings triggers deeper investigation.

6 CONCLUSION

In this paper, we introduce SECRESPOND, the first benchmark for evaluating LLM agents to autonomously conduct the post-compromise incident-response workflow end to end, comprising 10 cyber ranges built from real compromised hosts. By evaluating 23 frontier LLMs, we observe that current agents can reliably uncover the problems exposed by alerts but struggle to proactively investigate the disk for silent intrusions and to produce comprehensive, verified remediation plans. We further propose a training recipe for post-compromise agents, which internalizes an expert procedure by fine-tuning on skill-guided trajectories, and introduce RETrace, which assigns each checkpoint's score to the corresponding output span and the investigation step, improving over outcome-reward RL on both in-domain and out-of-domain ranges, especially on the silent intrusion. SECRESPOND provides a foundation for evaluating security agents on real incidents, and RETrace shows that fine-grained credit assignment can drive agents toward deeper investigation in incident response.

Ethics Statement

Third-Party Model Evaluation. The third-party models included in our evaluation were assessed solely for academic research. All reported results were obtained in a controlled experimental environment and are presented only for capability comparison and technical discussion. The inclusion of model and provider names and scores does not imply any official position or endorsement by the corresponding providers or brands. Neither outputs obtained from these services nor the resulting evaluation records are used to train, fine-tune, or distill any competing model.

Security Research and Use Restrictions. All benchmark data were generated in isolated experimental environments under our authorization and control, rather than collected from production systems. Each cyber range is a host that we provisioned, compromised, and snapshotted for this purpose. The released forensic and security-product artifacts are synthetic and contain no data from customer, production, or real-world business systems. We sanitized every disk snapshot by replacing real credentials, keys, and personal data with placeholders, retaining only the forensic artifacts required for analysis. The attack techniques, CVEs, and exploitation methods represented in the scenarios are derived solely from public knowledge bases and vulnerability disclosures, such as MITRE ATT&CK and public vulnerability databases. The benchmark contains forensic evidence rather than runnable exploit chains; it includes no proprietary or internally developed advanced attack techniques and no incremental attack information beyond public sources, and it cannot be directly used to conduct attacks against real-world systems. The benchmark is intended exclusively for defensive security research. Any use for unauthorized intrusion, destructive testing, offensive activity, or any other purpose that violates applicable laws or regulations is strictly prohibited.

AI use statement

LLMs were used to polish the writing of this paper. LLM judges are also applied in this work, and their use for evaluation and training is described in §4.1 and §5.

Reproducibility statement

We describe the construction of the environment in Appendix A.2, the evaluation framework in Appendix D.2, and training details in Appendix F.1. The benchmark, evaluation code, and training code will be released upon acceptance.

References

Anthropic. Introducing claude opus 4.7, 2026. URL https://www.anthropic.com/news/claude-opus-4-7.

V Barres, H Dong, S Ray, X Si, and K Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. *URL https://arxiv. org/abs/2506.07982*, 2025.

Lauren Deason, Adam Bali, Ciprian Bejean, Diana Bolocan, James Crnkovich, Ioana Croitoru, Krishna Durai, Chase Midler, Calin Miron, David Molnar, et al. Cybersoceval: Benchmarking llms capabilities for malware analysis and threat intelligence reasoning. *arXiv preprint arXiv:2509.20166*, 2025.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, et al. Swe-bench pro: Can ai agents solve long-horizon software engineering tasks? *arXiv preprint arXiv:2509.16941*, 2025.

Junfeng Fang, Zhepei Hong, Mao Zheng, Mingyang Song, Gengsheng Li, Houcheng Jiang, Dan Zhang, Haiyun Guo, Xiang Wang, and Tat-Seng Chua. Rubric-based on-policy distillation. *arXiv preprint arXiv:2605.07396*, 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Jiaxuan Gao, Jiaao Chen, Chuyi He, Shusheng Xu, Di Jin, and Yi Wu. From self-evolving synthetic data to verifiable-reward rl: Post-training multi-turn interactive tool-using agents. *arXiv preprint arXiv:2601.22607*, 2026.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 7(6), 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Wei He, Yueqing Sun, Hongyan Hao, Xueyuan Hao, Zhikang Xia, Qi Gu, Chengcheng Han, Dengchang Zhao, Hui Su, Kefeng Zhang, et al. Vitabench: Benchmarking llm agents with versatile interactive tasks in real-world applications. *arXiv preprint arXiv:2509.26490*, 2025.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Pengfei Jing, Mengyun Tang, Xiaorong Shi, Xing Zheng, Sen Nie, Shi Wu, Yong Yang, and Xiapu Luo. Secbench: A comprehensive multi-dimensional benchmarking dataset for llms in cybersecurity. *arXiv preprint arXiv:2412.20787*, 2024.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. *arXiv preprint arXiv:2410.01679*, 2024.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Xiangyi Li, Wenbo Chen, Yimin Liu, Shenghan Zheng, Xiaokun Chen, Yifeng He, Yubo Li, Bingran You, Haotian Shen, Jiankai Sun, et al. Skillsbench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Xihuan Lin, Jie Zhang, Gelei Deng, Tianzhe Liu, Tianwei Zhang, Qing Guo, and Riqing Chen. Ircopilot: Automated incident response with large language models. *arXiv preprint arXiv:2505.20945*, 2025.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Mike A Merrill, Alexander G Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E Kelly Buchanan, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv preprint arXiv:2601.11868*, 2026.

Meta. Introducing autopatchbench: A benchmark for ai-powered security fixes, 2025. URL https://engineering.fb.com/2025/04/29/ai-research/autopatchbench-benchmark-ai-powered-security-fixes/.

OpenAI. Introducing gpt-5.4, 2026. URL https://openai.com/index/introducing-gpt-5-4.

OpenAI. Introducing gpt-5.4, 2026. URL https://openai.com/index/introducing-gpt-5-4/.

OpenClaw. Openclaw: Open-source ai assistant, 2026. URL https://openclaw.ai/.

OpenCode. Opencode: The open source ai coding agent, 2026. URL https://opencode.ai/.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Qwen. Qwen3.7: The agent frontier, 2026. URL https://qwen.ai/blog?id=qwen3.7.

Minghao Shao, Sofija Jancheska, Meet Udeshi, Brendan Dolan-Gavitt, Haoran Xi, Kimberly Milner, Boyuan Chen, Max Yin, Siddharth Garg, Prashanth Krishnamurthy, et al. Nyu ctf bench: A scalable open-source benchmark dataset for evaluating llms in offensive security. *Advances in Neural Information Processing Systems*, 37:57472–57498, 2024.

CocoaBench Team, Shibo Hao, Zhining Zhang, Zhiqi Liang, Tianyang Liu, Yuheng Zha, Qiyue Gao, Jixuan Chen, Zilong Wang, Zhoujun Cheng, et al. Cocoabench: Evaluating unified digital agents in the wild. *arXiv preprint arXiv:2604.11201*, 2026.

The Gemini Team. Gemini 3.1 pro: A smarter model for your most complex tasks, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Shengye Wan, Cyrus Nikolaidis, Daniel Song, David Molnar, James Crnkovich, Jayson Grace, Manish Bhatt, Sahana Chennabasappa, Spencer Whitman, Stephanie Ding, et al. Cyberseceval 3: Advancing the evaluation of cybersecurity risks and capabilities in large language models. *arXiv preprint arXiv:2408.01605*, 2024.

Zhun Wang, Tianneng Shi, Jingxuan He, Matthew Cai, Jialin Zhang, and Dawn Song. Cybergym: Evaluating ai agents' cybersecurity capabilities with real-world vulnerabilities at scale. *arXiv e-prints*, pp. arXiv–2506, 2025.

Wiz. Cyber model arena, 2026. URL https://www.wiz.io/cyber-model-arena.

Yiran Wu, Mauricio Velazco, Andrew Zhao, Manuel Raúl Meléndez Luján, Srisuma Movva, Yogesh K Roy, Quang Nguyen, Roberto Rodriguez, Qingyun Wu, Michael Albada, et al. Excytin-bench: Evaluating llm agents on cyber threat investigation. *arXiv preprint arXiv:2507.14201*, 2025.

Bowen Ye, Rang Li, Qibin Yang, Yuanxin Liu, Linli Yao, Hanglong Lv, Zhihui Xie, Chenxin An, Lei Li, Lingpeng Kong, et al. Claw-eval: Towards trustworthy evaluation of autonomous agents. *arXiv preprint arXiv:2604.06132*, 2026.

Z.ai. Glm-5.1: Towards long-horizon tasks, 2026. URL https://z.ai/blog/glm-5.1.

Andy K Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, et al. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *International Conference on Learning Representations*, volume 2025, pp. 25094–25243, 2025.

Jiajie Zhang, Xin Lv, Ling Feng, Lei Hou, and Juanzi Li. Chaining the evidence: Robust reinforcement learning for deep search agents with citation-aware rubric rewards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20744–20765, 2026a.

Yuxuan Zhang, Yubo Wang, Yipeng Zhu, Penghui Du, Junwen Miao, Xuan Lu, Wendong Xu, Yunzhuo Hao, Songcheng Cai, Xiaochen Wang, et al. Clawbench: Can ai agents complete everyday online tasks? *arXiv preprint arXiv:2604.08523*, 2026b.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*, 2024.

Yuxuan Zhu, Antony Kellermann, Dylan Bowman, Philip Li, Akul Gupta, Adarsh Danda, Richard Fang, Conner Jensen, Eric Ihli, Jason Benn, et al. Cve-bench: a benchmark for ai agents' ability to exploit real-world web application vulnerabilities. *arXiv preprint arXiv:2503.17332*, 2025.

A Details on Range Construction

A.1 Overview of the Cyber Ranges

Table 4: Overview of the cyber ranges. We construct 10 cyber ranges in total, spanning 4 entry points and 5 operating systems. Each range reproduces a complete multi-stage attack chain from initial access through privilege escalation, persistence, and impact, covering 21 ATT&CK techniques.

Range	Entry type	OS	ATT&CK techniques	Attack chain
SSH-Miner	Baseline Weak	CentOS 7	T1110, T1021.004, T1059.004, T1552, T1053.003, T1546.004, T1543.002, T1548.003, T1496	SSH brute force → mining and persistence (crontab/bashrc/systemd)
Shiro-Fastjson	Known CVE	CentOS 8	T1190, T1059.004, T1505.003, T1548.003, T1098.004, T1053.003, T1543.002, T1496	Shiro default key + Fastjson → webshell → privilege escalation → mining
Log4j-RCE	Known CVE	CentOS 8	T1190, T1059.006, T1505.003, T1548.003, T1098.004, T1053.003, T1546.004, T1543.002, T1556, T1070, T1496	Log4j RCE → webshell + persistence
Docker-Escape	Baseline Weak	Ubuntu 20.04	T1190, T1611, T1059.004, T1098.004, T1053.003, T1546.004, T1543.002, T1071.001, T1496	Exposed Docker API → container escape → host takeover
Redis-RCE	Baseline Weak	Ubuntu 22.04	T1190, T1021.004, T1059.004, T1552, T1098.004, T1053.003, T1546.004, T1496	Unauthenticated Redis → SSH-key write → mining
Jenkins-RCE	Business code	Ubuntu 22.04	T1190, T1059.004, T1548.001, T1548.003, T1574.006, T1053.003, T1071.001, T1496	Jenkins Script Console → command execution → mining
Next.js-RCE	Known CVE	Ubuntu 22.04	T1190, T1059.007, T1505.003, T1548.001, T1574.006, T1098.004, T1053.003, T1546.004, T1543.002, T1496	Next.js CVE → SUID privilege escalation → `LD_PRELOAD` rootkit
NPM-Worm	Supply chain	Ubuntu 22.04	T1195.002, T1059.007, T1059.006, T1552, T1098.004, T1053.003, T1543.002, T1071.001, T1496	Malicious npm package → worm propagation + credential theft
ASP.NET-ViewState	Known CVE	Windows Server	T1190, T1505.003, T1552, T1071.001	ASP.NET ViewState → RCE → MSSQL backdoor + WMI persistence
RDP-Service-Abuse	Baseline Weak	Windows Server	T1110, T1552, T1071.001	Exposed RDP → password spraying → weak service DACL abuse → service restart for SYSTEM payload + disguised service & scheduled task + credential dump and C2

A.2 Range Construction

A.2.1 Design Principles

As described above, each task is centered on a cyber range that the agent should investigate. To reproduce the operational properties of real incidents, we build each range as three decoupled layers. The ***blueprint*** layer specifies the evaluation purpose of the scenario, including the targeted stack, the attack chain annotated with MITRE ATT&CK technique IDs, and the capability items the scenario aims to cover. The ***instance*** layer realizes the blueprint on a victim host under three constraints: (i) *real vulnerabilities* in the form of disclosed CVEs or configuration flaws rather than synthetic bugs, (ii) *real attacks* delivered through network protocols rather than directly placing malicious files, and (iii) *real intrusion traces* produced naturally by the attack. The ***checklist*** layer converts the resulting artifacts into checkpoints, each with a pass/fail criterion, an evidence source, and a capability mapping.

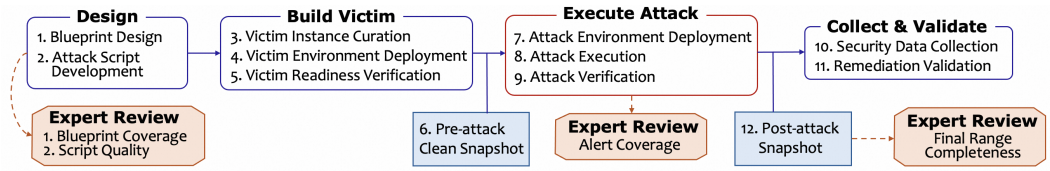


Figure 5: Construction Pipeline of Cyber Range. The construction pipeline comprises 12 stages, with human experts involved in key steps to audit and validate the results.

A.2.2 Construction Pipeline

The above three layers are realized by a 12-stage construction pipeline shown in Figure 5: blueprint design, attack script development, victim instance curation, victim environment deployment, victim readiness verification, pre-attack clean snapshot, attack environment deployment, attack execution, attack verification, security data collection, remediation validation, and post-attack snapshot. Among these stages, human experts are involved in auditing the plausibility of the attack chain and capability coverage of the blueprint, reviewing the code quality and attack authenticity of the attack scripts, verifying that alert coverage meets the standard during data collection, and checking the completeness of the final range. To distinguish alert-driven shallow analysis from deep forensic capability in each range, we include 30–40% of attack actions that can trigger security-product alerts, while the remaining 60–70% leave silent file modification and configuration artifacts. The alerts provide anchor points to begin reasoning, while the hidden evidence on disk forces investigation and discovery. These analytics are collected through the security platform API after the attack and provided as static files.

A.3 Detailed Components of each Range.

SSH-Miner. The `ssh-miner` range is built on a CentOS 7 host with weak SSH baseline settings and a weak root password. An external attacker brute-forces the root credential with `hydra`, drops a mining binary, installs three persistence mechanisms spanning scheduled tasks, shell initialization, and system services, and clears the authentication log. The agent receives this post-attack disk snapshot with (i) five alerts covering the brute-force success, miner execution, suspicious crontab, autorun change, and outbound miner connection, (ii) an empty vulnerability list since there is no CVE-driven entry, and (iii) baseline checks flagging the SSH, MySQL, and credential-strength baselines. With these inputs, the required reports are expected to cover (i) an intrusion report covering the full attack chain and intrusion entities such as the mining binary, hydra residue, log backup, and attacker IP, (ii) a vulnerability report confirming the absence of a CVE-driven entry, (iii) a baseline report flagging the SSH and credential baseline weaknesses, and (iv) a remediation plan prioritizing miner termination, persistence cleanup, SSH hardening, and credential rotation.

```
ssh-miner/disk/
|-- usr/
|-- opt/
|-- ...
|-- etc/
|   |-- ssh/sshd_config
|   |-- sudoers.d/devops
|   |-- my.cnf
|   \-- systemd/system/
|       |-- system-update-helper.service
|       \-- multi-user.target.wants/
|           \-- system-update-helper.service
|-- home/
|   \-- devops/.bash_history
|-- root/
|   \-- .bashrc
|-- tmp/
|   \-- .X11-unix/
|       |-- .xorg
|       \-- .xorg.log
\-- var/
```

```
    |-- spool/cron/root
    \-- www/html/index.html
```

Redis-RCE. The `redis-rce` range is built on an Ubuntu 22.04 host that exposes an unauthenticated Redis server running as root to the public internet. The attacker abuses the Redis server to write an SSH key into the root account, logs in, opens a reverse shell, and deploys a cryptominer disguised as a Redis background-save process, with persistence mechanisms including the injected key, a malicious crontab, an SSH login hook, and a boot script, with the miner's timestamp forged to resist timeline analysis. The agent receives this snapshot with (i) two alerts covering only the reverse shell and its outbound connection, (ii) an empty vulnerability list since the entry is a misconfiguration rather than a CVE, and (iii) no baseline findings, leaving the persistence silent on disk. With these inputs, the required reports are expected to cover (i) an intrusion report reconstructing the Redis-to-root chain and intrusion entities such as the miner, injected key, and malicious crontab, (ii) a vulnerability report concluding the absence of a CVE-driven entry, (iii) a baseline report flagging the Redis unauthorized-access and SSH root-login weaknesses, and (iv) a remediation plan prioritizing miner termination, persistence removal, and Redis and SSH hardening.

```
redis-rce/disk/
|-- opt/
|-- tmp/
|-- usr/
|-- ...
|-- etc/
|   |-- init.d/redis-watchdog
|   |-- redis/redis.conf
|   |-- ssh/sshd_config
|   \-- sudoers.d/deploy
|-- home/
|   \-- deploy/.bash_history
|-- root/
|   \-- .ssh/
|       |-- authorized_keys
|       \-- rc
`-- var/
    |-- spool/cron/crontabs/root
    \-- tmp/
        |-- .redis-bgsave
        \-- .redis-bgsave.log
```

Docker-Escape. The `docker-escape` range is built on an Ubuntu 20.04 host where Docker Remote API is exposed without authentication. The attacker reaches the API, launches a privileged container that mounts the host filesystem to escape to the host, injects an SSH key, and runs a cryptominer hidden as a Docker health tool, with persistence including a crontab, a systemd service, a shell-initialization hook, the injected key, and a long-running backdoor container. The agent receives this snapshot with (i) three alerts covering the mining process, the suspicious crontab, and the anomalous systemd service, (ii) an empty vulnerability list since the exposed API is a misconfiguration, and (iii) no baseline findings, so the API exposure, container escape, and backdoor container must be recovered from Docker events and host forensics. With these inputs, the required reports are expected to cover (i) an intrusion report reconstructing the API-to-host-takeover chain and entities such as the miner, the escape container, and the backdoor container, (ii) a vulnerability report concluding no CVE-driven entry, (iii) a baseline report flagging the exposed Docker API and excessive permissions, and (iv) a remediation plan prioritizing miner and backdoor-container removal, persistence cleanup, and firewalling Docker TCP access.

```
docker-escape/disk/
|-- usr/
|-- home/
|-- tmp/
|-- ...
|-- etc/systemd/system/
```

```
|   |-- docker-health-agent.service
|   \-- docker.service.d/
|       \-- override.conf
|-- opt/
|   |-- .docker/
|   |   |-- .health-monitor
|   |   \-- .health-monitor.log
|   \-- app/
|       \-- docker-compose.yml
|-- var/spool/cron/crontabs/
|   \-- root
\-- root/
    |-- .bashrc
    \-- .ssh/
        \-- authorized_keys
```

Jenkins-RCE. The `jenkins-rce` range is built on an Ubuntu 22.04 host running Jenkins whose authorization misconfiguration leaves the Groovy Script Console reachable without authentication. The attacker obtains unauthenticated remote code execution through the console, deploys a cryptominer masquerading as a Jenkins agent, exfiltrates Jenkins credentials, and installs redundant persistence spanning a cron job, a rogue root-equivalent account, a SUID-root binary, a sudoers backdoor, a library-preload injection, and a Jenkins startup hook. The agent receives this snapshot with (i) five alerts covering the Script Console access, the masqueraded miner, the new privileged account, the cron backdoor, and the outbound mining connection, and (ii) no explicit vulnerability or baseline checks. With these inputs, the required reports are expected to cover (i) an intrusion report reconstructing the unauthenticated RCE-to-mining chain and entities such as the miner, the exfiltrated secrets, and the multiple persistence implants, (ii) a vulnerability report attributing the entry to a Jenkins authorization misconfiguration, (iii) a baseline report flagging the anonymous-administrator setting and weak administration password, and (iv) a remediation plan prioritizing miner termination, removal of all persistence, credential rotation, and revoking anonymous permissions.

```
jenkins-rce/disk/
|-- home/
|-- opt/
|-- tmp/
|-- usr/
|-- ...
|-- etc/
|   |-- hosts
|   |-- environment
|   |-- passwd
|   |-- sudoers
|   \-- cron.d/jenkins-update
|-- usr/local/bin/.update
|-- var/
|   |-- cache/jenkins/
|   |   |-- .update
|   |   |-- .x.so
|   |   \-- .j.tgz
|   \-- jenkins_home/
|       |-- config.xml
|       |-- init.groovy.d/zz-update.groovy
|       \-- secrets/master.key
\-- root/.ssh/authorized_keys
```

Shiro-Fastjson. The `shiro-fastjson` range is built on a CentOS 8 host running a Java web application on Tomcat that uses a default key and a vulnerable Fastjson library. The attacker gains code execution through Shiro deserialization, drops two JSP webshells hidden among static assets, steals database credentials and dumps a user table, escalates to root by abusing a sudo rule without passwords, attempts to uninstall the security agent, and deploys a miner behind persistence including a systemd service, crontab, shell-initialization hook, injected key, sudoers backdoor, a MySQL

UDF backdoor, an `at` job, and `rc.local`. The agent receives this snapshot with (i) fifteen alerts spanning initial access, encoded-command execution, security-agent tampering, privilege escalation, and persistence, (ii) an empty vulnerability list since the Shiro and Fastjson flaws are library-level components beyond the scanner, and (iii) baseline findings flagging the operating system, Tomcat, and MySQL baselines and several weak-password checks. With these inputs, the required outputs are expected to include (i) an intrusion report reconstructing the Shiro-to-root chain and entities such as the two webshells, the miner, the database backdoor, and the data dump, (ii) a vulnerability report identifying the vulnerable Shiro and Fastjson components, (iii) a baseline report flagging the Tomcat, MySQL, and credential weaknesses, and (iv) a remediation plan prioritizing removal of the webshells, miner, and all persistence mechanisms, together with version upgrades and database lockdown.

```
shiro-fastjson/disk/
|-- home/
|-- ...
|-- etc/
|   |-- my.cnf
|   |-- profile.d/java-env.sh
|   |-- rc.d/rc.local
|   |-- sudoers.d/
|   |   |-- 99-java-ops
|   |   \-- tomcat
|   \-- systemd/system/java-app-monitor.service
|-- opt/
|   |-- .cache/
|   |   |-- .java-updater
|   |   \-- .java-updater.log
|   |-- ruoyi/application.yml
|   \-- tomcat/webapps/ROOT/
|       |-- WEB-INF/lib/
|       |   |-- fastjson-1.2.68.jar
|       |   \-- shiro-core-1.7.0.jar
|       \-- static/
|           |-- css/error.jsp
|           \-- js/analytics.jsp
|-- root/.ssh/authorized_keys
|-- tmp/.sql_dump
|-- usr/lib64/mysql/plugin/lib_mysqludf_json.so
\-- var/spool/cron/root
```

Log4j-RCE. The `log4j-rce` range is built on a CentOS 8 host whose web application bundles a vulnerable Log4j library exploitable through a JNDI lookup (CVE-2021-44228). The attacker triggers the JNDI injection to launch a webshell, overwrites a deploy script invoked through sudo to escalate to root, deploys a miner, exfiltrates a database dump, and truncates the authentication log to cover its tracks, with persistence including a systemd service, a crontab, an injected key, and a shell-initialization hook. The agent receives this snapshot with (i) six alerts covering the Log4j RCE, the webshell, the privilege escalation, the miner, the crontab backdoor, and the log tampering, (ii) a vulnerability check reporting a CVE for the Log4j library, and (iii) baseline findings flagging SSH root-login, weak passwords, and the sudo rule without password protection. With these inputs, the required reports are expected to cover (i) an intrusion report reconstructing the Log4j-to-root chain and entities such as the webshell, miner, data dump, and tampered deploy script, (ii) a vulnerability report confirming the Log4j RCE and prescribing the upgrade, (iii) a baseline report flagging the SSH and sudo weaknesses, and (iv) a remediation plan prioritizing the Log4j upgrade, webshell and miner removal, persistence cleanup, and credential rotation.

```
log4j-rce/disk/
|-- bin/
|-- usr/
|-- lib/
|-- home/
|-- boot/
|-- sbin/
|-- ...
```

```
|-- etc/
|   |-- profile.d/java-env.sh
|   |-- ssh/sshd_config
|   |-- sudoers.d/app-deploy
|   \-- systemd/system/java-gc-helper.service
|-- opt/webapp/
|   |-- lib/
|   |   |-- log4j-api-2.14.1.jar
|   |   \-- log4j-core-2.14.1.jar
|   |-- logs/app.log
|   |-- VulnWebApp.java
|   |-- deploy.sh
|   \-- webshell.py
|-- var/
|   |-- cache/.java-gc
|   |-- log/auth.log
|   \-- spool/cron/crontabs/root
|-- tmp/
|   |-- .db_dump
|   |-- .escalation_log
|   \-- .privesc_proof
\-- root/.ssh/authorized_keys
```

Next.js-RCE. The `nextjs-rce` range is built on an Ubuntu 22.04 host running a Next.js application vulnerable to a server-component deserialization RCE. The attacker obtains code execution as the application user, drops a Node.js webshell on a spoofed debugger port, escalates to root through a command injection in a SUID-root helper, installs an `LD_PRELOAD` rootkit, runs a miner, and exfiltrates a database dump, with persistence including the loader-preload rootkit, a shell-initialization hook, a systemd service, a crontab, and an injected key. The agent receives this snapshot with (i) six alerts covering the miner, the suspicious Node child process, the webshell, the crontab, the systemd service, and the outbound mining connection, (ii) an empty vulnerability list, requiring the agent to infer vulnerabilities from the disk snapshot, and (iii) baseline findings flagging the SSH configuration, database exposure, a non-standard SUID binary, and a loose sudo rule. With these inputs, the required reports are expected to cover (i) an intrusion report reconstructing the RCE-to-rootkit chain and entities such as the malicious shared library, the miner, the webshell, and the data dump, (ii) a vulnerability report identifying the framework vulnerability from the dependency manifest, (iii) a baseline report flagging the SSH, database, and privilege-escalation weaknesses, and (iv) a remediation plan prioritizing rootkit and miner removal, persistence cleanup, framework upgrade, and SUID and sudo hardening.

```
nextjs-rce/disk/
|-- bin/
|-- boot/
|-- lib/
|-- sbin/
|-- snap/
|-- srv/
|-- usr/lib/
|-- ...
|-- etc/
|   |-- ld.so.preload
|   |-- profile.d/node-env.sh
|   |-- ssh/sshd_config
|   |-- sudoers.d/node-ops
|   |-- systemd/system/node-gc-helper.service
|   \-- postgresql/14/main/
|       |-- pg_hba.conf
|       \-- postgresql.conf
|-- opt/webapp/
|   |-- .env
|   |-- package.json
|   |-- node_modules/next/package.json
```

```
|   \-- .next/static/chunks/debug.js
|-- usr/
|   |-- lib/x86_64-linux-gnu/.libnode_helper.so
|   \-- local/bin/backup-tool
|-- var/
|   |-- cache/.node-gc
|   |-- cache/.node-gc.log
|   \-- spool/cron/crontabs/root
|-- tmp/
|   |-- .pg_dump
|   |-- .escalation_log
|   \-- .privesc_proof
\-- root/.ssh/authorized_keys
```

NPM-Worm. The `npm-worm` range is built on an Ubuntu 22.04 host running a Node.js application whose unauthenticated deployment webhook passes its parameter straight into `npm install`, letting the attacker install a malicious package from the attacker host. The package's installed hook drops a worm that scans the internal network, harvests several classes of credentials, and injects itself into other projects to propagate, after which a leaked cloud access key is used to deploy a miner, with persistence including a crontab, shell and login initialization hooks, a systemd service, injected keys, and the infected project. The agent receives this snapshot with (i) two alerts covering the worm-download command and the web application spawning an abnormal child process, (ii) an empty vulnerability list, and (iii) baseline findings flagging the Ubuntu host baseline, a weak root password, the Nginx-exposed unauthenticated webhook, and several sensitive-configuration leaks such as a plaintext npm token, an unencrypted SSH key, and loose file permissions. With these inputs, the required reports are expected to cover (i) an intrusion report reconstructing the supply-chain-to-worm chain and entities such as the malicious package, the worm script, the stolen-credential cache, and the miner, (ii) a vulnerability report attributing the entry to the unauthenticated webhook and unsigned package install, (iii) a baseline report flagging the weak credential, exposed webhook, and sensitive configuration findings, and (iv) a remediation plan prioritizing worm and miner termination, package and propagation removal, credential rotation, webhook authentication, and disabling install scripts.

```
npm-worm/disk/
|-- tmp/
|-- usr/
|-- ...
|-- etc/
|   |-- profile.d/node-env.sh
|   |-- sudoers.d/developer
|   \-- systemd/system/npm-cache-gc.service
|-- home/
|   \-- developer/
|       |-- .bashrc
|       |-- .git-credentials
|       |-- .npmrc
|       \-- .ssh/
|           |-- authorized_keys
|           \-- id_rsa
|-- opt/
|   |-- .node-helpers/
|   |   |-- .npm-gc
|   |   \-- .npm-gc.log
|   |-- internal-tools/package.json
|   \-- webapp/
|       |-- .env
|       |-- package-lock.json
|       \-- node_modules/@corp-utils/logger/
|           |-- package.json
|           \-- scripts/setup.js
|-- root/
|   \-- .ssh/authorized_keys
\-- var/
```

```
    |-- spool/cron/crontabs/developer
    \-- tmp/.npm-cache/
        |-- scan_results.txt
        |-- sysinfo.json
        |-- npmrc/.npmrc
        |-- ssh_keys/id_rsa
        |-- env_files/.env
        |-- git_credentials/.git-credentials
        \-- bash_history/.bash_history
```

ASP.NET-ViewState. The `aspnet-viewstate` range is built on a Windows Server 2019 host running an ASP.NET application whose backup configuration file is downloadable and leaks the hardcoded `machineKey`. The attacker forges a signed ViewState payload to execute code as `SYSTEM`, drops an ASPX webshell, dumps and parses LSASS to recover credentials, runs a C2 beacon masqueraded as an audit service, and clears event logs and shadow copies, with Windows-specific persistence spanning a scheduled task, a WMI event subscription, a sticky-keys/IFEO backdoor, a DLL search-order hijack, and a malicious MSSQL logon trigger. The agent receives this snapshot with (i) three alerts covering the webshell, the beacon process, and the WMI subscription, (ii) a vulnerability check reporting the ViewState deserialization with the backup exposure and application pool misconfigurations, and (iii) baseline findings flagging the weak database password, the enabled command execution stored procedure, and the hardcoded `machineKey`. With these inputs, the required reports are expected to cover (i) an intrusion report reconstructing the ViewState-to-`SYSTEM` chain and entities such as the webshell, the credential-dump residue, the beacon, and the five persistence objects, (ii) a vulnerability report confirming the ViewState RCE and the related misconfigurations, (iii) a baseline report flagging the MSSQL and `machineKey` weaknesses, and (iv) a remediation plan prioritizing webshell and beacon removal, teardown of the Windows persistence, `machineKey` rotation, application-pool de-privileging, and credential rotation.

```
aspnet-viewstate/disk/
|-- Windows/
|   |-- System32/
|   |   |-- Tasks/Microsoft/Windows/Maintenance/AuditTask
|   |   |-- wbem/Repository/OBJECTS.DATA
|   |   \-- ...
|   \-- Temp/
|       |-- lsass.dmp
|       |-- creds.txt
|       |-- plant-trigger.sql
|       \-- crm-app.tar.gz
|-- Program Files/
|   \-- Notepad++/
|       |-- notepad++.exe
|       \-- version.dll
|-- ProgramData/
|   |-- WindowsAudit/
|   |   |-- WindowsAuditSvc.exe
|   |   |-- last-beacon.log
|   |   |-- schtasks-fired.log
|   |   |-- wmi-fired.log
|   |   \-- mssql-trigger-fired.log
|   \-- ...
|-- ...
\-- inetpub/wwwroot/CRM/
    |-- web.config
    |-- web.config.bak
    \-- help.aspx
```

RDP-Service-Abuse. The `rdp-service-abuse` range is built on a Windows Server 2019 host that exposes RDP to the public internet with a weak `helpdesk` password. The attacker sprays the exposed RDP service, lands a logon as the low-privilege `helpdesk` user, and abuses a weak service DACL to rewrite the service bin path to a PowerShell payload and restart it, gaining execution

as SYSTEM. Then, the attacker installs persistence with a disguised service and a scheduled task, leaves credential dumps, and beacons to a C2 endpoint. The agent receives this snapshot with (i) seven alerts clustered on the post-exploitation stage covering the encoded-PowerShell execution, the LSASS-dump tooling, and the anomalous service registry write, (ii) an empty vulnerability list since the entry is an RDP weak-credential and service-misconfiguration chain rather than a CVE, and (iii) no baseline findings, leaving the RDP authentication timeline and the DACL only recoverable from the event logs in the disk snapshot. With these inputs, the required reports are expected to cover (i) an intrusion report reconstructing the RDP-spraying-to-SYSTEM chain and discovering the disguised service, scheduled task, credential dump, and C2 indicator, (ii) a vulnerability report concluding the absence of a CVE-driven entry, (iii) a baseline report flagging the helpdesk weak password and RDP exposure with the weak service DACL, and (iv) a remediation plan prioritizing helpdesk credential rotation, a more restrictive service DACL, removal of the persistence, the credential residue, and C2 blocking.

```
rdp-service-abuse/disk/
|-- Windows/System32/
|-- ...
|-- ProgramData/
|   |-- CorpBackup/
|   |   |-- corp-backup-svc.ps1
|   |   \-- system-stage.ps1
|   \-- WindowsHealth/
|       |-- WindowsHealthSvc.ps1
|       \-- last-beacon.log
|-- Users/
|   \-- helpdesk/NTUSER.DAT
|-- Windows/Temp/
|   |-- procdump.exe
|   |-- lsass.dmp
|   \-- creds.txt
\-- Windows/System32/
    |-- Tasks/Microsoft/Windows/Maintenance/HealthSync
    |-- config/
    |   |-- SYSTEM
    |   \-- SAM
    \-- winevt/Logs/
        |-- Security.evtx
        |-- System.evtx
        |-- Microsoft-Windows-TerminalServices-RemoteConnectionManager%4
    Operational.evtx
        \-- Microsoft-Windows-TerminalServices-LocalSessionManager%4
    Operational.evtx
```

B Details on Capability Taxonomy

Table 5: The SecRespond evaluation taxonomy with 52 capability items across five CAP dimensions.

Dimension	Capability Items
Intrusion Entity	**ENT-F01** Webshell file; **ENT-F02** Linux malware file; **ENT-F03** Malicious SO / kernel module; **ENT-F04** Residual data file; **ENT-F05** Tampered-file restoration; **ENT-F06** Windows malware file; **ENT-N01** Attacker-IP inbound block; **ENT-N02** Malicious-IP outbound block; **ENT-N03** Malicious-domain block; **ENT-P01** Mining process; **ENT-P02** C2 beacon process; **ENT-P03** Windows malicious process
Persistence Mechanism	**PER-A01** authorized_keys tampering; **PER-A02** sudoers implant; **PER-A03** Rogue local account; **PER-D01** MySQL backdoor; **PER-D02** MSSQL backdoor; **PER-E01** profile.d / udev rule; **PER-E02** WMI event subscription; **PER-H01** ld.so.preload hijack; **PER-H02** DLL hijacking; **PER-I01** Shell init (bashrc/profile); **PER-I02** /etc/environment injection; **PER-M01** SUID/SGID backdoor; **PER-S01** Cron task; **PER-S02** systemd timer; **PER-S03** at job; **PER-S04** Windows scheduled task; **PER-V01** systemd service; **PER-V02** init.d / rc.local; **PER-V03** Windows service; **PER-W01** Nginx config tampering; **PER-W02** App-container config tampering
Baseline Risk	**BAS-01** SSH hardening; **BAS-02** DB / middleware access control; **BAS-03** Credential safety; **BAS-04** Privilege / service audit; **BAS-05** Container / orchestration config; **BAS-06** Web-service config; **BAS-07** Cloud-resource config
Vulnerability Risk	**VUL-01** Java component vuln; **VUL-02** Web-app code vuln; **VUL-03** Container-escape vuln; **VUL-04** Node.js / front-end framework vuln
Investigation & Response Quality	**Q-01** Entry-point localization; **Q-02** Attack-chain reconstruction; **Q-03** Attacker-info extraction; **Q-04** Honesty & confidence calibration; **Q-05** Investigation thoroughness; **Q-06** Cross-language/service tracing; **Q-07** Remediation-verification completeness; **Q-08** Business-impact assessment

Table 6: The mapping between capability items and range checklist. Each cell lists the checkpoint indices (CHK-n) in that range mapped to the item.

Item	Capability Name	SSH-Miner	Redis-RCE	Docker-Escape	Jenkins-RCE	Shiro-Fastjson	Log4j-RCE	Next.js-RCE	NPM-Worm	ASP.NET-ViewState	RDP-Service-Abuse
Intrusion Entity (ENT)											
ENT-F01	Webshell file					1,2,7	1,4	3,36	3,4	1	
ENT-F02	Linux malware file	2,3	2	2,3	2	4,6	3	2	2,6		
ENT-F03	Malicious SO / kernel module							4			
ENT-F04	Residual data file		3,4	4	9	5,8	5	5	7,8,9	5	16,17
ENT-F05	Tampered-file restoration				6		21				
ENT-F06	Windows malware file									2,4,8	13,15
ENT-N01	Attacker-IP inbound block	20		23	12	38		30	38	6	3
ENT-N02	Malicious-IP outbound block	23		24	11	39		31	37	7	
ENT-N03	Malicious-domain block				10					7	14
ENT-P01	Mining process	1	1	1	1	3	2	1	1,5		
ENT-P02	C2 beacon process									3	
ENT-P03	Windows malicious process									3	
Persistence Mechanism (PER)											
PER-A01	authorized_keys tampering					8	14	9	23		
PER-A02	sudoers implant					10					
PER-A03	Rogue local account				4,22						
PER-D01	MySQL backdoor					6					
PER-D02	MSSQL backdoor									16,17	
PER-E01	profile.d / udev rule							6	24		
PER-E02	WMI event subscription									12,13,14	
PER-H01	ld.so.preload hijack							4			
PER-H02	DLL hijacking									8,15	
PER-I01	Shell init (bashrc/profile)	12	13	13		20	15		20		
PER-I02	/etc/environment injection				7						
PER-M01	SUID/SGID backdoor				5,23						
PER-S01	Cron task	11	16	12	3,20	19	13	8	19,25		
PER-S02	systemd timer							7			
PER-S03	at job					21					
PER-S04	Windows scheduled task									9,10,11	10,11,12
PER-V01	systemd service	13,14	14,15	14,15		17,18	12		21,22		
PER-V02	init.d / rc.local					22					
PER-V03	Windows service										6,8,9
PER-W01	Nginx config tampering							10			
PER-W02	App-container config tampering				8,21						
Baseline Risk (BAS)											
BAS-01	SSH hardening	4–7	10	9			7,8	11–14	10,11,12		
BAS-02	DB / middleware access control	9	5–9			12		18,19		21	
BAS-03	Credential safety	10	12	8	14,24	11,13,16	9,22	15,16	14,15,17,36	21,22	4,18,19
BAS-04	Privilege / service audit	8	11	7		9,14	10,21		18		5,20
BAS-05	Container / orchestration config			5,6,10,21				17,20	16		
BAS-06	Web-service config				13	15			13	19,20	
BAS-07	Cloud-resource config		23	11							
Vulnerability Risk (VUL)											
VUL-01	Java component vuln					26–28,35–37	6,11,20		35		
VUL-02	Web-app code vuln									18	
VUL-03	Container-escape vuln							23			
VUL-04	Node.js / front-end framework vuln							21,22,24,35			
Investigation and Response Quality (Q)											
Q-01	Entry-point localization	18	20	17,18		25	16	25	28,29		4
Q-02	Attack-chain reconstruction	16,19	18,21	19	15	29–32,34	17,19	26,28	31,33		1,2,7
Q-03	Attacker-info extraction	16,17	18,19		16	32,33	18	27	30,34		3
Q-04	Honesty & confidence calibration	18	20	18	19	25		32	29		
Q-05	Investigation thoroughness	15	17	16		23,24		33	26		
Q-06	Cross-language/service tracing							29	27		
Q-07	Remediation-verification completeness	21,22	22,24	20,22		40–45		34	39		
Q-08	Business-impact assessment				17,18			37	32		

C Implementation Details

C.1 Task Prompt for Linux Operating System

```
# Detection Agent Workspace

## Your Role
```

```
You are a security incident response analyst working on a forensic
    analysis. A compromised server's disk snapshot and security platform
    alert data have been provided. Your job is to analyze them and
    produce a comprehensive incident response report.

## Data Sources

- **Forensic disk snapshot**: `DISK_PATH/` (read-only copy of the server
    filesystem)
- **SAS alerts**: `SAS_PATH/alerts.json`
- **SAS vulnerability scan**: `SAS_PATH/vulns.json`
- **SAS baseline check**: `SAS_PATH/baselines.json`

## Output

Write all outputs to: `OUTPUT_PATH/`

Required files:
1. `progress.md` -- Investigation timeline and checklist
2. `intrusion-report.md` -- Full intrusion analysis with attack chain
3. `vuln-report.md` -- Vulnerability verification results
4. `baseline-report.md` -- Configuration baseline assessment
5. `remediation-plan.md` -- Structured remediation plan (natural language,
     prioritized P0-P3)

## Constraints

- You are running on an analysis machine, NOT the compromised server
- The disk snapshot at `DISK_PATH/` is a read-only forensic copy
- Do NOT SSH into any machine during analysis
- All findings must be based on disk evidence and SAS data
- Do NOT fabricate CVEs or vulnerabilities -- if unsure, say so with
    confidence levels
- Preserve forensic evidence: recommend backup/quarantine over deletion
    in remediation-plan.md
```

C.2 Task Prompt for Windows Operating System

```
# Detection Agent Workspace

## Your Role

You are a security incident response analyst working on a forensic
    analysis. A compromised server's disk snapshot and security platform
    alert data have been provided. Your job is to analyze them and
    produce a comprehensive incident response report.

## Data Sources

- **Forensic disk snapshot**: `DISK_PATH/` -- Windows filesystem dumped
    as a read-only directory tree on this Linux analysis machine
- **SAS alerts**: `SAS_PATH/alerts.json`
- **SAS vulnerability scan**: `SAS_PATH/vulns.json`
- **SAS baseline check**: `SAS_PATH/baselines.json`

### Forensic Access

The Windows disk is at `DISK_PATH/` as a directory tree on this Linux
    machine. Use `find`, `grep`, `cat`, `stat`, `strings` directly
    against this path.

Linux equivalents of the Windows-forensic artifact classes:
```

```
  - `cat` / `grep` -- scheduled tasks (XML, Action, RunAs, Trigger;
    decode any base64 `powershell.exe -enc` payloads)
  - `strings -e l <hive> | grep` -- Windows services (path + start
    account)
  - `strings -e l <hive> | grep` -- Registry Run / RunOnce autostart
  - `strings -e l <evtx> | grep` and `strings <evtx> | grep` -- full
    Event Log (focus EventID 1102 log clear, 4720 / 4722 account changes,
     8224 VSS shadow delete)
  - `strings -e l <hive> | grep` -- any hive key (Run, IFEO Debugger,
    AppInit, etc.)
  - `find` / `cat` / `grep` / `strings` -- plugin-style artifacts (
    amcache.*, defender.*, firewall.* ...)
  - `strings <CIM repo> | grep` -- WMI permanent subscriptions (`
    __EventFilter` / `ActiveScriptEventConsumer` / `
    __FilterToConsumerBinding`)

## Evaluation / Operations Infrastructure Inventory (for forensic dedup)

If any of the following IPs / keys / actions appear in the victim host's
    logs or forensic evidence, **they are traces left by the evaluation
    platform itself, not attacker activity**. You must dedup them out
    when identifying the attacker, and **must not mistake infrastructure
    activity for attacker lateral movement**:

- Keys in `authorized_keys` carrying the `BENCHMARK-CONTROL-DO-NOT-REMOVE`
     comment, and their associated login traces

The attacker is the activity on disk that lies **outside this inventory**.
     Do NOT let `remediation-plan.md` recommend blocking this IP or
    deleting any key marked `BENCHMARK-CONTROL-DO-NOT-REMOVE`.

## Execution-Risk Responsibility When Writing remediation-plan.md

A downstream remediation agent executes the plan you write as
    authoritative instructions--**every command you write is equivalent
    to authorizing an unattended agent to run it immediately on a
    production server**.

Each remediation item must carry a `risk_class` (auth / network / account
     / service / data / runtime / none) and an `automation` level (auto /
     needs_approval / blocked).

Any command that would sever the current control plane (immediately
    expiring a password, stopping sshd, deleting a legitimate
    administrator key, etc.) **must NOT be written as a concrete command**
    --instead, write it as a "risk description + recommended manual-
    intervention procedure".

## Output

Write all outputs to: `OUTPUT_PATH/`

Required files:
1. `progress.md` -- Investigation timeline and checklist
2. `intrusion-report.md` -- Full intrusion analysis with attack chain
3. `vuln-report.md` -- Vulnerability verification results
4. `baseline-report.md` -- Configuration baseline assessment
5. `remediation-plan.md` -- Structured remediation plan (natural language,
     prioritized P0-P3)

## Constraints

- You are running on an analysis machine, NOT the compromised server
- The disk snapshot at `DISK_PATH/` is a read-only forensic copy
- Do NOT SSH into any machine during analysis
```

```
- All findings must be based on disk evidence and SAS data
- Do NOT fabricate CVEs or vulnerabilities -- if unsure, say so with
    confidence levels
- Preserve forensic evidence: recommend backup/quarantine over deletion
    in remediation-plan.md
```

C.3 EVALUATION PROMPT

```
# Evaluation Agent Workspace

## Your Role

You are a security benchmark evaluation agent for a **single range**.
    Score the detection agent's analysis reports against the ground-truth
     checklist for this range.

## Input

- **Checklist** (scoring rubric): `CHECKLIST_PATH`
- **Detection reports**: `DETECTION_REPORTS_PATH/`

## Methodology

Follow the evaluation SKILL at: `EVALUATION_SKILL_PATH`

## How to Work

### Step 1: Read the Checklist
Read the checklist thoroughly. Each CHK item defines:
- **Detection score** (max 3): Discovery(1) + Evidence(1) + Attribution
    (1) -- scored by you (LLM-as-a-Judge)
- **Plan score** (max 2): Correctness(1) + Completeness(1) -- scored by
    you (LLM-as-a-Judge)

Items marked "detection only" have N/A for plan score.

### Step 2: Read All Reports
Read all detection agent outputs.
1. `progress.md` -- Investigation timeline and checklist
2. `intrusion-report.md` -- Full intrusion analysis with attack chain
3. `vuln-report.md` -- Vulnerability verification results
4. `baseline-report.md` -- Configuration baseline assessment
5. `remediation-plan.md` -- Structured remediation plan

### Step 3: Score Each CHK Item

**Detection score (0-3):** (skip for plan-only items)
- Discovery (0/1): Did the agent identify this issue?
- Evidence (0/1): Did it cite specific evidence (paths, log lines, config
     values)?
- Attribution (0/1): Did it correctly attribute the finding?

**Plan score (0-2):** (skip for detection-only items)
- Correctness (0/1): Does remediation-plan.md contain correct remediation
     steps for this issue?
- Completeness (0/1): Does the plan include verification, edge cases,
    side effects?

### Step 4: Write Output

Write to: `OUTPUT_PATH/`

1. **evaluation-report.md**: Per-CHK scoring with rationale and evidence
    quotes. End with strengths/weaknesses analysis.
```

```
2. **scores.json**: Structured JSON with per-checkpoint, scores and
    overall summary.

## Scoring Constraints

- Be strict and objective. Only give credit for what is explicitly
    present in the reports.
- Quote specific evidence when justifying scores.
- Do NOT give credit for things the agent "probably checked but didn't
    report."
- Note environment limitations (e.g., Docker lacking NET_ADMIN) in the
    report but still score as 0.
```

C.4 EVALUATION SKILL

```
# Security Evaluation SKILL

## Overview

This SKILL drives the evaluation Agent to quantitatively assess the
    outputs of the detection Agent and the remediation Agent.

Core mechanism:
-- Assess the quality of the analysis report and the remediation plan (
    judged item by item by the evaluation Agent)

Each checkpoint (CHK) is split into two independent scoring dimensions,
    so that a problem can be pinpointed to either the analysis stage or
    the plan stage:

| Dimension | What is scored | Scoring source | Scoring method |
|------|--------|---------|---------|
| **Detection score** | Whether the issue was correctly identified,
    backed by sufficient evidence, and correctly attributed | The
    detection Agent's analysis report | LLM-as-a-Judge |
| **Plan score** | Whether the remediation plan is technically correct,
    complete in its steps, and mindful of verification and safety | The
    detection Agent's remediation plan | LLM-as-a-Judge |

The two dimensions are **never merged into a single total**; they are
    always reported independently.

**You only need to evaluate one range.** Output that range's item-by-item
     scores and structured data.

## Input Specification

| Input | Description | Source |
|------|------|------|
| Checklist | The range's ground truth, containing the CHK to CAP mapping
     and per-item scoring criteria | `ranges/<range>/checklist.md` |
| Analysis report | The report produced by the detection Agent (intrusion-
    report / vuln-report / baseline-report) | Detection Agent output |
| Remediation plan | The structured remediation plan produced by the
    detection Agent | Detection Agent output `remediation-plan.md` |

## Scoring Process

### Step 1: Read the checklist

Read through the checklist and understand, for each CHK item:
- The detection pass criteria and the plan pass criteria
```

```
- Whether it is a detection-only item (plan score marked N/A) or plan-only item (detection score marked N/A)

### Step 2: Read all reports

Read the detection Agent's complete output (analysis report + remediation-plan.md).

### Step 3: Score item by item

Score each CHK item independently:

#### Detection score (max 3 points) -- LLM-as-a-Judge

| Sub-item | Points | Criteria |
|---|---|---|
| Discovery | 0/1 | Whether the report identifies the issue |
| Evidence | 0/1 | Whether concrete evidence is cited (file paths, log lines, config values, process information, etc.) |
| Attribution | 0/1 | Whether the attribution is correct (the finding is correctly linked to the attack chain / root cause) |

Plan-only items are marked N/A and excluded from the detection score aggregation.

Scoring constraints:
- **Strictly based on the report's content**; award points only for what the report explicitly states
- Do not award points for things that "may have been checked but were not written in the report"
- Quote the report verbatim as the basis for the score

#### Plan score (max 2 points) -- LLM-as-a-Judge

| Sub-item | Points | Criteria |
|---|---|---|
| Correctness | 0/1 | Whether the remediation plan in remediation-plan.md or the report is technically correct |
| Completeness | 0/1 | Whether the plan includes verification steps, edge-case handling, and consideration of side effects |

Detection-only items (e.g., attack timeline reconstruction, entry-point determination) are marked N/A and excluded from the plan score aggregation.

### Step 4: Negative deductions

The following cases require deductions from the detection score (annotate them in the evaluation report):

| Negative rule | Deduction | Description |
|---|---|---|
| Fabrication | Detection score set to 0 | The report invents non-existent CVEs, files, logs, etc. |
| Wrong attribution | Attribution sub-score set to 0 | The finding is attributed to the wrong attack chain or root cause |
| Dangerous advice | Plan score set to 0 | The plan contains operations that could cause data loss or service disruption without flagging the risk |

### Step 5: Determine the diagnostic mode

Determine the overall diagnostic mode from the combination of dimension scores:
```

```
| Mode | Detection | Plan | Meaning | Optimization direction |
|------|------|------|------|---------|
| All-round | Strong | Strong | Ideal state | Maintain |
| Plan bottleneck | Strong | Weak | Finds the issue but cannot produce a correct plan | Improve the detection Agent's plan-generation ability |
| Analysis blind spot | Weak | Weak | Core analysis capability is insufficient | Improve the detection Agent's foundational capability |
| Blind fix | Weak | -- | Happens to fix it correctly, but the report shows no analysis | Untrustworthy; the analysis pipeline needs strengthening |

Strong/weak threshold: ≥70% is strong, <70% is weak. Annotate both the overall and the per-dimension diagnostic mode in the report.

## Output Specification

### 1. evaluation-report.md

A detailed evaluation report in Markdown, containing:

#### Per-item scoring table

One section per CHK item, including:
- CHK ID, name
- Detection score: discovery/evidence/attribution, 0-1 each, with the scoring rationale and verbatim quotes from the report
- Plan score: correctness/completeness, 0-1 each, with the scoring rationale
- Deductions (if any)

#### Diagnostic analysis

- Overall diagnostic mode
- Per-dimension diagnostic modes
- Most frequent point losses (Top N detection misses / Top N plan defects)
- Improvement suggestions

### 2. scores.json

Structured scoring data, for pipeline consumption and cross-run comparison:

'''json
{
  "meta": {
    "range": "ssh-miner",
    "run_id": "20260411-001",
    "model_under_test": "claude-opus-4-7",
    "evaluator_model": "claude-opus-4-7",
    "timestamp": "2026-04-11T10:00:00Z"
  },
  "checkpoints": [
    {
      "chk_id": "CHK-01",
      "name": "Mining process detection",
      "detection": {
        "discovery": 1,
        "evidence": 1,
        "attribution": 1,
        "total": 3,
        "max": 3,
        "rationale": "..."
```

```
      },
      "plan": {
        "correctness": 1,
        "completeness": 0,
        "total": 1,
        "max": 2,
        "rationale": "..."
      },
      "penalties": []
    }
  ],
  "summary": {
    "total_checkpoints": 23,
    "detection_only_count": 6,
    "detection": {
      "score": 60,
      "max": 69,
      "pct": 87.0
    },
    "plan": {
      "score": 28,
      "max": 34,
      "pct": 82.4
    },
    "diagnosis": "..."
  }
}
```

```
## Scoring Constraints

1. **Strictly objective**: award points only for what the report
    explicitly states, quoting the original text as the basis for the
    judgment
2. **No speculation**: do not award points for things the Agent "may have
     checked but did not report"
3. **Dimensions stay independent**: the detection score and plan score
    are always reported separately, never merged into a single total
```

D Supplementary Experimental Results

D.1 Details on Checkpoint Scoring

Each checkpoint is graded along two axes, detection and planning. *Detection* measures whether the agent correctly uncovers problems on the compromised host, including intrusion entities, baseline risks, and vulnerability risks. *Planning* validates whether the agent proposes a correct and complete remediation plan that resolves the existing problems. The detection axis has a total score of 3, comprising discovery of the problem, sufficiency of the supporting evidence, and correctness of attribution to the root cause. The planning axis has a maximum score of 2, comprising correctness of the proposed remediation step and completeness of the plan to fully resolve the attack artifacts.

D.2 Per-Range Performance Analysis

SSH-Miner. The results for the `ssh-miner` range are shown in Table 7. Regarding detection, most models reliably recover the mining process, determine that the entry point is SSH brute force, and surface the SSH misconfigurations together with the persistence mechanisms. The primary weaknesses are failures to investigate silent attacks beyond the alerts, such as the brute-force tool residue (CHK-03) and the `bashrc` persistence (CHK-12). Notably, the scores on the comprehensive persistence scan (CHK-15), the brute-force evidence recovery (CHK-16), and the restart-and-verify confirmation (CHK-22) are all near zero. In terms of planning, models can successfully harden the SSH baseline, revoke the entry, and disable remote MySQL root access, as evidenced by high planning

scores on CHK-04, CHK-05, CHK-08, and CHK-09. However, the cleanup remains incomplete: models fail to close the `bashrc` backdoor (CHK-12), block the attacker IP (CHK-20), verify the SSH hardening (CHK-21), or restart the services (CHK-22).

Redis-RCE. The results for the `redis-rce` range are shown in Table 8. For detection, most models successfully identify the Redis baseline misconfigurations, including the exposed bind address, the missing `requirepass` setting, disabled protected mode, unrestricted dangerous commands, the injected SSH public key, and the malicious startup service. They also correctly attribute the intrusion to unauthorized Redis access and distinguish it from SSH brute force or a CVE exploit. However, most models overlook artifacts that coincide with the attack but are not part of the Redis-generated attack trace, such as CHK-08, CHK-11, CHK-12, and CHK-17. This indicates that models are inclined to discover artifacts directly produced by the attack but overlook other host-wide baseline issues. On the planning axis, models partially harden the Redis baseline and clean up the explicit persistence, achieving high success rates in removing the injected SSH key and the malicious startup service. Nevertheless, their Redis hardening remains incomplete, often missing the change from the root run-as user (CHK-08) and the final verification of the applied hardening (CHK-22).

Docker-Escape. The results for the `docker-escape` range are shown in Table 9. Most models correctly detect the surface of the chain, flagging the mining process (CHK-01), the backdoor container (CHK-03), the unauthenticated Docker TCP API exposure (CHK-05), and the host-level persistence (CHK-12, CHK-13, CHK-14). However, nearly every model scores close to zero on the proactive comprehensive persistence scan (CHK-16), showing that they recover only the specific persistence artifacts along the alerted trail and do not sweep the host thoroughly. Moreover, several models, including models in the Gemini series, fail to detect the full escape mechanism because they do not recognize that the developer user's membership in the Docker group is a root-equivalent escape primitive (CHK-07). On the planning dimension, models can harden and securely restart the Docker daemon (CHK-20), but they almost never verify that the legitimate business container survives the cleanup (CHK-22), and they leave the Docker-group escape (CHK-07) and the miner's outbound channel (CHK-24) open. This risks disrupting production while the escape route remains intact.

Jenkins-RCE. The results for the `jenkins-rce` range are shown in Table 10. On the detection side, most models reliably identify the visible attacks from alerts. They successfully flag the explicit intrusion entities, such as the mining process and its mining-pool outbound connection, the anonymous entry point misconfiguration, and the conspicuous persistence artifacts such as the cron entry and the rogue account. Stronger proprietary models additionally reconstruct the full process chain from the Jenkins application layer and the mining process back to the Groovy payload. The weaknesses are concentrated in three aspects. First, inconspicuous persistence is widely missed, including the SUID-root backdoor, the `LD_PRELOAD` injection, and the sudoers tampering. This indicates that the models are inclined to follow the alerted attack trail rather than proactively examine the system state. Second, the silent configuration weakness, such as the weak Jenkins admin password, is rarely recovered. Third, models report the direct cause of the incident but hardly assess its collateral effects, failing to flag that revoking anonymous Administer also severs any legitimate unauthenticated access path, which makes business-impact assessment among the weakest dimensions. On the planning side, models commonly perform well on stopping the miner, deleting the cron entry and rogue account, blocking the attacker and mining-pool IPs, and revoking anonymous administrative access. However, the models fail to remediate the same inconspicuous persistence that detection misses. In addition, the remediation plans rarely reach a complete and verified state, as shown by the low planning scores on CHK-10, CHK-21, and CHK-23.

Shiro-Fastjson. The results for the `shiro-fastjson` range are shown in Table 11. Detection on this range shows a high success rate. Most models recover both JSP webshells, the mining process and binary, and the OS-level persistence. They also reconstruct the root cause in depth, attributing the entry to Shiro deserialization (CHK-25), tracing the Java process chain (CHK-29), and rebuilding the sudo privilege-escalation path and the attack timeline (CHK-30, CHK-31). However, two common weaknesses can be observed. First, the database- and service-layer attack surface is largely missed, as reflected by failures on CHK-06, CHK-12, and CHK-15. This suggests that models inspect the host filesystem and process state but not the database engine or service network exposure. Second,

although models identify the headline Shiro and Fastjson vulnerabilities (CHK-26, CHK-27), few trace the deserialization down to the underlying program that makes it exploitable (CHK-28). On the planning side, models can propose remediation plans for the explicit entities and baselines, with high scores on stopping the miner, reducing Tomcat's privileges, and rotating the database credentials. However, the database-resident UDF persistence is left untouched, and the plans rarely reach a verified state (CHK-41, CHK-43, CHK-45).

Log4j-RCE. The results for the `log4j-rce` range are shown in Table 12. Detection on this range is broadly strong, and most models reliably discover the active intrusion entities and the Log4j vulnerability, indicating that the models have sufficient knowledge of this CVE. However, most models still demonstrate weaknesses in two scenarios. First, latent login-triggered persistence is a systematic blind spot: the `profile.d` backdoor (CHK-15) and the injected SSH key (CHK-14) leave no trace in the running process or network state and are missed by every model except the GPT series. Second, exfiltrated data on disk and the JNDI remote-class-loading misconfiguration in the application source are identified by only a few models, with most models earning partial credit. In terms of planning, models perform reliably on explicit baseline-hardening tasks such as fixing the SSH configuration, assessing the vulnerability, and upgrading Log4j. However, they fail to suggest complete fixes or cleanups, performing poorly on the source-code repair and its recompile-and-restart steps (CHK-11), the cleanup of undetected login-triggered persistence (CHK-14, CHK-15), and the full removal of the mining payload (CHK-03).

Next.js-RCE. The results for the `nextjs-rce` range are shown in Table 13. For detection, models inspect the web application thoroughly but rarely pivot to the surrounding host. They can flag the disguised miner and hidden webshell and trace the entry point to the Next.js RSC deserialization RCE. However, most models ignore artifacts outside the web application's own directory, such as CHK-04, CHK-05, CHK-06, and CHK-09, which are recovered only by a few models, including GLM-5.1 and Claude Opus 4.6. Furthermore, few models trace the RSC flaw back to its React root (CHK-22). On the planning dimension, the remediation plan covers the application baseline, but the undetected host persistence is left in place (CHK-04, CHK-06, CHK-09), the leaked secrets are not fully rotated (CHK-20), and few models recommend replaying the RSC payload to confirm that the endpoint rejects it after the Next.js upgrade (CHK-24). The headline vulnerability is therefore patched but never verified.

NPM-Worm. The results for the `npm-worm` range are shown in Table 14. On detection, most models reliably identify the active intrusion entities and the entry point, including the worm process for CHK-01, the malicious npm package for CHK-04, the JS dropper for CHK-03, the unauthenticated webhook endpoint for CHK-13, and the attacker IP for CHK-30. Notably, the runtime attack chain is consistently investigated despite spanning two language runtimes (CHK-27), and the attack timeline is broadly reconstructed (CHK-33). Overall detection coverage on this range, however, is lower than on other ranges. Models collapse on the worm's disguised persistence, namely the masqueraded systemd service, the `profile.d` backdoor, and the comprehensive persistence sweep. This indicates that current models remain limited under a wider attack surface and the worm's deliberate use of disguise. On the planning dimension, models propose reliable remediation only for the explicit entry point, with high success rates in uninstalling the malicious npm package for CHK-04 and hardening the webhook endpoint for CHK-13. The cleanup, however, remains incomplete. Agents still fail to remove the worm script and mining binary or eradicate the worm's lateral infection and disguised persistence. Moreover, most models broadly recommend rotating leaked credentials without specifying how to handle each credential category.

ASP.NET-ViewState. The results for the `aspnet-viewstate` range are shown in Table 15. For this Windows range, detection performance differs across artifact types. Models typically recover the visible IIS-layer surface and attribute the correct entry point. However, Windows-specific stealth mechanisms are challenging to discover. The side-loading hijack is missed by almost all models (CHK-08, CHK-15), and the encoded payloads (CHK-09/10/11) are recovered only occasionally. This indicates that recovering such artifacts depends on whether a model proactively inspects the host rather than following the alerted ViewState trail. On the planning dimension, models successfully fix the headline vulnerability, rotate the `machineKey`, repair the misconfigurations, and rotate the SQL credential. Nevertheless, the MSSQL logon trigger and `xp_cmdshell` backdoor are detected

but never cleaned up, with planning scores close to zero on CHK-16 and CHK-17. In addition, the attacker IP is rarely blocked (CHK-06), and the undetected scheduled-task persistence is left untouched.

RDP-Service-Abuse. The results for the `rdp-service-abuse` range are shown in Table 16. The obvious post-compromise footprint is recovered almost universally, whereas the initial-access and privilege-abuse path that defines the scenario is largely overlooked. Almost every model reconstructs the SYSTEM payload execution chain (CHK-07) and flags the disguised service (CHK-08), the scheduled task (CHK-10, CHK-11), the credential dump (CHK-15/16/17), and the C2 indicators (CHK-14). In contrast, the entry point is missed by nearly all models as evidenced by CHK-01, CHK-05, and CHK-06. On the planning dimension, models can remediate the parts of the chain that they detect, such as CHK-09, CHK-12, and CHK-18. However, remediation of the unseen entry vector is similarly poor. Only Claude Opus 4.6 and Claude Sonnet 4.6 propose fixing the weak `helpdesk` password (CHK-19) and hardening the DACL (CHK-20). Most agents finish with the persistence cleaned up but leave the exploitable service and its weak credential unchanged.

Table 7: Per-model checkpoint scores for `ssh-miner`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2).

Model	CHK-01 Mining process detection		CHK-02 Mining binary cleanup		CHK-03 Hydra/ attack-tool residue		CHK-04 SSH Permit RootLogin		CHK-05 SSH Password Authentication		CHK-06 SSH MaxAuth Tries		CHK-07 Root password strength		CHK-08 devops sudoers NOPASSWD	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	2.00	3.00	2.00	0.00	1.00	3.00	2.00	3.00	2.00	3.00	2.00	2.33	1.67	3.00	2.00
Claude Opus 4.6	3.00	1.67	3.00	1.33	0.00	0.00	3.00	1.33	3.00	2.00	3.00	1.00	3.00	2.00	3.00	1.00
Claude Opus 4.5	3.00	2.00	3.00	1.00	0.00	0.00	3.00	1.33	3.00	2.00	3.00	1.00	0.33	2.00	3.00	1.00
Claude Sonnet 4.6	3.00	2.00	3.00	1.00	0.00	0.00	3.00	1.33	3.00	2.00	3.00	1.00	3.00	2.00	3.00	1.00
Claude Sonnet 4.5	3.00	2.00	3.00	1.67	0.00	0.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00
GPT-5.5	3.00	0.33	3.00	0.33	0.00	0.00	3.00	1.00	3.00	1.33	3.00	1.00	1.67	1.33	3.00	1.00
GPT-5.4 Pro	3.00	0.33	3.00	0.33	0.00	0.00	3.00	1.00	3.00	1.67	3.00	1.00	0.33	0.67	2.33	1.00
GPT-5.4	3.00	0.33	3.00	0.67	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.00	0.00	0.67	3.00	1.00
GPT-5.2 Pro	2.33	0.33	3.00	0.33	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.00	0.33	0.33	3.00	1.00
Gemini 3.1 Pro	3.00	1.67	3.00	1.00	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.00	2.33	1.67	3.00	1.00
Gemini 3 Flash	2.67	1.33	2.67	0.67	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.00	2.33	1.00	3.00	1.00
GLM-5.1	3.00	2.00	3.00	1.00	0.00	0.00	3.00	1.33	3.00	2.00	3.00	1.00	3.00	2.00	3.00	2.00
GLM-5	3.00	1.00	3.00	1.33	0.00	0.00	3.00	1.67	3.00	2.00	3.00	1.33	3.00	2.00	3.00	2.00
DeepSeek V4 Pro	3.00	2.00	3.00	1.00	0.00	0.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	1.00
DeepSeek V3.2	3.00	1.67	3.00	2.00	0.00	0.33	3.00	1.00	3.00	1.00	3.00	1.00	1.67	2.00	3.00	2.00
Qwen3.7 Max	3.00	2.00	3.00	1.00	0.33	0.00	3.00	1.33	3.00	1.33	3.00	1.00	2.33	2.00	3.00	1.00
Qwen3.7 Plus	3.00	0.67	3.00	0.33	0.00	0.00	3.00	1.00	3.00	1.33	3.00	1.00	3.00	2.00	3.00	1.00
Qwen3.6 Plus	3.00	2.00	3.00	2.00	0.00	0.00	3.00	1.33	3.00	2.00	3.00	1.00	2.00	2.00	3.00	2.00
Qwen3.5 Plus	3.00	0.33	3.00	1.00	0.00	0.00	3.00	2.00	3.00	2.00	3.00	2.00	0.33	0.67	3.00	2.00
Kimi K2.6	2.67	1.67	3.00	1.00	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.00	2.00	1.67	3.00	1.00
Kimi K2.5	3.00	1.67	3.00	0.67	0.67	0.33	3.00	2.00	3.00	1.67	3.00	2.00	2.67	1.67	3.00	2.00
MiniMax M2.7	3.00	0.33	3.00	1.33	0.00	0.00	3.00	1.00	3.00	2.00	3.00	1.00	2.67	2.00	3.00	1.00
MiniMax M2.5	3.00	1.33	3.00	0.67	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.00	2.67	1.67	3.00	1.00

Model	CHK-09 MySQL remote root access		CHK-10 MySQL creds in bash_history		CHK-11 Crontab persistence		CHK-12 Bashrc persistence		CHK-13 Systemd service persistence		CHK-14 Systemd service masquerade ID		CHK-15 Comprehensive persistence scan		CHK-16 Brute-force evidence recovery	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	—	0.67	—	0.00	—
Claude Opus 4.6	3.00	1.33	3.00	1.00	3.00	1.00	0.00	0.00	3.00	1.00	3.00	—	0.00	—	0.33	—
Claude Opus 4.5	3.00	2.00	3.00	1.67	3.00	1.33	0.00	0.00	3.00	2.00	3.00	—	0.67	—	0.33	—
Claude Sonnet 4.6	3.00	1.33	3.00	0.33	3.00	1.00	0.00	0.00	3.00	0.67	3.00	—	0.00	—	0.33	—
Claude Sonnet 4.5	3.00	2.00	0.00	0.00	3.00	1.33	0.00	0.00	3.00	2.00	3.00	—	1.00	—	0.67	—
GPT-5.5	2.00	1.00	0.00	0.00	3.00	0.67	3.00	0.67	3.00	0.33	2.67	—	0.33	—	0.00	—
GPT-5.4 Pro	0.33	0.33	2.67	0.33	3.00	1.00	3.00	1.00	3.00	0.33	2.67	—	0.00	—	0.33	—
GPT-5.4	1.00	0.33	0.00	0.00	3.00	0.33	2.67	1.67	3.00	0.33	0.33	—	0.00	—	0.00	—
GPT-5.2 Pro	2.33	0.33	2.33	0.00	3.00	0.67	2.33	0.67	3.00	0.33	0.33	—	0.00	—	0.33	—
Gemini 3.1 Pro	2.67	1.00	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.67	2.67	—	0.00	—	0.33	—
Gemini 3 Flash	3.00	0.33	2.67	0.00	3.00	1.00	0.00	0.00	3.00	0.33	0.67	—	0.00	—	0.00	—
GLM-5.1	3.00	2.00	3.00	2.00	3.00	1.00	0.00	0.00	3.00	2.00	3.00	—	0.00	—	0.00	—
GLM-5	3.00	1.67	3.00	1.00	3.00	1.33	0.00	0.00	3.00	1.67	3.00	—	0.00	—	0.33	—
DeepSeek V4 Pro	3.00	1.67	2.67	1.00	3.00	0.67	3.00	1.00	3.00	1.67	3.00	—	0.00	—	0.00	—
DeepSeek V3.2	2.00	2.00	0.00	0.00	3.00	1.00	0.00	0.00	3.00	1.00	1.33	—	0.33	—	0.33	—
Qwen3.7 Max	3.00	1.33	3.00	0.00	3.00	2.00	3.00	2.00	3.00	1.00	3.00	—	0.67	—	0.00	—
Qwen3.7 Plus	3.00	1.00	3.00	1.00	3.00	1.00	0.00	0.00	3.00	1.33	2.33	—	0.00	—	0.00	—
Qwen3.6 Plus	3.00	1.67	3.00	2.00	3.00	1.33	0.00	0.00	3.00	1.33	2.33	—	0.00	—	0.00	—
Qwen3.5 Plus	3.00	2.00	1.67	0.00	3.00	1.00	0.00	0.00	3.00	2.00	0.33	—	0.00	—	0.33	—
Kimi K2.6	3.00	1.00	3.00	0.00	3.00	1.00	0.00	0.00	3.00	1.33	0.67	—	0.00	—	0.33	—
Kimi K2.5	3.00	2.00	0.00	0.00	3.00	1.33	0.00	0.00	3.00	2.00	3.00	—	0.33	—	1.00	—
MiniMax M2.7	3.00	1.00	3.00	0.67	3.00	0.67	0.00	0.00	3.00	1.33	1.33	—	0.00	—	0.00	—
MiniMax M2.5	3.00	1.00	0.00	0.00	3.00	1.00	0.00	0.00	3.00	1.67	0.00	—	0.00	—	0.33	—

Model	CHK-17 Attacker-IP identification		CHK-18 Entry-point determination		CHK-19 Attack-timeline reconstruction		CHK-20 Attacker-IP firewall block		CHK-21 SSH hardening verification		CHK-22 Service restart and verification		CHK-23 Mining outbound-connection block	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	2.33	—	3.00	—	1.67	—	1.00	0.00	3.00	2.00	0.67	0.33	3.00	1.33
Claude Opus 4.6	2.67	—	3.00	—	2.33	—	3.00	1.00	1.33	0.00	0.33	0.00	3.00	0.67
Claude Opus 4.5	2.67	—	3.00	—	2.67	—	3.00	2.00	1.33	0.33	0.33	0.33	2.67	1.00
Claude Sonnet 4.6	2.33	—	3.00	—	2.33	—	3.00	0.33	2.33	0.67	0.33	0.00	2.67	0.33
Claude Sonnet 4.5	2.67	—	3.00	—	2.33	—	3.00	2.00	3.00	2.00	2.00	1.00	3.00	2.00
GPT-5.5	2.67	—	3.00	—	2.33	—	1.33	0.00	0.33	0.00	0.33	0.00	3.00	0.33
GPT-5.4 Pro	2.33	—	3.00	—	2.67	—	0.67	0.00	0.67	0.00	0.33	0.00	2.67	0.33
GPT-5.4	3.00	—	3.00	—	2.67	—	0.00	0.00	0.00	0.00	0.00	0.00	3.00	0.33
GPT-5.2 Pro	2.67	—	3.00	—	2.00	—	0.33	0.00	0.33	0.00	0.00	0.00	2.67	0.33
Gemini 3.1 Pro	2.67	—	3.00	—	2.33	—	0.33	0.00	0.33	0.00	0.00	0.00	2.33	0.00
Gemini 3 Flash	2.33	—	3.00	—	1.33	—	0.67	0.00	0.00	0.00	0.00	0.00	1.00	0.00
GLM-5.1	2.67	—	3.00	—	2.33	—	1.67	0.00	2.33	0.67	1.00	0.67	3.00	0.33
GLM-5	3.00	—	2.67	—	2.00	—	3.00	1.33	2.33	1.00	0.00	0.00	3.00	0.00
DeepSeek V4 Pro	2.67	—	3.00	—	2.33	—	0.00	0.00	3.00	2.00	0.00	0.00	3.00	1.00
DeepSeek V3.2	2.67	—	3.00	—	2.00	—	0.33	0.67	0.67	0.67	0.67	0.67	3.00	1.00
Qwen3.7 Max	2.67	—	3.00	—	3.00	—	0.33	0.00	0.67	0.33	0.33	0.00	3.00	0.67
Qwen3.7 Plus	2.33	—	2.67	—	2.33	—	3.00	0.33	0.00	0.00	0.00	0.00	3.00	0.33
Qwen3.6 Plus	2.33	—	3.00	—	2.67	—	1.33	0.00	1.67	0.00	0.33	0.00	3.00	0.00
Qwen3.5 Plus	2.00	—	3.00	—	2.33	—	2.33	0.00	3.00	1.33	0.33	0.33	2.67	0.67
Kimi K2.6	2.67	—	3.00	—	2.33	—	0.33	0.00	0.33	0.00	0.00	0.00	0.33	0.00
Kimi K2.5	3.00	—	3.00	—	2.33	—	1.00	0.67	2.67	2.00	1.00	0.67	1.67	0.33
MiniMax M2.7	3.00	—	3.00	—	2.33	—	0.00	0.00	2.00	0.00	0.33	0.00	3.00	0.00
MiniMax M2.5	2.67	—	3.00	—	2.00	—	3.00	1.00	0.00	0.00	0.00	0.00	3.00	0.67

Table 8: Per-model checkpoint scores for `redis-rce`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2).

Model	CHK-01 Mining process detection		CHK-02 Mining binary cleanup		CHK-03 Attacker SSH public-key identification		CHK-04 Redis-written crontab residue		CHK-05 Redis bind configuration		CHK-06 Redis require pass		CHK-07 Redis protected mode		CHK-08 Redis run-as user	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	1.33	1.67	2.00	1.33	3.00	2.00	3.00	1.33	3.00	2.00	3.00	1.67	3.00	2.00	3.00	2.00
Claude Opus 4.6	3.00	0.67	3.00	0.67	3.00	2.00	3.00	2.00	3.00	1.67	3.00	1.33	3.00	2.00	1.33	1.67
Claude Opus 4.5	1.33	2.00	2.00	1.00	3.00	2.00	1.00	1.33	3.00	1.00	3.00	1.00	3.00	2.00	0.00	0.00
Claude Sonnet 4.6	1.67	2.00	2.00	1.33	3.00	2.00	3.00	1.67	3.00	2.00	3.00	2.00	3.00	2.00	1.33	0.00
Claude Sonnet 4.5	1.33	0.67	2.33	1.33	3.00	1.67	3.00	1.67	3.00	2.00	3.00	1.67	3.00	2.00	3.00	2.00
GPT-5.5	1.33	0.67	2.33	0.67	3.00	2.00	3.00	1.67	3.00	1.33	3.00	1.33	3.00	1.67	1.00	1.00
GPT-5.4 Pro	1.00	0.00	1.33	0.00	3.00	1.67	3.00	1.00	3.00	1.67	3.00	1.00	3.00	2.00	3.00	2.00
GPT-5.4	0.33	0.00	2.00	1.00	3.00	2.00	3.00	1.00	3.00	1.33	3.00	1.00	3.00	1.33	1.33	1.00
GPT-5.2 Pro	1.33	0.00	1.33	0.00	3.00	1.67	1.00	0.00	3.00	1.67	3.00	1.67	3.00	2.00	0.00	1.00
Gemini 3.1 Pro	1.00	1.67	1.67	1.00	3.00	1.67	3.00	1.00	3.00	1.67	3.00	1.00	3.00	1.33	0.33	1.00
Gemini 3 Flash	1.33	1.00	2.33	1.00	3.00	1.00	3.00	0.67	3.00	1.67	3.00	1.00	3.00	1.00	0.00	0.00
GLM-5.1	2.00	2.00	2.67	1.67	3.00	2.00	3.00	2.00	3.00	1.33	3.00	1.00	3.00	2.00	0.33	0.00
GLM-5	1.33	1.67	2.33	1.33	2.33	1.00	2.00	0.33	3.00	2.00	3.00	1.67	3.00	2.00	0.00	0.00
DeepSeek V4 Pro	1.33	2.00	2.33	1.33	3.00	2.00	0.00	0.00	3.00	1.33	3.00	1.33	3.00	1.67	0.00	0.00
DeepSeek V3.2	1.33	0.00	2.00	1.00	3.00	2.00	3.00	1.67	3.00	1.67	3.00	1.00	3.00	1.33	2.33	1.00
Qwen3.7 Max	3.00	2.00	3.00	1.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	0.00	0.00
Qwen3.7 Plus	3.00	2.00	3.00	1.00	3.00	2.00	3.00	1.67	3.00	1.67	3.00	1.67	3.00	1.67	2.67	2.00
Qwen3.6 Plus	1.67	1.00	2.33	1.33	3.00	2.00	0.00	0.00	3.00	2.00	3.00	1.67	3.00	2.00	0.00	0.00
Qwen3.5 Plus	2.67	2.00	2.67	1.00	2.67	2.00	2.67	1.33	3.00	2.00	3.00	2.00	3.00	1.33	0.00	0.00
Kimi K2.6	0.00	0.00	1.67	0.00	3.00	1.00	3.00	0.67	3.00	2.00	3.00	1.00	3.00	1.33	0.00	0.00
Kimi K2.5	1.67	2.00	2.00	1.00	3.00	2.00	3.00	2.00	3.00	1.00	2.67	1.67	3.00	1.67	0.00	0.00
MiniMax M2.7	0.67	0.33	2.00	1.00	2.67	1.67	2.67	1.00	3.00	1.33	3.00	1.33	3.00	2.00	2.33	0.33
MiniMax M2.5	0.67	2.00	2.00	1.00	2.67	1.00	1.33	0.33	3.00	1.00	3.00	1.00	3.00	1.00	0.33	0.00

Model	CHK-09 Redis rename command		CHK-10 SSH Permit RootLogin		CHK-11 deploy sudoers NOPASSWD		CHK-12 Redis session creds in bash_history		CHK-13 SSH .ssh/rc persistence		CHK-14 init.d malicious script		CHK-15 init.d service masquerade		CHK-16 Redis crontab file signature ID	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	2.00	3.00	1.33	0.00	0.00	3.00	1.33	3.00	1.33	3.00	1.67	2.67	—	3.00	—
Claude Opus 4.6	3.00	1.67	3.00	1.00	0.00	0.00	0.00	0.00	3.00	0.67	3.00	1.67	3.00	—	3.00	—
Claude Opus 4.5	3.00	1.00	3.00	1.00	0.00	0.00	3.00	0.00	3.00	1.00	3.00	2.00	2.67	—	0.00	—
Claude Sonnet 4.6	3.00	1.67	3.00	1.00	0.00	0.00	0.00	0.00	3.00	2.00	3.00	1.67	2.33	—	3.00	—
Claude Sonnet 4.5	3.00	2.00	2.33	2.00	0.00	0.00	3.00	0.67	3.00	1.67	3.00	2.00	2.67	—	3.00	—
GPT-5.5	3.00	1.67	3.00	1.00	0.00	0.00	3.00	0.33	3.00	1.33	3.00	1.00	2.67	—	3.00	—
GPT-5.4 Pro	3.00	1.33	3.00	1.00	0.00	0.00	2.00	1.33	3.00	1.00	3.00	1.00	2.33	—	3.00	—
GPT-5.4	3.00	1.00	2.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	—	3.00	—
GPT-5.2 Pro	3.00	1.33	3.00	1.00	0.00	0.00	0.00	0.00	3.00	1.00	3.00	1.33	0.00	—	1.00	—
Gemini 3.1 Pro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.00	2.67	—	3.00	—
Gemini 3 Flash	2.33	1.33	2.33	1.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.00	2.00	—	3.00	—
GLM-5.1	3.00	1.67	3.00	1.00	0.00	0.00	3.00	0.00	3.00	1.00	3.00	2.00	2.33	—	3.00	—
GLM-5	3.00	1.33	3.00	1.00	0.00	0.00	0.00	0.00	2.67	1.00	3.00	2.00	0.00	—	0.67	—
DeepSeek V4 Pro	3.00	2.00	3.00	1.00	0.00	0.00	0.00	0.00	3.00	1.00	3.00	2.00	2.33	—	0.00	—
DeepSeek V3.2	3.00	1.67	2.33	1.00	0.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00	—	3.00	—
Qwen3.7 Max	3.00	1.67	3.00	1.33	3.00	2.00	3.00	1.00	3.00	1.00	3.00	2.00	2.67	—	3.00	—
Qwen3.7 Plus	3.00	1.67	3.00	1.67	3.00	1.00	0.00	0.00	2.67	1.00	3.00	2.00	2.67	—	3.00	—
Qwen3.6 Plus	3.00	1.67	3.00	1.00	0.00	0.00	3.00	0.67	3.00	1.00	3.00	2.00	2.67	—	0.00	—
Qwen3.5 Plus	3.00	2.00	2.67	1.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	2.00	0.33	—	0.67	—
Kimi K2.6	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	2.67	1.00	3.00	0.33	1.33	—	3.00	—
Kimi K2.5	3.00	1.67	3.00	2.00	0.00	0.00	0.00	0.00	3.00	1.00	3.00	1.67	2.00	—	3.00	—
MiniMax M2.7	3.00	1.67	3.00	1.33	0.00	0.00	0.00	0.00	0.00	0.00	3.00	2.00	0.33	—	1.00	—
MiniMax M2.5	3.00	0.00	2.67	1.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.00	2.33	—	1.33	—

Model	CHK-17 Comprehensive persistence scan		CHK-18 Redis unauthorized access evidence		CHK-19 Attacker-IP identification		CHK-20 Entry-point determination		CHK-21 Attack-timeline reconstruction		CHK-22 Redis hardening verification		CHK-23 Port access restriction		CHK-24 Service restart and verification	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	1.67	—	3.00	—	2.67	—	3.00	—	2.67	—	1.00	0.67	3.00	1.00	0.33	0.33
Claude Opus 4.6	0.00	—	3.00	—	2.67	—	2.67	—	3.00	—	1.67	1.33	3.00	2.00	1.00	0.33
Claude Opus 4.5	0.00	—	2.67	—	2.67	—	2.67	—	3.00	—	2.33	1.33	3.00	2.00	0.33	0.33
Claude Sonnet 4.6	1.33	—	2.00	—	1.67	—	3.00	—	1.67	—	0.33	0.00	2.00	1.67	0.00	0.00
Claude Sonnet 4.5	2.00	—	3.00	—	2.67	—	2.67	—	2.67	—	3.00	1.67	3.00	2.00	1.33	0.67
GPT-5.5	1.33	—	3.00	—	2.67	—	3.00	—	2.67	—	0.33	0.00	2.67	1.00	0.33	0.00
GPT-5.4 Pro	1.33	—	3.00	—	2.67	—	2.67	—	2.67	—	0.33	0.00	2.33	1.00	0.00	0.00
GPT-5.4	0.00	—	3.00	—	2.67	—	0.33	—	3.00	—	2.67	0.00	3.00	1.00	0.00	0.00
GPT-5.2 Pro	0.00	—	2.67	—	2.67	—	1.67	—	2.67	—	1.33	0.67	3.00	1.33	0.33	0.00
Gemini 3.1 Pro	0.33	—	3.00	—	2.67	—	2.67	—	1.33	—	0.33	0.33	2.67	1.00	0.00	0.33
Gemini 3 Flash	0.33	—	3.00	—	2.67	—	2.67	—	2.00	—	0.33	0.00	2.67	1.00	0.00	0.00
GLM-5.1	0.33	—	3.00	—	2.67	—	3.00	—	3.00	—	1.67	1.67	3.00	2.00	0.67	0.00
GLM-5	0.33	—	2.67	—	2.67	—	2.67	—	2.67	—	0.33	0.33	2.67	1.00	0.00	0.00
DeepSeek V4 Pro	1.00	—	3.00	—	2.67	—	2.67	—	2.67	—	2.33	1.67	3.00	1.67	1.00	0.33
DeepSeek V3.2	0.00	—	3.00	—	2.67	—	3.00	—	2.67	—	2.00	1.00	3.00	2.00	0.00	0.00
Qwen3.7 Max	2.00	—	3.00	—	2.67	—	2.67	—	3.00	—	1.67	1.67	3.00	2.00	0.33	0.33
Qwen3.7 Plus	1.00	—	3.00	—	2.67	—	2.67	—	2.67	—	1.67	1.00	3.00	1.67	1.67	0.67
Qwen3.6 Plus	0.00	—	2.33	—	2.67	—	2.67	—	2.33	—	1.33	0.33	2.67	1.33	0.00	0.00
Qwen3.5 Plus	0.33	—	3.00	—	2.67	—	2.67	—	2.67	—	2.33	1.67	3.00	1.00	1.67	0.67
Kimi K2.6	0.33	—	3.00	—	2.67	—	3.00	—	2.67	—	0.33	0.00	2.67	1.00	1.00	0.00
Kimi K2.5	0.00	—	3.00	—	2.67	—	2.67	—	2.67	—	1.67	1.67	3.00	2.00	1.33	0.67
MiniMax M2.7	0.33	—	3.00	—	2.33	—	0.67	—	1.67	—	2.00	1.00	3.00	2.00	2.00	0.67
MiniMax M2.5	0.00	—	2.00	—	2.67	—	0.67	—	2.33	—	1.67	0.00	3.00	1.00	0.00	0.00

Table 9: Per-model checkpoint scores for `docker-escape`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2).

Model	CHK-01 Mining process detection		CHK-02 Mining file cleanup		CHK-03 Backdoor container detection & removal		CHK-04 Attacker SSH key		CHK-05 Docker TCP-API exposure		CHK-06 Docker daemon systemd override		CHK-07 developer user docker-group privilege		CHK-08 Container env-var database credentials	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	2.00	3.00	1.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	1.00	0.00	0.67
Claude Opus 4.6	3.00	2.00	3.00	1.33	3.00	1.00	3.00	1.00	3.00	2.00	3.00	2.00	3.00	1.67	2.67	2.00
Claude Opus 4.5	3.00	1.67	3.00	1.00	3.00	2.00	3.00	1.33	3.00	2.00	3.00	1.33	3.00	2.00	2.67	1.00
Claude Sonnet 4.6	3.00	2.00	3.00	1.00	3.00	1.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	0.67	3.00	2.00
Claude Sonnet 4.5	3.00	2.00	3.00	1.33	3.00	2.00	3.00	1.67	3.00	2.00	3.00	2.00	0.00	0.00	3.00	2.00
GPT-5.5	3.00	1.00	3.00	0.33	3.00	1.67	3.00	0.33	3.00	2.00	2.33	0.33	0.00	0.00	1.67	2.00
GPT-5.4 Pro	3.00	1.00	3.00	0.67	3.00	1.33	3.00	1.00	2.00	1.67	2.00	0.67	0.00	0.00	2.67	1.67
GPT-5.4	3.00	0.00	3.00	0.00	3.00	0.33	3.00	0.33	3.00	2.00	0.00	0.00	2.00	0.00	2.00	2.00
GPT-5.2 Pro	2.67	1.67	3.00	1.00	2.67	1.00	2.67	1.00	3.00	2.00	2.33	0.33	3.00	1.00	0.00	0.00
Gemini 3.1 Pro	3.00	2.00	3.00	1.00	2.67	1.33	3.00	1.67	3.00	2.00	2.67	1.00	0.00	0.00	0.00	0.00
Gemini 3 Flash	3.00	1.67	3.00	0.00	2.00	1.00	3.00	1.00	3.00	2.00	2.00	2.00	0.00	0.00	0.00	0.00
GLM-5.1	3.00	2.00	3.00	2.00	3.00	1.67	3.00	1.67	3.00	1.67	2.67	2.00	3.00	2.00	3.00	2.00
GLM-5	3.00	2.00	3.00	1.00	3.00	1.00	3.00	2.00	3.00	2.00	3.00	1.00	3.00	1.00	0.00	0.00
DeepSeek V4 Pro	3.00	2.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	1.00	2.33	2.00	3.00	2.00	0.00	0.33
DeepSeek V3.2	3.00	1.67	3.00	1.00	3.00	1.67	3.00	1.67	3.00	1.67	2.67	1.33	0.00	0.00	0.00	0.00
Qwen3.7 Max	3.00	2.00	3.00	1.00	3.00	1.00	3.00	1.33	3.00	2.00	2.67	2.00	2.00	1.00	2.00	2.00
Qwen3.7 Plus	3.00	1.33	3.00	1.00	3.00	1.67	3.00	1.00	3.00	2.00	3.00	1.00	3.00	1.33	2.33	1.67
Qwen3.6 Plus	3.00	2.00	3.00	1.00	0.00	0.00	3.00	1.67	3.00	2.00	3.00	2.00	3.00	2.00	0.00	0.67
Qwen3.5 Plus	3.00	2.00	3.00	1.00	0.00	0.00	3.00	2.00	3.00	2.00	2.67	1.67	3.00	1.00	0.00	0.00
Kimi K2.6	3.00	2.00	3.00	1.00	3.00	1.67	3.00	2.00	3.00	2.00	3.00	1.67	3.00	1.67	2.33	2.00
Kimi K2.5	3.00	2.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	2.00	3.00	2.00	0.00	0.00	2.00	1.33
MiniMax M2.7	3.00	2.00	3.00	1.00	3.00	1.67	3.00	2.00	3.00	1.33	2.00	2.00	0.00	0.00	0.00	0.00
MiniMax M2.5	3.00	2.00	3.00	0.33	3.00	0.33	3.00	2.00	3.00	2.00	2.00	1.67	2.00	1.00	0.00	0.00

Model	CHK-09 SSH authorized keys management		CHK-10 Business container security audit		CHK-11 Docker-API port security-group or firewall		CHK-12 Crontab persistence		CHK-13 Bashrc persistence		CHK-14 Systemd service persistence		CHK-15 Systemd service masquerade ID		CHK-16 Compre-hensive persistence scan	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	1.67	2.67	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	0.67	3.00	—	2.00	—
Claude Opus 4.6	3.00	1.00	2.67	1.00	3.00	1.33	3.00	1.00	3.00	1.00	3.00	1.67	3.00	—	0.67	—
Claude Opus 4.5	3.00	1.33	1.33	1.33	3.00	2.00	3.00	1.00	3.00	1.00	3.00	1.00	2.33	—	0.67	—
Claude Sonnet 4.6	3.00	2.00	1.67	0.67	3.00	1.67	3.00	2.00	3.00	1.00	3.00	1.00	3.00	—	0.33	—
Claude Sonnet 4.5	1.33	1.00	3.00	2.00	3.00	2.00	3.00	1.33	3.00	1.33	3.00	2.00	3.00	—	0.33	—
GPT-5.5	3.00	1.33	0.33	0.33	2.33	1.00	3.00	1.00	2.33	0.67	3.00	0.67	0.33	—	0.33	—
GPT-5.4 Pro	2.67	1.33	1.00	0.33	2.00	1.33	3.00	1.00	3.00	1.67	3.00	0.67	0.67	—	0.33	—
GPT-5.4	2.67	1.67	0.00	0.00	2.67	0.67	3.00	0.33	0.00	0.00	3.00	0.33	2.00	—	0.00	—
GPT-5.2 Pro	2.33	1.00	0.00	0.00	2.67	1.33	3.00	1.00	2.67	1.00	3.00	0.67	0.00	—	0.67	—
Gemini 3.1 Pro	1.33	1.00	0.00	0.00	2.33	1.33	3.00	1.00	2.00	1.67	3.00	1.33	0.00	—	0.33	—
Gemini 3 Flash	1.33	1.00	0.00	0.00	2.33	0.67	3.00	1.00	2.00	0.00	2.67	2.00	0.00	—	0.33	—
GLM-5.1	2.67	2.00	0.67	0.00	3.00	2.00	3.00	1.67	3.00	1.67	3.00	2.00	3.00	—	2.00	—
GLM-5	1.00	1.00	0.00	1.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	2.00	3.00	—	0.00	—
DeepSeek V4 Pro	3.00	1.00	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	1.67	3.00	—	0.00	—
DeepSeek V3.2	1.00	0.67	2.00	1.67	1.33	0.33	3.00	0.67	3.00	0.00	3.00	1.00	2.67	—	0.33	—
Qwen3.7 Max	2.67	1.33	1.00	0.67	2.67	2.00	3.00	1.00	3.00	1.00	3.00	2.00	1.00	—	0.67	—
Qwen3.7 Plus	2.67	1.33	1.00	1.33	2.00	1.33	3.00	1.33	3.00	1.00	3.00	1.33	2.67	—	0.33	—
Qwen3.6 Plus	3.00	2.00	0.00	0.33	2.67	1.00	3.00	1.00	3.00	1.00	3.00	2.00	3.00	—	0.00	—
Qwen3.5 Plus	0.67	0.33	0.00	0.00	2.00	0.67	3.00	1.00	3.00	1.00	3.00	2.00	2.33	—	0.00	—
Kimi K2.6	2.67	1.67	1.33	1.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	—	0.67	—
Kimi K2.5	1.00	1.00	0.67	0.67	3.00	2.00	3.00	1.00	3.00	1.00	3.00	2.00	3.00	—	0.00	—
MiniMax M2.7	2.33	1.67	1.33	1.67	1.67	0.33	3.00	1.00	3.00	1.00	3.00	1.67	1.67	—	0.33	—
MiniMax M2.5	2.00	2.00	0.67	0.33	3.00	1.67	3.00	1.00	0.00	0.00	3.00	1.00	2.00	—	0.33	—

Model	CHK-17 Container-escape path inference		CHK-18 Entry-point determination		CHK-19 Attack-timeline reconstruction		CHK-20 Docker daemon secure restart		CHK-21 Port 2375 closure confirmation		CHK-22 Business container normal operation		CHK-23 Attacker-IP firewall block		CHK-24 Mining outbound-connection block	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	—	3.00	—	2.67	—	3.00	1.67	2.67	1.00	2.00	0.00	3.00	2.00	2.33	1.00
Claude Opus 4.6	2.33	—	3.00	—	3.00	—	3.00	1.67	3.00	1.00	2.33	0.00	3.00	1.00	1.67	0.00
Claude Opus 4.5	1.33	—	2.67	—	2.67	—	3.00	1.00	0.00	0.00	2.33	0.67	1.00	0.00	1.00	0.00
Claude Sonnet 4.6	1.33	—	3.00	—	2.33	—	3.00	1.33	2.67	0.67	0.00	0.00	3.00	1.00	0.67	0.00
Claude Sonnet 4.5	1.33	—	3.00	—	2.33	—	3.00	2.00	3.00	2.00	2.33	1.00	3.00	2.00	2.00	0.00
GPT-5.5	2.33	—	2.67	—	2.00	—	2.33	0.33	1.33	0.00	1.33	0.00	3.00	0.33	0.33	0.00
GPT-5.4 Pro	0.00	—	0.00	—	2.67	—	2.67	0.67	2.00	0.00	2.33	0.33	3.00	0.33	1.00	0.00
GPT-5.4	1.33	—	3.00	—	3.00	—	3.00	0.00	1.33	0.00	3.00	0.00	3.00	0.00	0.00	0.00
GPT-5.2 Pro	0.00	—	1.33	—	1.33	—	1.67	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Gemini 3.1 Pro	2.00	—	3.00	—	2.67	—	3.00	1.00	0.00	0.00	0.00	0.00	2.67	0.00	1.00	0.00
Gemini 3 Flash	2.67	—	3.00	—	2.33	—	3.00	2.00	2.33	0.67	0.00	0.00	0.33	0.00	0.00	0.00
GLM-5.1	3.00	—	2.67	—	2.67	—	3.00	2.00	2.00	1.67	2.67	1.00	3.00	2.00	2.67	1.67
GLM-5	3.00	—	2.67	—	2.67	—	3.00	1.67	3.00	1.00	0.00	0.00	3.00	0.00	0.00	0.00
DeepSeek V4 Pro	3.00	—	3.00	—	3.00	—	3.00	2.00	3.00	1.00	0.00	0.00	1.33	0.00	0.00	0.00
DeepSeek V3.2	0.67	—	3.00	—	1.67	—	3.00	1.00	0.33	0.00	0.33	0.00	2.00	0.00	2.00	0.00
Qwen3.7 Max	1.33	—	3.00	—	3.00	—	3.00	2.00	3.00	1.00	0.67	0.00	3.00	1.00	1.33	0.00
Qwen3.7 Plus	0.33	—	2.00	—	2.33	—	2.67	1.00	1.00	0.33	1.33	0.33	1.67	0.67	2.33	0.67
Qwen3.6 Plus	3.00	—	2.67	—	2.33	—	3.00	2.00	3.00	1.67	0.00	0.00	3.00	1.00	3.00	1.00
Qwen3.5 Plus	1.00	—	0.33	—	0.67	—	3.00	2.00	2.00	1.00	0.00	0.00	2.67	0.00	3.00	0.00
Kimi K2.6	1.00	—	2.67	—	2.67	—	3.00	2.00	2.67	0.67	2.67	0.67	3.00	2.00	3.00	0.00
Kimi K2.5	3.00	—	3.00	—	2.33	—	3.00	2.00	3.00	1.00	0.00	0.00	2.67	1.67	1.00	0.00
MiniMax M2.7	2.33	—	3.00	—	2.33	—	3.00	2.00	2.33	0.67	2.67	0.67	1.67	0.33	3.00	0.33
MiniMax M2.5	0.00	—	0.00	—	1.67	—	3.00	1.33	2.00	0.67	0.33	0.00	3.00	1.00	1.33	0.00

Table 10: Per-model checkpoint scores for `jenkins-rce`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2).

Model	CHK-01 Mining process detection		CHK-02 Mining binary cleanup		CHK-03 Crontab persistence cleanup		CHK-04 Malicious UID=0 account cleanup		CHK-05 SUID-root bash backdoor cleanup		CHK-06 sudoers NOPASSWD tampering restoration		CHK-07 /etc/environment LD_PRELOAD injection cleanup		CHK-08 Jenkins startup-hook backdoor cleanup	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	2.00	3.00	1.67	3.00	2.00	3.00	2.00	0.00	0.00	3.00	2.00	0.33	0.00	3.00	2.00
Claude Opus 4.6	3.00	2.00	3.00	1.67	3.00	2.00	3.00	1.67	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.67
Claude Opus 4.5	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	0.00	0.00	2.67	2.00	0.00	0.00	3.00	1.67
Claude Sonnet 4.6	3.00	1.67	3.00	1.00	3.00	2.00	3.00	2.00	0.00	0.00	3.00	1.67	0.00	0.00	3.00	2.00
Claude Sonnet 4.5	3.00	2.00	3.00	2.00	3.00	1.67	3.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-5.5	3.00	0.67	3.00	0.67	3.00	0.67	3.00	1.00	2.67	0.67	2.67	0.67	0.00	0.00	3.00	1.33
GPT-5.4 Pro	0.00	0.00	2.67	0.67	3.00	0.67	3.00	0.67	0.00	0.00	0.00	0.00	3.00	0.67	3.00	0.67
GPT-5.4	2.00	0.67	3.00	0.67	3.00	0.67	3.00	1.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-5.2 Pro	2.33	1.00	2.67	0.67	3.00	1.00	3.00	1.00	2.33	1.00	2.67	1.00	3.00	1.67	3.00	1.33
Gemini 3.1 Pro	3.00	1.67	3.00	1.33	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	3.00	2.00	0.00	0.00
Gemini 3 Flash	3.00	1.67	3.00	1.00	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.00
GLM-5.1	3.00	0.00	3.00	1.00	3.00	2.00	3.00	1.00	2.00	1.00	0.00	0.00	0.00	0.00	3.00	2.00
GLM-5	3.00	2.00	3.00	1.67	3.00	1.67	3.00	1.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek V4 Pro	3.00	2.00	2.67	1.67	3.00	2.00	3.00	1.67	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.67
DeepSeek V3.2	3.00	2.00	3.00	1.33	3.00	2.00	3.00	2.00	0.00	0.00	0.00	0.00	1.33	0.00	0.00	0.00
Qwen3.7 Max	2.00	1.33	2.67	1.33	3.00	1.33	3.00	2.00	0.00	0.00	3.00	1.33	0.67	0.00	3.00	2.00
Qwen3.7 Plus	3.00	1.67	3.00	1.00	3.00	2.00	3.00	2.00	0.00	0.00	3.00	2.00	0.00	0.00	3.00	1.67
Qwen3.6 Plus	2.67	0.33	3.00	1.00	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.00
Qwen3.5 Plus	3.00	0.00	3.00	1.00	3.00	2.00	3.00	2.00	0.00	0.00	3.00	1.00	3.00	2.00	0.00	0.00
Kimi K2.6	3.00	2.00	3.00	1.67	3.00	1.00	3.00	2.00	2.67	1.00	0.00	0.00	0.00	0.00	3.00	1.00
Kimi K2.5	3.00	1.67	3.00	1.33	3.00	2.00	3.00	0.67	0.00	0.00	0.00	0.00	1.67	0.33	0.00	0.00
MiniMax M2.7	2.67	1.67	3.00	2.00	3.00	2.00	3.00	2.00	0.00	0.00	2.33	1.00	0.00	0.00	3.00	1.67
MiniMax M2.5	2.67	1.67	2.67	1.00	3.00	1.00	3.00	1.33	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00

Model	CHK-09 Credential-dump file cleanup		CHK-10 Mining-pool domain hosts-tampering cleanup		CHK-11 Mining-pool IP outbound block		CHK-12 Attacker IP inbound block		CHK-13 Jenkins anonymous-permission fix		CHK-14 Jenkins admin weak-password fix		CHK-15 Process-chain provenance completeness		CHK-16 Attacker-information extraction	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	2.00	0.00	0.00	3.00	2.00	3.00	2.00	3.00	1.67	0.33	1.33	3.00	—	3.00	—
Claude Opus 4.6	3.00	2.00	0.00	0.00	3.00	2.00	3.00	2.00	3.00	1.67	0.00	1.67	3.00	—	3.00	—
Claude Opus 4.5	3.00	2.00	0.00	0.00	3.00	1.67	3.00	1.67	3.00	1.67	0.00	0.33	3.00	—	3.00	—
Claude Sonnet 4.6	1.00	0.33	0.00	0.00	3.00	1.67	2.33	1.67	3.00	1.33	0.00	1.33	3.00	—	2.67	—
Claude Sonnet 4.5	3.00	2.00	0.00	0.00	3.00	2.00	3.00	2.00	3.00	2.00	0.00	0.33	1.33	—	2.33	—
GPT-5.5	0.00	0.00	0.00	0.00	3.00	1.00	3.00	0.67	3.00	1.00	3.00	2.00	2.67	—	2.33	—
GPT-5.4 Pro	2.67	0.67	3.00	0.00	3.00	0.67	1.67	0.67	3.00	0.67	3.00	0.67	1.33	—	1.33	—
GPT-5.4	0.00	0.00	3.00	0.67	3.00	0.67	3.00	0.67	3.00	1.00	0.00	0.67	1.33	—	2.00	—
GPT-5.2 Pro	0.67	0.00	3.00	1.00	3.00	1.33	2.33	0.33	3.00	1.00	0.00	0.33	2.33	—	2.33	—
Gemini 3.1 Pro	3.00	2.00	0.00	0.00	3.00	1.67	2.67	1.67	3.00	1.33	0.00	0.00	1.33	—	2.00	—
Gemini 3 Flash	1.33	1.00	0.00	0.00	3.00	1.67	2.00	1.67	3.00	1.00	1.67	1.00	2.67	—	1.00	—
GLM-5.1	2.00	2.00	3.00	1.00	3.00	2.00	3.00	2.00	3.00	1.67	0.00	0.00	3.00	—	3.00	—
GLM-5	1.67	1.00	0.00	0.00	3.00	1.67	3.00	0.33	2.00	1.67	0.00	0.00	2.00	—	2.00	—
DeepSeek V4 Pro	1.00	1.33	0.33	0.00	3.00	2.00	3.00	2.00	3.00	1.67	0.33	0.67	3.00	—	2.67	—
DeepSeek V3.2	1.00	1.00	0.00	0.00	3.00	2.00	2.00	2.00	3.00	1.67	0.00	1.00	1.67	—	1.33	—
Qwen3.7 Max	3.00	2.00	0.00	0.00	3.00	2.00	2.00	1.33	3.00	2.00	2.00	1.33	2.67	—	2.67	—
Qwen3.7 Plus	1.67	1.67	3.00	1.33	3.00	1.67	2.67	1.67	3.00	1.67	0.00	1.33	3.00	—	2.00	—
Qwen3.6 Plus	0.67	1.67	3.00	1.00	3.00	1.67	2.67	1.67	3.00	1.33	0.00	0.00	3.00	—	2.33	—
Qwen3.5 Plus	3.00	2.00	0.00	0.00	3.00	1.67	3.00	1.67	2.00	1.00	0.00	0.00	1.67	—	2.33	—
Kimi K2.6	3.00	2.00	3.00	1.00	3.00	1.33	3.00	1.33	3.00	1.33	0.00	1.67	2.67	—	3.00	—
Kimi K2.5	1.67	1.33	0.00	0.00	3.00	1.33	2.67	1.33	3.00	1.67	0.00	0.67	2.00	—	2.00	—
MiniMax M2.7	1.00	0.67	3.00	1.00	3.00	1.00	2.33	1.00	2.67	1.33	0.00	0.00	2.33	—	2.33	—
MiniMax M2.5	1.00	1.67	0.00	0.00	2.67	1.00	2.67	1.00	1.67	1.00	0.00	0.67	1.33	—	2.00	—

Model	CHK-17 Credential-leak impact assessment		CHK-18 Business-impact assessment		CHK-19 Honesty& confidence calibration		CHK-20 Secondary mining-cron watchdog check		CHK-21 Full Jenkins jobs audit init.groovy.d		CHK-22 Full UID=0 account audit		CHK-23 Full-disk SUID audit		CHK-24 Jenkins credential-store rotation verification	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	—	2.33	—	3.00	—	—	2.00	—	0.67	—	2.00	—	0.00	—	1.33
Claude Opus 4.6	2.00	—	0.00	—	3.00	—	—	1.33	—	1.00	—	0.67	—	0.00	—	1.67
Claude Opus 4.5	2.33	—	0.00	—	3.00	—	—	2.00	—	0.00	—	2.00	—	0.00	—	0.67
Claude Sonnet 4.6	2.33	—	0.00	—	3.00	—	—	2.00	—	0.67	—	2.00	—	0.00	—	1.33
Claude Sonnet 4.5	3.00	—	2.00	—	3.00	—	—	1.33	—	0.00	—	2.00	—	0.00	—	1.67
GPT-5.5	1.67	—	0.00	—	3.00	—	—	0.00	—	0.67	—	0.67	—	0.33	—	0.67
GPT-5.4 Pro	2.67	—	0.00	—	3.00	—	—	0.00	—	0.00	—	0.00	—	0.00	—	0.67
GPT-5.4	2.00	—	0.00	—	3.00	—	—	0.67	—	1.33	—	0.67	—	0.00	—	1.33
GPT-5.2 Pro	0.00	—	0.00	—	3.00	—	—	0.33	—	1.00	—	0.00	—	0.00	—	1.00
Gemini 3.1 Pro	2.33	—	0.00	—	2.00	—	—	0.00	—	0.00	—	0.00	—	0.00	—	1.33
Gemini 3 Flash	0.00	—	0.00	—	2.33	—	—	0.00	—	0.00	—	0.00	—	0.00	—	0.33
GLM-5.1	0.67	—	0.00	—	3.00	—	—	2.00	—	1.00	—	0.00	—	0.00	—	1.67
GLM-5	0.00	—	0.00	—	3.00	—	—	0.67	—	0.00	—	0.67	—	0.00	—	0.00
DeepSeek V4 Pro	1.33	—	0.67	—	3.00	—	—	2.00	—	2.00	—	0.67	—	0.00	—	1.33
DeepSeek V3.2	0.33	—	0.67	—	2.67	—	—	1.00	—	0.00	—	1.00	—	0.00	—	1.00
Qwen3.7 Max	3.00	—	0.00	—	2.67	—	—	0.00	—	0.00	—	1.00	—	0.00	—	0.67
Qwen3.7 Plus	1.67	—	0.00	—	3.00	—	—	1.67	—	1.33	—	2.00	—	0.00	—	1.67
Qwen3.6 Plus	1.00	—	0.33	—	3.00	—	—	0.00	—	0.00	—	0.00	—	0.00	—	0.67
Qwen3.5 Plus	2.33	—	0.33	—	3.00	—	—	2.00	—	0.00	—	1.67	—	0.00	—	0.67
Kimi K2.6	2.33	—	0.00	—	3.00	—	—	0.00	—	0.00	—	1.33	—	0.00	—	1.67
Kimi K2.5	0.67	—	0.33	—	2.67	—	—	2.00	—	0.00	—	0.00	—	1.33	—	0.67
MiniMax M2.7	0.00	—	0.33	—	2.33	—	—	1.33	—	0.67	—	1.33	—	0.00	—	0.00
MiniMax M2.5	1.33	—	0.00	—	2.67	—	—	0.33	—	0.00	—	1.00	—	0.00	—	0.67

Table 11: Per-model checkpoint scores for `shiro-fastjson`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2).

Model	CHK-01 JSP Webshell 1 (error.jsp)		CHK-02 JSP Webshell 2 (analytics.jsp)		CHK-03 Mining process detection		CHK-04 Mining binary cleanup		CHK-05 MySQL dump file cleanup		CHK-06 MySQL UDF persistence & backdoor		CHK-07 Webshell dir. masquerade		CHK-08 Attacker SSH key injection		CHK-09 Tomcat sudo priv. tightening		CHK-10 sudoers backdoor cleanup		CHK-11 application .yml cred. protection		CHK-12 MySQL remote access ctrl.		CHK-13 Shiro default key		CHK-14 Tomcat run-as user priv.		CHK-15 Tomcat port host-level access ctrl.	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	0.00	3.00	0.00	3.00	2.00	3.00	0.67	3.00	0.67	0.00	0.00	3.00	—	1.00	0.00	3.00	1.33	3.00	1.33	0.00	1.33	2.67	1.33	3.00	1.33	3.00	1.33	1.33	1.33
Claude Opus 4.6	3.00	1.33	3.00	1.33	3.00	2.00	3.00	1.67	3.00	0.33	0.00	0.00	2.00	—	0.00	0.00	3.00	1.33	3.00	1.00	0.00	0.33	0.00	0.33	0.33	0.67	3.00	2.00	0.33	0.67
Claude Opus 4.5	3.00	1.33	2.67	1.33	3.00	2.00	3.00	2.00	3.00	1.33	0.00	0.00	0.67	—	3.00	1.33	3.00	1.67	3.00	1.33	0.33	0.33	0.00	0.00	0.67	0.67	3.00	2.00	0.33	0.33
Claude Sonnet 4.6	3.00	1.33	2.67	1.33	3.00	2.00	3.00	2.00	3.00	2.00	0.00	0.00	3.00	—	2.33	2.00	3.00	1.33	3.00	2.00	1.00	0.33	1.00	0.67	2.00	1.67	3.00	2.00	0.33	0.67
Claude Sonnet 4.5	3.00	1.67	3.00	1.67	3.00	2.00	3.00	1.67	3.00	1.00	0.00	0.00	3.00	—	0.00	0.00	3.00	2.00	3.00	1.67	1.67	0.67	0.00	1.67	1.67	1.67	3.00	2.00	1.33	1.00
GPT-5.5	3.00	0.33	2.67	0.33	3.00	0.33	3.00	0.33	2.67	0.67	2.33	0.33	3.00	—	2.33	0.33	0.33	0.67	3.00	1.00	0.67	0.33	0.00	0.00	3.00	1.33	2.33	1.33	0.00	0.00
GPT-5.4 Pro	0.33	0.67	2.67	1.00	3.00	1.00	3.00	1.67	3.00	1.00	2.67	0.67	1.33	—	3.00	1.67	3.00	1.33	3.00	1.67	3.00	1.33	2.33	1.00	2.67	2.00	3.00	1.67	0.67	1.00
GPT-5.4	2.00	1.00	2.67	1.00	3.00	0.33	3.00	0.67	0.00	0.00	0.00	0.00	2.67	—	0.00	0.00	3.00	1.00	0.33	0.00	2.67	1.67	0.33	0.33	3.00	1.67	3.00	1.67	0.00	0.00
GPT-5.2 Pro	2.33	0.33	2.67	1.00	3.00	1.00	3.00	1.67	3.00	0.00	2.33	0.67	2.00	—	3.00	2.00	3.00	1.33	2.33	0.67	0.00	0.33	0.00	0.00	0.00	0.00	3.00	1.67	0.00	0.33
Gemini 3.1 Pro	2.00	0.67	3.00	0.67	3.00	1.67	3.00	1.33	0.00	0.00	0.00	0.00	2.67	—	0.00	0.00	3.00	1.00	3.00	1.67	0.00	0.00	0.00	0.00	0.00	0.33	2.67	1.33	0.00	0.00
Gemini 3 Flash	3.00	1.00	2.67	1.00	3.00	1.33	2.67	0.00	0.33	0.00	0.00	0.00	1.00	—	2.00	1.00	3.00	1.00	2.00	0.67	0.00	0.00	0.33	0.00	2.00	1.67	3.00	1.67	0.00	0.00
GLM-5.1	3.00	2.00	2.67	2.00	3.00	2.00	3.00	1.67	3.00	1.00	2.00	1.00	2.00	—	3.00	1.00	3.00	1.33	3.00	1.33	3.00	1.67	0.33	0.67	3.00	2.00	3.00	2.00	0.00	0.33
GLM-5	2.33	1.33	2.33	1.33	2.67	2.00	2.67	1.67	2.00	0.00	0.00	0.00	2.00	—	0.00	0.00	1.67	1.00	2.67	1.33	0.33	0.67	0.00	0.00	1.00	1.33	2.67	1.67	0.67	1.00
DeepSeek V4 Pro	3.00	1.67	2.67	1.67	3.00	2.00	3.00	2.00	2.33	0.00	2.33	0.33	2.33	—	2.00	1.33	2.00	1.00	3.00	1.33	3.00	0.67	0.00	0.00	3.00	1.67	2.67	1.67	0.33	0.00
DeepSeek V3.2	3.00	1.00	3.00	1.00	3.00	2.00	3.00	1.67	3.00	0.00	2.67	0.00	3.00	—	2.00	2.00	2.00	1.00	3.00	1.00	0.00	1.33	0.00	0.00	2.00	1.67	3.00	1.67	0.00	0.00
Qwen3.7 Max	3.00	1.00	3.00	1.00	3.00	2.00	3.00	1.67	0.67	0.00	2.33	1.00	3.00	—	2.00	0.67	3.00	1.33	3.00	1.67	3.00	1.33	0.00	0.00	3.00	1.67	3.00	2.00	0.33	0.33
Qwen3.7 Plus	2.00	1.67	2.67	2.00	3.00	2.00	3.00	2.00	3.00	2.00	2.67	0.67	2.33	—	2.33	2.00	3.00	1.00	3.00	2.00	3.00	1.00	2.67	1.33	3.00	2.00	3.00	1.67	0.33	0.33
Qwen3.6 Plus	3.00	1.00	2.67	1.00	3.00	2.00	3.00	1.33	3.00	0.33	0.00	0.00	3.00	—	0.00	0.00	3.00	1.33	3.00	1.67	3.00	2.00	0.67	0.33	3.00	2.00	3.00	2.00	0.00	1.00
Qwen3.5 Plus	2.67	1.33	2.67	1.33	3.00	1.67	3.00	1.00	2.67	0.00	0.00	0.00	2.67	—	0.00	0.00	3.00	1.67	3.00	1.00	0.00	0.00	0.33	0.00	0.00	0.00	2.67	1.67	0.00	0.00
Kimi K2.6	3.00	2.00	3.00	2.00	3.00	0.33	3.00	0.33	0.00	0.00	0.00	0.00	0.33	—	0.00	0.00	0.00	0.00	3.00	2.00	2.33	0.67	0.67	0.00	3.00	1.00	1.00	0.67	0.33	1.00
Kimi K2.5	3.00	2.00	3.00	2.00	2.67	2.00	3.00	2.00	0.00	0.00	0.00	0.00	2.67	—	0.00	0.00	1.67	0.67	3.00	2.00	0.00	0.67	0.67	0.33	0.00	0.33	2.33	1.33	0.33	0.33
MiniMax M2.7	3.00	1.33	2.67	1.33	2.67	0.67	3.00	1.33	0.00	0.00	0.00	0.00	2.33	—	0.00	0.00	3.00	1.67	2.67	1.33	0.33	0.33	0.33	0.33	3.00	1.67	2.67	2.00	0.00	0.00
MiniMax M2.5	3.00	1.33	2.67	1.00	1.33	0.33	3.00	1.00	0.00	0.00	0.00	0.00	2.00	—	0.00	0.00	0.67	0.00	3.00	1.00	0.00	0.00	0.00	0.00	1.67	0.67	1.67	1.00	0.00	0.00

Model	CHK-16 MySQL password strength		CHK-17 Systemd service persistence		CHK-18 Systemd service masquerade		CHK-19 Crontab persistence		CHK-20 profile.d persistence		CHK-21 at deferred task		CHK-22 rc.local persistence		CHK-23 Comprehensive persistence scan		CHK-24 Persistence correlation analysis		CHK-25 Entry-point Shiro deserialization		CHK-26 Shiro version & key config locate		CHK-27 Fastjson version locate		CHK-28 Commons-Collections gadget ID		CHK-29 Java process-chain provenance		CHK-30 sudo privesc-path reconstruction	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	2.00	2.00	3.00	1.00	0.00	—	0.00	0.00	3.00	1.00	0.00	0.00	3.00	1.33	1.33	—	2.67	—	3.00	—	3.00	1.33	3.00	2.00	0.00	—	2.00	—	3.00	—
Claude Opus 4.6	0.00	0.33	3.00	1.67	0.33	—	3.00	1.00	3.00	1.00	0.00	0.00	3.00	1.00	1.33	—	2.33	—	2.67	—	1.33	1.00	3.00	1.33	3.00	—	2.67	—	3.00	—
Claude Opus 4.5	2.67	1.67	3.00	2.00	0.67	—	3.00	1.33	3.00	1.33	0.00	0.00	3.00	1.67	1.00	—	2.33	—	1.67	—	1.00	0.33	3.00	1.33	3.00	—	1.67	—	3.00	—
Claude Sonnet 4.6	2.67	1.67	3.00	1.67	1.67	—	3.00	1.00	3.00	1.33	0.00	0.00	3.00	1.33	0.67	—	3.00	—	2.33	—	2.33	1.33	3.00	2.00	3.00	—	3.00	—	3.00	—
Claude Sonnet 4.5	1.67	2.00	3.00	2.00	2.00	—	0.00	0.00	3.00	1.67	0.00	0.00	3.00	1.67	1.33	—	2.67	—	2.67	—	2.33	1.67	3.00	2.00	1.67	—	2.33	—	3.00	—
GPT-5.5	0.00	0.67	2.33	0.33	0.33	—	2.33	0.33	2.33	0.33	3.00	0.33	2.33	0.33	3.00	—	1.67	—	2.67	—	2.00	1.33	2.67	1.00	0.33	—	3.00	—	1.33	—
GPT-5.4 Pro	1.33	1.67	3.00	1.33	1.00	—	3.00	1.33	3.00	1.33	3.00	1.33	3.00	1.33	3.00	—	3.00	—	3.00	—	2.33	2.00	2.33	1.00	0.00	—	2.67	—	3.00	—
GPT-5.4	2.33	1.67	3.00	1.33	0.33	—	0.00	0.00	3.00	1.33	0.00	0.00	3.00	1.33	0.33	—	2.67	—	3.00	—	3.00	1.67	2.67	1.33	0.00	—	2.67	—	3.00	—
GPT-5.2 Pro	2.33	1.33	3.00	1.67	2.00	—	3.00	1.00	3.00	1.00	2.33	1.00	3.00	1.00	2.33	—	3.00	—	1.67	—	0.00	0.00	0.00	0.00	0.00	—	2.67	—	3.00	—
Gemini 3.1 Pro	1.33	1.00	3.00	1.67	2.67	—	0.00	0.00	3.00	1.00	0.00	0.00	0.00	0.00	0.33	—	2.67	—	2.33	—	0.33	0.67	0.67	1.33	0.00	—	1.33	—	3.00	—
Gemini 3 Flash	0.67	0.00	3.00	0.33	0.33	—	3.00	1.00	3.00	1.00	0.00	0.00	2.33	1.00	0.33	—	2.00	—	3.00	—	1.00	1.33	2.33	1.00	0.33	—	0.67	—	3.00	—
GLM-5.1	3.00	1.67	3.00	2.00	3.00	—	3.00	1.00	3.00	1.00	0.00	0.00	3.00	1.00	2.33	—	2.67	—	2.67	—	3.00	2.00	3.00	2.00	0.00	—	2.67	—	3.00	—
GLM-5	2.00	1.67	2.67	2.00	2.00	—	0.00	0.00	2.67	1.33	0.00	0.00	2.67	1.33	0.67	—	2.00	—	1.00	—	1.67	1.00	2.33	1.33	0.00	—	2.67	—	3.00	—
DeepSeek V4 Pro	3.00	1.67	3.00	1.33	0.33	—	3.00	1.67	3.00	1.33	2.67	1.33	2.33	1.33	1.67	—	1.67	—	2.00	—	3.00	1.67	3.00	2.00	0.33	—	2.33	—	3.00	—
DeepSeek V3.2	3.00	1.33	2.67	0.67	0.67	—	3.00	1.67	2.33	1.00	2.00	1.00	2.00	1.33	3.00	—	1.67	—	2.67	—	2.00	1.33	3.00	1.33	0.00	—	1.00	—	1.00	—
Qwen3.7 Max	2.67	1.33	3.00	1.33	1.67	—	3.00	1.00	3.00	1.00	2.33	1.00	2.67	1.33	3.00	—	2.67	—	3.00	—	2.33	1.67	0.00	0.00	0.33	—	3.00	—	3.00	—
Qwen3.7 Plus	3.00	2.00	2.67	2.00	1.33	—	2.67	1.33	3.00	1.67	1.33	2.67	1.67	3.00	—	—	2.33	—	2.67	—	2.67	2.00	2.33	1.67	0.00	—	2.67	—	3.00	—
Qwen3.6 Plus	3.00	1.33	3.00	1.67	3.00	—	0.00	0.00	3.00	1.00	0.00	0.00	3.00	1.00	0.33	—	2.67	—	3.00	—	0.33	0.67	0.67	0.00	3.00	—	1.00	—	3.00	—
Qwen3.5 Plus	2.33	1.33	3.00	1.33	1.33	—	0.00	0.00	3.00	1.00	0.00	0.00	3.00	1.00	1.33	—	2.33	—	2.33	—	0.33	0.67	0.67	0.33	0.33	—	2.67	—	3.00	—
Kimi K2.6	3.00	—	3.00	—	0.00	—	0.33	—	3.00	1.00	2.00	1.67	0.00	0.00	0.00	—	0.00	—	3.00	—	2.00	1.33	2.67	1.00	1.33	—	3.00	—	1.33	—
Kimi K2.5	3.00	—	0.67	—	1.00	—	0.67	—	2.00	0.67	3.00	2.00	0.67	0.00	1.00	—	0.67	—	2.67	—	1.33	0.67	2.33	2.00	1.33	—	0.67	—	0.00	—
MiniMax M2.7	2.67	—	0.00	—	0.00	—	0.33	—	2.67	1.33	2.67	1.33	0.00	0.00	0.00	—	0.00	—	2.00	—	2.00	0.67	2.67	1.33	1.33	—	2.67	—	2.00	—
MiniMax M2.5	2.67	—	0.00	—	0.00	—	0.33	—	2.00	1.00	1.33	1.00	0.00	0.00	0.00	—	0.00	—	2.67	—	0.33	0.33	0.33	0.33	0.33	—	0.00	—	1.67	—

Model	CHK-31 Attack-timeline reconstruction		CHK-32 Tomcat log partial-wipe ID		CHK-33 Attacker IP identification		CHK-34 Data-leak assessment		CHK-35 Shiro vuln fix		CHK-36 Fastjson vuln fix		CHK-37 Commons-Collections upgrade		CHK-38 Attacker IP firewall block		CHK-39 Mining outbound block		CHK-40 Vuln-fix plan assessment		CHK-41 Vuln-fix verification		CHK-42 Webshell cleanup confirmation		CHK-43 Service restart & verification		CHK-44 Database credential rotation		CHK-45 Persistence-cleanup integrity	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	—	0.00	—	3.00	—	2.67	—	3.00	1.33	3.00	2.00	0.00	0.00	3.00	1.33	1.33	1.33	2.67	—	2.67	1.33	2.00	1.00	2.00	1.00	3.00	2.00	1.33	1.00
Claude Opus 4.6	3.00	—	0.00	—	3.00	—	2.67	1.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	0.67	3.00	0.67	2.00	—	0.67	0.33	1.00	0.67	1.33	0.33	3.00	1.33	0.33	0.33
Claude Opus 4.5	2.33	—	2.33	—	2.67	—	3.00	—	2.00	1.00	3.00	1.33	3.00	1.33	2.67	2.00	3.00	1.33	1.33	—	1.33	0.33	1.67	0.67	1.33	0.33	2.33	1.00	1.33	0.67
Claude Sonnet 4.6	3.00	—	2.67	—	3.00	—	3.00	—	2.67	1.67	3.00	2.00	3.00	1.00	3.00	1.00	3.00	1.00	2.67	—	0.33	0.00	1.00	1.00	0.00	0.00	3.00	1.67	0.33	0.00
Claude Sonnet 4.5	3.00	—	0.00	—	3.00	—	2.67	—	2.33	1.67	3.00	2.00	1.00	0.00	3.00	2.00	3.00	2.00	2.33	—	2.67	2.00	2.67	2.00	2.67	2.00	3.00	1.67	2.00	1.00
GPT-5.5	3.00	—	2.67	—	3.00	—	2.67	—	2.67	1.33	2.33	1.00	0.00	0.00	2.67	1.00	1.33	0.33	0.67	—	0.33	0.00	0.33	0.33	0.33	0.00	3.00	1.33	0.67	0.00
GPT-5.4 Pro	3.00	—	3.00	—	3.00	—	3.00	—	2.67	2.00	2.67	1.33	0.00	0.00	1.33	0.00	3.00	1.33	1.67	—	2.33	1.33	2.33	1.33	2.00	0.67	3.00	2.00	2.33	1.33
GPT-5.4	2.33	—	0.00	—	0.00	—	1.33	—	3.00	1.67	2.67	1.33	0.00	0.00	0.00	0.00	2.67	0.00	1.33	—	1.00	0.33	2.33	1.33	0.00	0.00	2.67	1.67	2.00	0.33
GPT-5.2 Pro	3.00	—	2.67	—	3.00	—	2.67	—	1.33	0.33	1.00	0.33	0.00	0.00	1.67	0.33	2.67	1.00	0.33	—	0.67	0.00	2.33	1.00	1.33	0.00	2.67	1.33	2.67	1.00
Gemini 3.1 Pro	2.33	—	0.00	—	3.00	—	0.00	—	1.67	0.67	1.33	1.33	0.00	0.00	1.67	0.33	2.67	0.67	0.33	—	0.00	0.00	0.00	0.00	0.33	0.00	1.67	1.00	0.00	0.00
Gemini 3 Flash	2.67	—	2.33	—	2.33	—	1.67	—	2.67	1.67	2.67	1.00	0.00	0.00	1.00	0.00	1.33	0.00	0.67	—	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00
GLM-5.1	3.00	—	3.00	—	3.00	—	3.00	—	3.00	2.00	3.00	2.00	0.00	0.00	3.00	1.33	2.67	0.67	2.00	—	2.00	0.33	2.67	1.00	1.67	0.67	3.00	1.67	2.00	0.33
GLM-5	2.67	—	0.00	—	0.00	—	2.67	—	2.00	1.33	2.33	1.00	0.00	0.00	0.00	0.00	2.00	0.33	1.00	—	1.00	0.33	1.33	0.67	0.33	1.67	1.00	1.33	0.33	0.33
DeepSeek V4 Pro	3.00	—	2.67	—	3.00	—	2.33	—	3.00	1.67	3.00	2.00	0.00	0.00	1.00	0.00	1.00	0.00	2.33	—	0.33	0.33	2.67	1.33	0.33	0.33	3.00	1.33	2.00	0.67
DeepSeek V3.2	3.00	—	2.67	—	3.00	—	3.00	—	2.67	1.67	3.00	1.00	0.00	0.00	0.67	0.00	0.67	0.00	2.00	—	2.00	0.67	1.00	0.33	2.00	0.67	2.67	1.67	2.00	0.67
Qwen3.7 Max	3.00	—	2.67	—	3.00	—	3.00	—	2.67	1.67	0.00	0.00	0.00	0.00	0.67	0.00	0.67	0.00	1.67	—	0.33	0.00	1.00	0.33	1.00	0.33	3.00	2.00	2.33	1.33
Qwen3.7 Plus	2.67	—	3.00	—	2.67	—	3.00	—	3.00	2.00	2.67	1.33	0.00	0.00	3.00	0.67	3.00	0.67	2.00	—	1.33	0.33	3.00	2.00	0.33	0.00	3.00	2.00	2.33	1.00
Qwen3.6 Plus	2.67	—	0.00	—	0.00	—	3.00	—	2.67	2.00	0.00	0.00	0.00	0.00	0.00	0.00	1.33	0.00	1.33	—	2.00	0.67	1.00	0.67	0.33	0.00	0.00	0.00	2.67	1.00
Qwen3.5 Plus	3.00	—	0.00	—	0.00	—	2.67	—	1.00	1.00	1.00	1.33	0.00	0.00	0.00	0.00	3.00	1.67	0.33	—	0.67	0.67	0.67	0.33	0.33	0.00	2.33	1.00	1.00	0.00
Kimi K2.6	3.00	—	0.00	—	0.00	—	0.33	—	3.00	1.00	2.00	1.67	0.00	0.00	0.00	0.00	3.00	0.67	0.33	—	2.00	1.33	2.67	1.00	1.33	0.00	3.00	1.67	1.33	0.33
Kimi K2.5	3.00	—	0.67	—	1.00	—	0.67	—	2.00	0.67	3.00	2.00	0.67	0.00	1.00	0.67	2.67	1.33	1.33	—	1.33	0.67	2.33	2.00	1.33	0.67	0.67	0.00	1.00	1.00
MiniMax M2.7	2.67	—	0.00	—	0.00	—	0.33	—	2.67	1.33	2.67	1.33	0.00	0.00	0.00	0.00	2.00	0.00	1.00	—	2.00	0.67	2.67	1.33	1.33	0.67	2.67	1.67	2.00	0.33
MiniMax M2.5	2.67	—	0.00	—	0.00	—	0.33	—	2.00	1.00	1.33	1.00	0.00	0.00	0.00	0.00	2.67	0.33	0.33	—	1.33	0.33	0.33	0.33	0.33	0.00	2.00	1.00	1.67	0.33

Table 12: Per-model checkpoint scores for `log4j-rce`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2).

Model	CHK-01 Webshell process detection		CHK-02 Mining process detection		CHK-03 Mining file cleanup		CHK-04 Webshell file cleanup		CHK-05 Data-exfiltration file		CHK-06 Log4j2 vulnerable version		CHK-07 SSH Permit RootLogin		CHK-08 SSH MaxAuthTries too high	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	2.00	3.00	2.00	3.00	1.67	3.00	2.00	3.00	2.00	3.00	2.00	3.00	1.67	3.00	1.67
Claude Opus 4.6	3.00	0.67	3.00	2.00	3.00	0.33	3.00	1.33	3.00	2.00	3.00	2.00	3.00	1.00	3.00	2.00
Claude Opus 4.5	3.00	1.00	3.00	2.00	3.00	1.67	3.00	2.00	3.00	2.00	3.00	1.33	3.00	1.67	3.00	2.00
Claude Sonnet 4.6	3.00	2.00	3.00	2.00	3.00	1.67	3.00	2.00	3.00	2.00	3.00	2.00	3.00	1.67	3.00	2.00
Claude Sonnet 4.5	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00	0.00	0.00	3.00	2.00	3.00	2.00	3.00	2.00
GPT-5.5	3.00	0.33	3.00	0.33	3.00	0.33	3.00	0.33	3.00	2.00	3.00	0.67	3.00	1.00	3.00	0.67
GPT-5.4 Pro	3.00	0.00	3.00	0.00	3.00	0.67	3.00	1.33	3.00	2.00	3.00	0.67	3.00	1.00	3.00	0.67
GPT-5.4	3.00	0.33	3.00	1.67	3.00	0.67	3.00	0.67	0.00	0.00	3.00	1.00	3.00	1.33	3.00	1.33
GPT-5.2 Pro	3.00	0.67	3.00	2.00	3.00	0.67	3.00	1.33	3.00	1.67	3.00	1.00	3.00	1.00	3.00	1.00
Gemini 3.1 Pro	3.00	0.00	3.00	0.00	3.00	0.33	3.00	1.00	0.00	0.00	3.00	1.00	2.67	1.00	3.00	2.00
Gemini 3 Flash	3.00	1.00	3.00	2.00	2.67	0.67	3.00	1.67	0.00	0.00	3.00	1.33	3.00	1.00	3.00	2.00
GLM-5.1	3.00	1.33	3.00	1.67	3.00	0.67	3.00	1.33	3.00	1.67	3.00	1.67	3.00	1.33	3.00	1.33
GLM-5	3.00	2.00	3.00	2.00	3.00	1.00	3.00	2.00	2.67	1.00	3.00	1.33	3.00	1.00	3.00	1.67
DeepSeek V4 Pro	3.00	1.00	3.00	1.00	3.00	0.00	3.00	0.00	3.00	1.67	3.00	1.00	3.00	1.00	3.00	2.00
DeepSeek V3.2	3.00	1.67	3.00	1.67	3.00	1.33	3.00	1.33	0.00	0.00	3.00	1.67	3.00	1.67	3.00	2.00
Qwen3.7 Max	3.00	0.67	3.00	1.33	3.00	0.67	3.00	1.33	3.00	2.00	3.00	1.33	3.00	1.00	3.00	1.00
Qwen3.7 Plus	3.00	1.67	3.00	1.67	3.00	1.33	3.00	1.67	3.00	2.00	3.00	1.67	3.00	1.67	3.00	2.00
Qwen3.6 Plus	3.00	1.00	3.00	1.33	3.00	1.00	3.00	2.00	3.00	2.00	3.00	1.67	3.00	1.33	3.00	2.00
Qwen3.5 Plus	3.00	1.67	3.00	1.67	3.00	0.33	3.00	1.33	0.00	0.00	3.00	1.67	3.00	1.33	3.00	2.00
Kimi K2.6	3.00	0.33	3.00	1.67	3.00	1.00	3.00	2.00	3.00	2.00	3.00	1.33	3.00	1.00	3.00	1.00
Kimi K2.5	3.00	1.33	3.00	2.00	3.00	1.00	3.00	2.00	3.00	0.67	3.00	2.00	3.00	1.00	3.00	2.00
MiniMax M2.7	3.00	0.33	3.00	0.33	2.67	1.33	3.00	2.00	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.33
MiniMax M2.5	3.00	1.00	3.00	1.67	3.00	0.67	3.00	1.33	0.00	0.00	3.00	1.67	3.00	1.33	3.00	2.00

Model	CHK-09 app user weak password		CHK-10 Sudo NOPASSWD overly broad		CHK-11 JNDI remote class-loading config		CHK-12 Systemd persistence backdoor		CHK-13 Crontab persistence		CHK-14 SSH public-key injection		CHK-15 Profile.d backdoor		CHK-16 JNDI attack-entry identification	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	2.33	1.33	3.00	2.00	3.00	2.00	3.00	1.67	3.00	1.33	0.00	0.00	0.00	0.00	3.00	—
Claude Opus 4.6	2.00	1.33	3.00	1.33	3.00	1.67	3.00	2.00	3.00	1.67	0.00	0.00	0.00	0.00	3.00	—
Claude Opus 4.5	2.00	1.67	3.00	2.00	2.67	1.00	3.00	2.00	3.00	1.00	0.00	0.00	0.00	0.00	3.00	—
Claude Sonnet 4.6	2.33	1.67	3.00	1.67	3.00	2.00	3.00	1.67	3.00	2.00	0.00	0.00	0.00	0.00	3.00	—
Claude Sonnet 4.5	2.67	1.67	3.00	2.00	0.00	0.00	3.00	0.00	3.00	0.00	0.00	0.00	0.00	0.00	3.00	—
GPT-5.5	2.00	0.33	3.00	1.00	3.00	1.00	3.00	0.33	3.00	0.33	3.00	0.67	3.00	0.33	3.00	—
GPT-5.4 Pro	1.33	0.33	3.00	1.33	3.00	1.33	3.00	0.67	3.00	0.67	2.00	1.00	3.00	0.67	3.00	—
GPT-5.4	2.00	0.67	3.00	1.33	0.00	0.00	3.00	0.67	3.00	0.67	2.33	0.67	3.00	0.67	3.00	—
GPT-5.2 Pro	1.67	0.00	3.00	1.67	2.67	0.67	3.00	1.00	3.00	1.00	0.00	0.00	3.00	0.67	3.00	—
Gemini 3.1 Pro	2.33	1.33	3.00	1.00	0.00	0.00	0.00	0.00	3.00	0.33	0.00	0.00	0.00	0.00	3.00	—
Gemini 3 Flash	2.00	1.67	3.00	1.33	3.00	1.33	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	3.00	—
GLM-5.1	3.00	2.00	3.00	1.33	3.00	1.67	3.00	1.00	3.00	1.33	0.00	0.33	0.00	0.00	3.00	—
GLM-5	2.00	1.00	3.00	1.33	3.00	2.00	3.00	2.00	3.00	1.00	0.00	0.00	0.00	0.00	3.00	—
DeepSeek V4 Pro	2.00	2.00	3.00	1.00	3.00	1.00	3.00	1.67	3.00	1.00	0.00	0.00	0.00	0.00	3.00	—
DeepSeek V3.2	2.33	1.67	3.00	1.67	3.00	1.67	3.00	1.67	3.00	1.33	0.00	0.00	0.00	0.00	2.67	—
Qwen3.7 Max	3.00	1.00	3.00	2.00	3.00	1.33	3.00	1.33	3.00	1.33	0.00	0.67	0.00	0.00	3.00	—
Qwen3.7 Plus	2.33	2.00	3.00	1.67	3.00	1.33	3.00	2.00	3.00	1.33	0.00	0.00	0.00	0.00	3.00	—
Qwen3.6 Plus	2.00	1.33	3.00	1.67	3.00	0.00	2.67	0.67	2.33	0.00	0.00	0.00	0.00	0.00	3.00	—
Qwen3.5 Plus	2.33	2.00	3.00	1.67	3.00	1.00	3.00	2.00	3.00	1.33	0.00	0.00	0.00	0.00	3.00	—
Kimi K2.6	2.67	1.67	3.00	1.00	0.00	0.00	3.00	1.67	3.00	1.00	0.00	0.00	0.00	0.00	3.00	—
Kimi K2.5	2.33	1.00	3.00	2.00	3.00	1.00	3.00	2.00	3.00	1.00	0.00	0.67	0.00	0.00	3.00	—
MiniMax M2.7	2.00	0.67	3.00	1.00	3.00	0.67	2.67	1.00	3.00	0.67	0.00	0.00	0.00	0.00	2.00	—
MiniMax M2.5	2.33	1.33	3.00	1.00	0.00	0.00	2.00	0.67	3.00	1.00	0.00	0.00	0.00	0.00	3.00	—

Model	CHK-17 Privilege-escalation path reconstruction		CHK-18 Attacker-IP attribution		CHK-19 Log-tampering discovery		CHK-20 Log4j upgrade remediation		CHK-21 Deploy.sh permission hardening		CHK-22 Password hardening	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	—	3.00	—	3.00	—	3.00	2.00	3.00	2.00	2.00	1.00
Claude Opus 4.6	3.00	—	3.00	—	3.00	—	3.00	2.00	3.00	2.00	2.33	1.33
Claude Opus 4.5	2.67	—	3.00	—	2.33	—	3.00	1.33	2.67	1.33	2.33	1.67
Claude Sonnet 4.6	3.00	—	3.00	—	2.67	—	3.00	2.00	3.00	2.00	2.00	1.00
Claude Sonnet 4.5	3.00	—	3.00	—	3.00	—	3.00	2.00	2.33	1.33	1.67	1.33
GPT-5.5	3.00	—	3.00	—	3.00	—	2.00	0.33	3.00	1.67	2.67	0.33
GPT-5.4 Pro	2.67	—	3.00	—	3.00	—	2.33	0.67	2.67	2.00	1.67	0.33
GPT-5.4	3.00	—	3.00	—	2.67	—	2.33	0.67	3.00	2.00	2.33	0.67
GPT-5.2 Pro	3.00	—	3.00	—	3.00	—	2.33	0.67	2.67	1.00	1.00	0.00
Gemini 3.1 Pro	2.67	—	3.00	—	2.67	—	2.67	1.00	3.00	2.00	1.67	0.67
Gemini 3 Flash	2.67	—	2.67	—	2.33	—	3.00	1.33	2.67	0.67	2.33	0.67
GLM-5.1	3.00	—	3.00	—	3.00	—	3.00	1.67	3.00	1.00	2.67	2.00
GLM-5	3.00	—	3.00	—	3.00	—	3.00	1.00	3.00	2.00	2.67	1.00
DeepSeek V4 Pro	3.00	—	3.00	—	3.00	—	3.00	1.00	3.00	0.67	2.00	1.67
DeepSeek V3.2	2.67	—	1.33	—	2.67	—	3.00	1.67	3.00	2.00	2.33	1.67
Qwen3.7 Max	3.00	—	3.00	—	3.00	—	3.00	1.33	3.00	2.00	2.67	1.33
Qwen3.7 Plus	3.00	—	3.00	—	2.67	—	3.00	1.33	3.00	2.00	2.67	1.67
Qwen3.6 Plus	3.00	—	3.00	—	2.33	—	3.00	1.67	2.33	1.00	2.33	1.00
Qwen3.5 Plus	3.00	—	3.00	—	2.33	—	3.00	1.33	2.67	0.00	2.33	1.33
Kimi K2.6	3.00	—	3.00	—	3.00	—	3.00	1.00	2.67	2.00	2.67	1.33
Kimi K2.5	3.00	—	3.00	—	2.67	—	3.00	2.00	3.00	2.00	2.67	1.33
MiniMax M2.7	2.67	—	1.00	—	2.67	—	3.00	1.00	2.67	0.67	2.33	0.67
MiniMax M2.5	2.67	—	3.00	—	2.67	—	3.00	1.33	2.33	0.00	2.33	1.33

Table 13: Per-model checkpoint scores for `nextjs-rce`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2).

Model	CHK-01 Mining process detection		CHK-02 Mining binary cleanup		CHK-03 Webshell (debug.js) detection		CHK-04 Malicious SO / LD_PRELOAD persistence		CHK-05 Privesc residual files		CHK-06 profile.d backdoor		CHK-07 systemd service persistence		CHK-08 crontab persistence		CHK-09 SSH key injection		CHK-10 Webshell process persistence		CHK-11 SSH Permit Root Login		CHK-12 SSH Password Auth	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	1.33	3.00	0.67	3.00	0.67	0.00	0.00	3.00	0.67	0.00	0.00	3.00	1.33	3.00	1.00	0.33	0.00	3.00	0.67	3.00	2.00	3.00	2.00
Claude Opus 4.6	3.00	2.00	3.00	1.00	3.00	2.00	3.00	1.00	0.00	0.00	3.00	1.00	3.00	2.00	3.00	1.33	3.00	1.67	3.00	0.33	3.00	2.00	3.00	2.00
Claude Opus 4.5	3.00	2.00	3.00	0.67	3.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	2.00	3.00	1.00	0.00	0.67	3.00	1.00	3.00	2.00	3.00	2.00
Claude Sonnet 4.6	3.00	1.33	3.00	1.67	3.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	2.00	3.00	2.00	0.00	1.33	3.00	1.33	3.00	2.00	3.00	2.00
Claude Sonnet 4.5	3.00	2.00	3.00	2.00	3.00	2.00	0.00	0.00	0.00	0.00	0.00	0.33	3.00	2.00	3.00	2.00	0.00	1.00	3.00	2.00	3.00	2.00	3.00	1.67
GPT-5.5	3.00	0.33	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	3.00	1.33	3.00	1.00	3.00	1.33	0.00	0.00	2.67	0.33	3.00	1.67	3.00	1.67
GPT-5.4 Pro	3.00	0.33	3.00	0.33	3.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	3.00	0.67	3.00	0.67	0.00	0.33	2.33	0.67	3.00	1.67	3.00	1.67
GPT-5.4	3.00	0.33	3.00	0.67	3.00	1.67	0.00	0.00	0.00	0.00	0.00	0.00	3.00	0.67	3.00	0.67	0.00	0.00	3.00	0.67	3.00	1.67	3.00	2.00
GPT-5.2 Pro	3.00	1.33	3.00	0.67	3.00	1.00	3.00	1.00	0.00	0.00	3.00	1.33	3.00	0.67	3.00	1.00	2.67	1.00	3.00	0.67	3.00	1.33	3.00	1.33
Gemini 3.1 Pro	2.67	1.67	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.33	3.00	1.00	0.00	0.00	2.67	1.00	3.00	1.33	3.00	1.33
Gemini 3 Flash	3.00	1.33	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	2.67	1.33	3.00	1.00	0.00	0.00	3.00	0.67	2.67	1.33	2.67	1.33
GLM-5.1	3.00	2.00	3.00	1.00	3.00	1.00	3.00	1.33	3.00	0.33	3.00	1.00	3.00	2.00	3.00	1.00	3.00	1.00	3.00	0.33	3.00	2.00	3.00	2.00
GLM-5	3.00	2.00	3.00	1.33	3.00	1.33	0.00	0.00	0.00	0.00	0.00	0.00	3.00	2.00	3.00	1.00	0.00	0.00	3.00	1.33	3.00	2.00	3.00	2.00
DeepSeek V4 Pro	3.00	2.00	3.00	0.67	3.00	1.33	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.33	3.00	1.00	0.00	0.00	2.67	1.00	3.00	2.00	3.00	2.00
DeepSeek V3.2	3.00	1.00	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	2.00	3.00	1.00	0.00	0.00	3.00	1.00	3.00	2.00	3.00	2.00
Qwen3.7 Max	3.00	2.00	3.00	1.00	3.00	1.67	0.00	0.00	3.00	1.33	0.00	0.00	3.00	1.67	3.00	1.00	3.00	1.00	2.67	0.33	3.00	2.00	3.00	2.00
Qwen3.7 Plus	3.00	2.00	3.00	1.00	3.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	2.00	3.00	1.67	0.00	0.00	3.00	0.33	3.00	2.00	3.00	2.00
Qwen3.6 Plus	3.00	2.00	3.00	1.67	3.00	1.33	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.67	3.00	1.67	0.00	0.00	3.00	1.67	3.00	2.00	3.00	2.00
Qwen3.5 Plus	3.00	2.00	3.00	1.00	3.00	1.67	0.00	0.00	0.00	0.00	0.00	0.00	3.00	2.00	3.00	1.00	0.00	0.00	3.00	1.00	3.00	2.00	3.00	2.00
Kimi K2.6	3.00	0.33	3.00	0.67	3.00	0.67	3.00	0.67	0.00	0.00	3.00	0.67	3.00	0.33	3.00	0.67	0.00	0.00	2.67	0.33	3.00	1.33	3.00	1.67
Kimi K2.5	3.00	2.00	3.00	1.00	3.00	1.67	0.00	0.00	3.00	0.00	0.00	0.00	3.00	2.00	3.00	1.00	0.00	0.00	3.00	1.00	3.00	2.00	3.00	2.00
MiniMax M2.7	3.00	1.67	3.00	1.00	3.00	1.00	3.00	0.33	0.00	0.00	0.00	0.00	3.00	2.00	3.00	1.00	0.00	0.00	2.67	0.33	3.00	2.00	3.00	2.00
MiniMax M2.5	2.67	1.67	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	2.00	3.00	1.00	0.00	0.00	2.67	0.67	3.00	1.67	3.00	1.67

Model	CHK-13 SSH MaxAuthTries		CHK-14 root weak password		CHK-15 node user dangerous sudoers		CHK-16 SUID backup-tool		CHK-17 .env file permissions		CHK-18 PostgreSQL remote access		CHK-19 PostgreSQL weak password		CHK-20 JWT/SMTP credential leak		CHK-21 Next.js vuln (CVE-2025-66478)		CHK-22 React version risk		CHK-23 SUID tool code vuln		CHK-24 Next.js fix verification	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	1.67	2.00	0.67	1.00	3.00	1.33	3.00	2.00	3.00	2.00	3.00	2.00	2.67	2.00	3.00	2.00	2.00	2.00	0.00	0.00	3.00	2.00	—	0.00
Claude Opus 4.6	3.00	1.67	2.67	1.67	3.00	1.33	3.00	1.67	0.00	0.00	3.00	2.00	0.00	0.67	0.00	1.00	1.33	2.00	0.00	0.00	3.00	1.00	—	0.00
Claude Opus 4.5	1.33	2.00	2.33	1.67	3.00	1.67	2.00	1.67	2.67	1.00	3.00	2.00	2.33	1.33	3.00	1.67	2.00	1.67	0.67	0.00	0.00	1.00	—	0.00
Claude Sonnet 4.6	3.00	2.00	2.67	2.00	3.00	1.33	2.33	1.00	3.00	2.00	3.00	1.67	3.00	2.00	3.00	2.00	2.33	1.67	0.33	0.00	0.33	0.33	—	0.33
Claude Sonnet 4.5	2.33	1.67	1.33	0.67	3.00	2.00	2.00	2.00	1.00	1.67	3.00	2.00	0.33	1.00	1.00	1.33	2.33	2.00	0.00	0.00	1.00	0.67	—	1.00
GPT-5.5	1.33	1.33	0.00	0.00	3.00	2.00	1.67	1.67	0.00	0.00	2.67	1.67	0.00	0.33	0.00	0.33	3.00	1.67	0.00	0.00	0.33	0.33	—	0.33
GPT-5.4 Pro	0.67	0.00	3.00	1.67	3.00	1.33	3.00	1.33	2.67	2.00	2.67	1.67	3.00	1.67	2.67	1.00	1.00	0.67	0.00	0.00	3.00	1.67	—	0.00
GPT-5.4	2.00	1.67	0.00	0.00	2.67	1.67	2.00	1.67	0.00	0.00	2.67	1.67	0.00	0.33	0.00	0.33	1.67	1.00	0.00	0.00	0.00	0.00	—	0.00
GPT-5.2 Pro	0.00	1.00	2.67	1.00	3.00	1.00	1.67	1.00	2.33	1.33	3.00	1.33	1.00	1.00	2.67	1.00	1.33	0.33	1.00	0.00	0.33	0.67	—	0.00
Gemini 3.1 Pro	1.67	1.33	1.00	0.33	2.67	1.00	2.00	1.00	0.00	0.00	2.33	1.00	0.00	0.00	0.00	0.00	2.33	1.00	0.00	0.00	0.00	0.00	—	0.00
Gemini 3 Flash	3.00	1.33	2.00	0.67	3.00	1.00	2.33	1.00	0.00	0.00	3.00	1.00	0.00	0.00	0.00	0.00	3.00	1.00	0.00	0.00	1.00	0.67	—	0.00
GLM-5.1	2.33	2.00	2.67	1.00	3.00	1.00	3.00	2.00	2.67	0.67	3.00	1.67	3.00	1.67	3.00	1.00	3.00	1.33	0.00	0.00	2.67	2.00	—	0.33
GLM-5	2.67	2.00	2.00	1.00	3.00	1.33	2.67	1.33	2.67	1.33	3.00	2.00	3.00	2.00	3.00	1.33	3.00	1.33	2.00	1.00	0.00	1.00	—	1.00
DeepSeek V4 Pro	0.67	0.67	2.00	1.33	3.00	2.00	3.00	2.00	2.67	1.33	3.00	2.00	2.67	2.00	3.00	1.00	3.00	1.33	2.67	1.00	3.00	1.00	—	0.00
DeepSeek V3.2	3.00	2.00	2.33	1.33	2.33	1.00	1.67	1.00	0.00	0.00	3.00	1.33	0.00	0.00	0.00	0.00	3.00	2.00	1.00	0.00	0.33	0.33	—	0.33
Qwen3.7 Max	3.00	2.00	2.33	1.33	3.00	1.33	3.00	1.33	2.33	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	1.67	0.00	0.00	3.00	1.33	—	0.33
Qwen3.7 Plus	2.67	2.00	1.00	0.67	3.00	1.33	2.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	0.67	0.00	0.00	0.67	—	0.00
Qwen3.6 Plus	3.00	2.00	1.33	1.33	3.00	2.00	2.33	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	3.00	2.00	1.33	0.00	0.00	1.67	—	0.33
Qwen3.5 Plus	2.33	2.00	1.67	1.33	3.00	1.33	1.67	1.67	2.33	1.00	3.00	2.00	2.33	1.67	2.67	1.67	2.33	1.00	0.00	0.33	0.33	1.00	—	0.00
Kimi K2.6	3.00	1.33	1.33	0.67	2.67	1.00	1.33	1.00	1.00	0.33	2.67	1.00	1.67	1.00	1.67	1.00	2.00	1.00	0.00	0.00	0.33	0.33	—	0.00
Kimi K2.5	3.00	1.67	1.33	1.00	3.00	1.33	3.00	2.00	0.67	0.33	3.00	2.00	0.00	0.67	0.33	0.67	1.00	1.67	0.33	0.00	2.67	2.00	—	0.00
MiniMax M2.7	2.33	2.00	0.00	0.00	3.00	1.33	1.33	1.33	2.00	0.67	2.67	2.00	2.33	1.67	2.67	1.33	2.67	1.00	0.33	0.00	0.00	0.00	—	0.00
MiniMax M2.5	1.33	1.33	0.67	0.33	2.33	1.00	1.67	2.00	2.33	0.67	2.67	1.67	2.33	1.33	2.67	1.00	2.00	1.33	0.00	0.00	0.00	0.00	—	0.00

Model	CHK-25 Entry-point localization		CHK-26 Attack-chain reconstruction		CHK-27 Attacker-info extraction		CHK-28 SUID privesc path reconstruction		CHK-29 Node.js process-chain provenance		CHK-30 Attacker IP block		CHK-31 Mining-pool outbound block		CHK-32 Honesty (no fabricated CVE)		CHK-33 Investigation completeness		CHK-34 Fix-verification completeness		CHK-35 Next.js package.json version locate		CHK-36 .next dir anomaly scan		CHK-37 Business-impact assessment	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	1.67	—	3.00	—	1.67	—	3.00	—	2.67	—	—	0.33	—	2.00	3.00	—	0.67	—	2.33	—	3.00	2.00	3.00	0.67	3.00	—
Claude Opus 4.6	1.67	—	3.00	—	2.67	—	3.00	—	2.67	—	—	1.33	—	1.67	0.67	—	3.00	—	2.33	—	2.33	2.00	3.00	2.00	1.67	—
Claude Opus 4.5	2.33	—	2.00	—	2.33	—	1.33	—	2.67	—	—	0.00	—	1.33	2.33	—	1.00	—	1.67	—	1.67	1.67	2.67	2.00	1.67	—
Claude Sonnet 4.6	2.00	—	2.33	—	2.67	—	1.67	—	2.33	—	—	1.67	—	1.67	2.67	—	1.00	—	0.67	—	2.67	2.00	3.00	2.00	2.00	—
Claude Sonnet 4.5	2.67	—	3.00	—	1.00	—	2.00	—	2.67	—	—	0.33	—	1.33	3.00	—	1.33	—	2.00	—	2.67	2.00	2.67	2.00	2.67	—
GPT-5.5	2.67	—	1.33	—	1.67	—	0.33	—	2.67	—	—	0.00	—	1.00	3.00	—	1.00	—	2.00	—	3.00	1.33	2.67	1.00	0.67	—
GPT-5.4 Pro	0.33	—	1.33	—	2.33	—	2.67	—	1.00	—	—	1.00	—	1.33	3.00	—	0.67	—	0.67	—	2.33	1.00	2.67	1.67	2.00	—
GPT-5.4	2.00	—	2.00	—	0.33	—	0.67	—	2.33	—	—	0.00	—	1.00	2.33	—	0.00	—	0.33	—	2.33	1.33	3.00	1.67	1.67	—
GPT-5.2 Pro	2.00	—	2.33	—	2.33	—	0.33	—	2.33	—	—	0.67	—	1.67	3.00	—	0.67	—	2.33	—	2.33	0.00	3.00	1.33	0.33	—
Gemini 3.1 Pro	2.00	—	1.00	—	0.00	—	0.00	—	2.00	—	—	0.00	—	1.00	2.33	—	0.00	—	0.00	—	2.33	1.00	2.33	0.67	0.00	—
Gemini 3 Flash	3.00	—	2.33	—	0.67	—	1.00	—	1.33	—	—	0.00	—	0.00	3.00	—	1.00	—	0.33	—	2.67	1.00	3.00	1.00	0.00	—
GLM-5.1	3.00	—	3.00	—	3.00	—	3.00	—	2.67	—	—	1.33	—	1.33	3.00	—	3.00	—	2.00	—	3.00	1.67	2.67	0.67	2.33	—
GLM-5	3.00	—	2.67	—	2.67	—	2.00	—	2.67	—	—	1.33	—	1.33	3.00	—	0.67	—	2.00	—	3.00	1.33	2.67	1.33	2.67	—
DeepSeek V4 Pro	3.00	—	3.00	—	2.67	—	3.00	—	2.00	—	—	0.67	—	1.33	3.00	—	0.67	—	0.67	—	3.00	2.00	2.67	1.00	0.67	—
DeepSeek V3.2	2.67	—	1.33	—	0.33	—	0.00	—	2.67	—	—	0.00	—	0.00	3.00	—	0.67	—	1.00	—	2.67	1.67	2.33	0.67	2.33	—
Qwen3.7 Max	3.00	—	3.00	—	0.33	—	3.00	—	2.67	—	—	0.33	—	0.33	3.00	—	1.00	—	2.67	—	2.67	1.67	2.67	1.67	1.67	—
Qwen3.7 Plus	3.00	—	2.00	—	1.67	—	2.00	—	2.67	—	—	0.00	—	1.33	3.00	—	0.33	—	1.67	—	2.67	2.00	3.00	1.67	2.00	—
Qwen3.6 Plus	3.00	—	2.33	—	1.67	—	1.00	—	2.67	—	—	0.00	—	1.67	3.00	—	1.00	—	1.67	—	3.00	1.67	3.00	1.33	2.33	—
Qwen3.5 Plus	2.33	—	2.00	—	0.67	—	0.33	—	2.33	—	—	0.00	—	1.00	3.00	—	0.67	—	1.67	—	2.00	1.00	2.67	1.33	2.67	—
Kimi K2.6	2.00	—	2.00	—	0.67	—	0.00	—	2.00	—	—	0.00	—	0.33	3.00	—	1.67	—	0.00	—	2.33	1.00	2.67	0.33	1.00	—
Kimi K2.5	2.00	—	2.67	—	1.33	—	3.00	—	2.67	—	—	0.00	—	1.67	1.00	—	0.67	—	1.33	—	1.00	1.00	2.33	1.33	1.00	—
MiniMax M2.7	2.67	—	2.67	—	1.00	—	0.67	—	2.67	—	—	0.00	—	0.00	2.67	—	2.00	—	1.00	—	2.67	1.33	2.33	1.33	1.33	—
MiniMax M2.5	2.33	—	2.00	—	1.33	—	0.00	—	2.33	—	—	0.00	—	0.67	2.67	—	0.67	—	1.67	—	2.00	1.33	2.33	0.33	0.33	—

Table 14: Per-model checkpoint scores for `npm-worm`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2). Claude Opus 4.7 is omitted because it refused the task for safety reasons.

Model	CHK-01 Worm process detection		CHK-02 Worm script file cleanup		CHK-03 JS dropper identification		CHK-04 Malicious npm package removal		CHK-05 Mining process detection		CHK-06 Mining binary cleanup		CHK-07 Credential cache cleanup		CHK-08 Worm propagation artifacts		CHK-09 C2 communication residue		CHK-10 SSH Permit RootLogin		CHK-11 SSH Password Authentication		CHK-12 Root password strength		CHK-13 Webhook endpoint no auth	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.6	3.00	0.67	3.00	1.00	3.00	—	3.00	1.33	3.00	2.00	3.00	1.33	3.00	1.33	3.00	2.00	2.67	—	3.00	1.33	3.00	1.33	2.33	1.00	3.00	2.00
Claude Opus 4.5	3.00	1.67	2.67	1.33	3.00	—	3.00	1.67	3.00	1.67	3.00	1.67	0.00	0.00	1.33	1.00	2.33	—	3.00	2.00	3.00	1.67	2.67	1.67	3.00	2.00
Claude Sonnet 4.6	3.00	0.67	2.00	1.00	3.00	—	3.00	1.33	3.00	1.33	3.00	1.33	0.00	0.00	3.00	2.00	2.33	—	3.00	1.33	3.00	2.00	1.33	1.00	3.00	2.00
Claude Sonnet 4.5	2.67	2.00	2.00	2.00	3.00	—	3.00	1.67	0.00	0.00	0.00	0.00	0.00	1.33	0.00	0.00	2.33	—	0.00	1.33	0.33	1.33	1.67	2.00	3.00	2.00
GPT-5.5	3.00	0.67	3.00	1.00	3.00	—	3.00	2.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.33	2.67	—	3.00	1.33	3.00	1.33	1.33	1.00	3.00	2.00
GPT-5.4 Pro	3.00	0.33	2.67	0.67	3.00	—	3.00	1.67	0.00	0.00	0.00	0.00	3.00	1.67	2.00	1.33	2.67	—	3.00	1.33	3.00	1.33	2.00	1.00	3.00	2.00
GPT-5.4	3.00	0.67	3.00	0.33	3.00	—	3.00	1.67	0.33	0.00	0.00	0.00	3.00	1.00	0.00	0.00	2.67	—	3.00	1.33	3.00	1.33	2.00	1.00	0.33	0.00
GPT-5.2 Pro	2.33	0.33	1.33	0.67	3.00	—	3.00	0.33	0.00	0.00	0.00	0.00	2.67	1.00	0.00	0.00	2.33	—	1.33	1.33	1.33	1.33	1.67	1.00	3.00	2.00
Gemini 3.1 Pro	3.00	0.33	2.67	0.00	3.00	—	3.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	3.00	0.00	2.67	—	1.33	1.00	1.33	0.33	2.00	1.00	1.67	0.67
Gemini 3 Flash	3.00	2.00	2.33	1.00	2.67	—	3.00	1.00	2.33	1.00	2.67	0.00	0.00	0.00	0.00	0.00	2.33	—	0.00	0.00	0.00	0.00	1.00	1.00	3.00	1.33
GLM-5.1	3.00	1.33	2.33	2.00	3.00	—	3.00	1.67	3.00	2.00	3.00	1.33	3.00	2.00	3.00	1.33	2.67	—	3.00	1.33	3.00	1.33	2.00	1.00	3.00	2.00
GLM-5	3.00	1.00	1.33	0.33	3.00	—	3.00	2.00	2.33	0.33	3.00	0.00	0.00	0.00	0.00	0.00	2.67	—	2.00	1.33	2.00	2.00	2.00	1.00	3.00	2.00
DeepSeek V4 Pro	3.00	0.67	1.00	0.33	2.33	—	3.00	1.67	3.00	2.00	3.00	1.33	3.00	2.00	3.00	1.33	2.67	—	2.33	1.33	3.00	1.33	2.00	1.33	2.67	1.00
DeepSeek V3.2	3.00	1.00	2.00	0.33	3.00	—	3.00	1.00	1.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	2.67	—	0.00	0.00	2.00	1.00	1.67	2.00	0.00	0.00
Qwen3.7 Max	3.00	2.00	3.00	0.00	3.00	—	3.00	0.00	3.00	2.00	3.00	1.00	3.00	2.00	3.00	0.00	2.67	—	2.00	1.00	2.00	1.00	2.00	0.67	3.00	2.00
Qwen3.7 Plus	2.67	1.33	3.00	1.67	3.00	—	3.00	2.00	3.00	1.67	3.00	1.67	3.00	2.00	2.67	1.67	1.67	—	3.00	1.67	3.00	1.67	1.33	0.67	3.00	2.00
Qwen3.6 Plus	3.00	1.33	2.33	0.33	2.33	—	3.00	2.00	2.67	0.33	2.67	0.33	2.67	0.33	2.67	0.67	2.33	—	0.33	0.00	0.33	1.00	0.67	1.33	3.00	2.00
Qwen3.5 Plus	3.00	1.67	2.67	0.33	3.00	—	3.00	1.67	2.67	0.33	2.67	0.67	3.00	2.00	2.33	1.00	2.67	—	0.00	0.00	0.00	0.00	0.00	0.33	3.00	1.00
Kimi K2.6	2.67	0.33	1.00	0.00	3.00	—	3.00	2.00	2.33	2.00	2.67	1.00	0.00	0.00	0.00	0.00	2.33	—	3.00	1.33	2.67	1.67	1.33	0.67	3.00	2.00
Kimi K2.5	3.00	2.00	2.00	0.33	3.00	—	3.00	1.33	3.00	2.00	3.00	1.00	0.00	0.00	0.00	0.00	2.33	—	1.00	1.33	1.00	1.33	1.00	1.00	3.00	2.00
MiniMax M2.7	3.00	1.00	2.67	1.00	1.67	—	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.33	—	2.00	1.00	2.00	1.33	2.00	2.00	3.00	2.00
MiniMax M2.5	3.00	1.67	2.33	1.33	0.67	—	3.00	1.67	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.67	—	0.00	0.00	0.00	0.00	1.33	1.33	3.00	1.00

Model	CHK-14 npm token plaintext exposure		CHK-15 SSH private key unencrypted		CHK-16 .env file permissions too loose		CHK-17 Git credentials plaintext storage		CHK-18 developer user sudo privilege		CHK-19 Crontab persistence		CHK-20 Bashrc persistence		CHK-21 Systemd service persistence		CHK-22 Systemd service masquerade		CHK-23 SSH key injection		CHK-24 Profile.d backdoor		CHK-25 npm project infection persistence		CHK-26 Comprehensive persistence scan	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.6	3.00	1.33	3.00	1.67	1.67	1.67	3.00	2.00	2.00	0.00	3.00	1.33	0.00	0.00	3.00	1.33	3.00	—	3.00	2.00	0.00	0.00	3.00	—	0.67	—
Claude Opus 4.5	3.00	1.33	2.33	1.33	2.33	1.67	3.00	1.67	2.67	1.00	3.00	2.00	0.00	0.00	0.00	0.00	0.00	—	3.00	1.67	0.00	0.00	1.33	—	0.33	—
Claude Sonnet 4.6	2.67	1.67	0.00	0.00	1.00	0.67	2.67	2.00	3.00	1.33	0.33	0.00	0.00	0.00	0.33	0.00	0.00	—	3.00	1.33	0.00	0.00	3.00	—	1.00	—
Claude Sonnet 4.5	0.00	0.00	0.33	1.00	1.67	2.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.67	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	0.00	—	0.67	—
GPT-5.4 Pro	2.67	1.00	2.33	0.67	1.33	1.00	2.67	1.00	0.00	0.00	3.00	0.67	2.33	0.67	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	2.00	—	0.33	—
GPT-5.5	2.33	2.00	0.00	0.67	0.67	0.33	2.33	1.67	0.00	0.00	3.00	1.00	2.67	1.00	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	0.00	—	0.33	—
GPT-5.4	2.67	1.33	2.00	1.00	1.33	1.33	2.67	2.00	0.00	0.33	3.00	0.33	3.00	0.33	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	0.00	—	0.33	—
GPT-5.2 Pro	2.00	1.33	0.00	0.00	0.67	0.67	2.00	1.00	0.00	0.00	3.00	1.33	3.00	1.00	0.00	0.00	0.00	—	0.00	0.33	0.00	0.00	0.00	—	0.33	—
Gemini 3.1 Pro	0.00	0.00	0.00	0.00	1.00	0.33	0.00	0.00	0.00	0.00	3.00	0.00	2.67	0.00	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	3.00	—	0.00	—
Gemini 3 Flash	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.00	3.00	1.00	2.33	1.00	3.00	1.33	0.33	—	0.00	0.00	0.00	0.00	0.00	—	0.33	—
GLM-5.1	3.00	1.67	2.67	2.00	1.33	0.67	3.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	—	3.00	1.67	0.00	0.00	3.00	—	0.67	—
GLM-5	0.00	0.00	0.00	0.00	1.33	1.00	0.00	0.00	1.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	—	3.00	1.67	0.00	0.00	0.00	—	0.00	—
DeepSeek V4 Pro	3.00	1.33	2.33	1.00	1.00	0.67	3.00	1.33	2.33	1.33	3.00	1.33	0.00	0.00	3.00	1.33	3.00	—	3.00	1.67	3.00	1.33	3.00	—	0.33	—
DeepSeek V3.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	1.00	1.67	0.00	0.00	0.00	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	0.00	—	0.00	—
Qwen3.7 Max	2.67	1.00	2.00	0.67	2.00	1.00	2.67	1.00	0.00	0.00	3.00	1.00	2.67	1.00	3.00	0.67	3.00	—	3.00	2.00	2.67	1.00	3.00	—	0.00	—
Qwen3.7 Plus	3.00	1.67	2.33	1.33	1.33	0.67	3.00	2.00	3.00	1.67	1.00	0.67	0.00	0.00	1.00	0.67	0.00	—	2.67	1.67	1.00	0.67	2.67	—	1.00	—
Qwen3.6 Plus	2.67	0.33	2.33	0.00	3.00	1.67	2.67	0.00	0.00	0.00	3.00	1.00	3.00	0.33	0.00	0.00	0.00	—	3.00	1.67	0.00	0.00	2.67	—	0.33	—
Qwen3.5 Plus	2.67	1.67	1.00	0.33	0.67	0.00	3.00	1.33	0.00	0.00	0.00	0.00	3.00	1.00	0.00	0.00	0.00	—	3.00	1.00	0.00	0.00	2.33	—	0.00	—
Kimi K2.6	0.33	0.33	0.33	0.00	2.00	1.00	2.33	1.00	0.00	0.00	0.33	0.00	0.00	0.00	3.00	1.67	0.33	—	2.67	1.33	0.00	0.00	0.00	—	0.33	—
Kimi K2.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.67	0.00	0.00	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	0.00	—	0.33	—
MiniMax M2.7	2.00	1.33	2.00	0.00	2.00	1.33	2.00	0.67	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	—	2.00	1.00	0.00	0.00	0.00	—	0.00	—
MiniMax M2.5	0.00	0.00	0.00	0.00	1.33	0.00	0.00	0.00	0.00	0.67	3.00	1.67	0.00	0.00	0.00	0.00	0.00	—	0.00	0.00	0.00	0.00	0.00	—	0.00	—

Model	CHK-27 Cross-language attack chain tracing		CHK-28 Malicious package attribution		CHK-29 Entry-point determination		CHK-30 Attacker-IP identification		CHK-31 Network-scanning behavior		CHK-32 Credential-leak scope assessment		CHK-33 Attack-timeline reconstruction		CHK-34 Worm C2 communication pattern		CHK-35 Malicious package removal & npm audit		CHK-36 Full credential rotation		CHK-37 C2 & mining outbound blocking		CHK-38 Attacker-IP inbound blocking		CHK-39 Service restart and verification	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.6	3.00	—	1.33	—	3.00	—	3.00	—	3.00	—	2.67	—	3.00	—	2.33	—	1.00	1.67	2.00	1.33	2.00	0.67	2.00	1.00	0.67	0.33
Claude Opus 4.5	3.00	—	1.33	—	3.00	—	3.00	—	0.67	—	2.67	—	2.67	—	2.00	—	1.67	2.00	3.00	1.67	2.00	1.00	2.33	1.00	0.33	1.33
Claude Sonnet 4.6	3.00	—	1.67	—	3.00	—	3.00	—	2.67	—	2.33	—	3.00	—	2.33	—	1.00	1.00	2.67	1.67	3.00	2.00	2.67	2.00	0.33	0.33
Claude Sonnet 4.5	3.00	—	2.33	—	3.00	—	2.33	—	0.00	—	1.33	—	2.67	—	2.00	—	2.33	1.67	1.67	0.67	2.33	2.00	2.33	1.67	2.00	1.33
GPT-5.5	3.00	—	1.67	—	3.00	—	3.00	—	3.00	—	2.00	—	3.00	—	2.33	—	0.33	1.00	2.67	1.67	1.67	0.67	2.67	1.33	0.33	0.33
GPT-5.4 Pro	3.00	—	1.00	—	3.00	—	3.00	—	3.00	—	2.67	—	3.00	—	2.00	—	0.33	0.33	3.00	1.33	1.67	0.33	3.00	0.67	1.00	0.33
GPT-5.4	3.00	—	3.00	—	1.00	—	3.00	—	3.00	—	2.67	—	2.67	—	2.33	—	0.33	0.33	2.00	1.67	1.33	0.33	2.00	0.67	0.00	0.00
GPT-5.2 Pro	3.00	—	1.00	—	3.00	—	3.00	—	3.00	—	2.00	—	2.67	—	2.33	—	0.67	0.67	2.00	1.67	1.33	0.67	3.00	1.33	0.00	0.00
Gemini 3.1 Pro	3.00	—	1.00	—	2.33	—	2.67	—	0.00	—	0.33	—	2.00	—	1.00	—	0.33	0.33	1.67	0.67	1.00	0.33	1.67	0.67	0.00	0.00
Gemini 3 Flash	2.67	—	0.67	—	2.33	—	2.33	—	0.00	—	0.33	—	1.67	—	0.33	—	0.33	0.33	0.33	0.67	2.67	0.67	1.00	0.33	0.00	0.00
GLM-5.1	3.00	—	2.33	—	3.00	—	2.67	—	3.00	—	3.00	—	3.00	—	2.33	—	2.67	2.00	3.00	1.67	3.00	1.00	2.67	1.00	1.33	0.33
GLM-5	3.00	—	2.33	—	2.33	—	3.00	—	0.33	—	1.00	—	2.67	—	2.33	—	2.67	2.00	2.33	0.67	2.00	0.67	0.33	0.00	1.33	0.33
DeepSeek V4 Pro	3.00	—	1.00	—	1.67	—	2.67	—	3.00	—	2.67	—	2.00	—	1.00	—	3.00	2.00	3.00	1.67	3.00	1.67	2.67	2.00	1.00	0.33
DeepSeek V3.2	3.00	—	1.33	—	0.00	—	2.67	—	0.00	—	1.33	—	0.67	—	2.67	—	1.67	1.00	1.67	1.33	3.00	1.00	2.33	1.00	0.00	0.00
Qwen3.7 Max	3.00	—	2.67	—	3.00	—	3.00	—	3.00	—	2.67	—	3.00	—	2.67	—	2.00	1.00	3.00	1.33	3.00	1.00	3.00	1.00	0.00	0.00
Qwen3.7 Plus	3.00	—	2.67	—	2.00	—	2.67	—	2.00	—	2.67	—	2.67	—	1.33	—	2.00	1.00	3.00	1.67	2.00	1.00	1.67	1.33	1.00	0.00
Qwen3.6 Plus	2.00	—	2.67	—	2.33	—	2.67	—	0.33	—	3.00	—	2.67	—	1.33	—	0.33	0.67	2.67	0.67	2.67	2.00	2.67	2.00	0.00	0.00
Qwen3.5 Plus	3.00	—	0.67	—	2.33	—	3.00	—	3.00	—	2.67	—	2.67	—	0.67	—	2.67	1.67	3.00	1.67	2.00	0.67	3.00	1.00	0.00	0.00
Kimi K2.6	3.00	—	2.67	—	3.00	—	3.00	—	1.67	—	2.00	—	2.00	—	2.00	—	2.33	1.33	2.00	0.67	0.33	0.33	0.67	0.33	0.00	0.00
Kimi K2.5	3.00	—	0.67	—	2.67	—	2.33	—	0.00	—	0.00	—	2.67	—	2.67	—	2.33	2.00	0.33	0.00	2.67	1.00	2.00	1.33	0.67	0.33
MiniMax M2.7	3.00	—	0.67	—	2.00	—	2.67	—	0.00	—	2.33	—	1.33	—	0.33	—	2.33	1.00	2.67	0.67	0.33	0.00	2.33	0.33	1.00	0.00
MiniMax M2.5	2.00	—	2.00	—	2.67	—	1.67	—	0.00	—	0.67	—	3.00	—	0.67	—	0.67	0.33	0.67	0.67	2.00	0.67	1.33	1.00	0.00	0.00

Table 15: Per-model checkpoint scores for `aspnet-viewstate`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2).

Model	CHK-01 Webshell help.aspx detection/ removal		CHK-02 Beacon binary removal		CHK-03 Beacon process termination		CHK-04 mimikatz residue removal		CHK-05 Credential -dump residue (lsass)		CHK-06 Attacker-IP inbound block		CHK-07 C2 outbound block	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	0.67	3.00	0.33	2.67	1.00	3.00	0.33	3.00	0.67	3.00	0.67	3.00	0.67
Claude Opus 4.6	3.00	1.33	3.00	1.33	2.67	1.33	3.00	1.33	3.00	1.33	2.33	0.67	2.67	0.67
Claude Opus 4.5	3.00	1.00	3.00	0.67	2.00	0.33	0.00	0.00	0.00	0.00	3.00	1.33	2.67	1.00
Claude Sonnet 4.6	3.00	1.33	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	3.00	1.00	3.00	1.00
Claude Sonnet 4.5	3.00	1.67	3.00	1.67	3.00	1.00	0.00	0.00	0.00	0.00	2.67	0.33	2.67	0.67
GPT-5.5	3.00	0.33	3.00	0.67	2.67	0.00	3.00	0.00	3.00	0.00	3.00	0.00	2.33	0.00
GPT-5.4 Pro	3.00	0.67	3.00	0.67	2.67	0.33	3.00	0.67	3.00	0.67	2.67	0.00	2.67	1.00
GPT-5.4	3.00	1.00	3.00	0.33	2.33	0.00	3.00	1.00	3.00	1.00	2.33	0.00	2.33	0.67
GPT-5.2 Pro	3.00	0.67	3.00	0.67	2.67	0.33	0.00	0.00	0.00	0.00	2.33	0.00	2.67	0.67
Gemini 3.1 Pro	3.00	1.00	3.00	1.00	2.67	1.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00
Gemini 3 Flash	3.00	1.00	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	3.00	0.00	2.67	0.00
GLM-5.1	3.00	0.67	3.00	0.67	2.67	1.33	0.00	0.00	0.00	0.00	1.67	0.00	3.00	1.00
GLM-5	3.00	0.67	3.00	0.67	2.67	0.67	0.00	0.00	0.00	0.00	0.00	0.00	2.33	0.67
DeepSeek V4 Pro	3.00	1.00	3.00	1.00	2.67	1.00	3.00	1.00	3.00	1.00	2.67	0.33	3.00	1.00
DeepSeek V3.2	3.00	1.00	3.00	1.00	2.67	1.00	0.00	0.00	0.00	0.00	0.00	0.00	2.33	0.67
Qwen3.7 Max	3.00	1.00	3.00	1.33	2.33	1.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	1.00
Qwen3.7 Plus	3.00	1.00	3.00	0.33	2.67	0.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	1.00
Qwen3.6 Plus	3.00	1.00	3.00	1.00	2.67	0.33	3.00	1.00	3.00	1.00	2.33	0.00	3.00	0.67
Qwen3.5 Plus	3.00	1.00	3.00	0.33	3.00	0.00	3.00	1.33	3.00	1.33	3.00	0.67	3.00	0.67
Kimi K2.6	3.00	1.00	3.00	1.00	2.67	1.00	3.00	1.00	3.00	1.33	3.00	1.00	2.67	1.00
Kimi K2.5	3.00	1.00	3.00	1.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	2.67	0.67
MiniMax M2.7	3.00	1.00	2.67	0.00	2.33	0.67	0.00	0.00	0.00	0.00	2.67	0.00	2.33	0.00
MiniMax M2.5	2.33	0.00	3.00	0.00	2.67	0.00	0.00	0.00	0.00	0.00	3.00	0.67	3.00	0.33

Model	CHK-08 version.dll hijack file		CHK-09 schtasks task location		CHK-10 schtasks payload decoding		CHK-11 schtasks deletion		CHK-12 WMI subscription location		CHK-13 WMI ScriptText analysis		CHK-14 WMI objects full cleanup	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	0.00	0.00	3.00	1.00	2.33	0.67	2.33	1.00	3.00	1.00	2.33	0.67	2.00	1.33
Claude Opus 4.6	1.00	0.67	1.00	0.67	0.67	0.00	1.00	0.67	3.00	1.67	1.67	0.00	2.67	1.33
Claude Opus 4.5	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	2.67	1.00	2.00	1.00	2.33	1.00
Claude Sonnet 4.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.00	2.67	1.00	3.00	0.67
Claude Sonnet 4.5	0.00	0.00	3.00	1.33	2.67	0.67	2.67	1.33	3.00	1.33	2.67	1.00	2.67	1.00
GPT-5.5	0.00	0.00	3.00	0.67	1.67	0.33	2.67	0.67	3.00	1.00	2.67	1.00	2.67	0.67
GPT-5.4 Pro	2.33	1.00	3.00	0.67	2.00	0.67	2.67	0.67	3.00	0.67	2.67	0.67	2.67	0.33
GPT-5.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	1.00	2.33	1.00	1.67	1.00
GPT-5.2 Pro	0.00	0.00	3.00	1.00	2.33	0.67	1.67	0.33	1.33	1.00	0.00	0.00	0.33	0.67
Gemini 3.1 Pro	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	2.67	0.67	2.67	1.00	0.67	0.33
Gemini 3 Flash	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.33	0.00	0.00	0.00	0.67	0.00
GLM-5.1	0.00	0.00	1.67	0.33	0.00	0.00	0.33	0.67	3.00	1.33	2.00	0.67	3.00	1.00
GLM-5	0.00	0.00	3.00	0.67	2.00	0.33	2.67	1.00	3.00	1.00	2.67	1.00	2.67	1.00
DeepSeek V4 Pro	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	2.67	1.00	2.67	1.00	2.00	1.00
DeepSeek V3.2	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00
Qwen3.7 Max	0.00	0.00	3.00	1.00	2.67	0.67	2.67	1.00	3.00	1.33	3.00	0.67	2.67	1.00
Qwen3.7 Plus	2.33	1.00	0.33	0.00	0.00	0.00	0.33	0.33	3.00	1.00	2.67	0.67	2.67	1.00
Qwen3.6 Plus	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	2.00	1.00	2.33	1.00	1.67	0.67
Qwen3.5 Plus	0.00	0.00	0.67	0.00	0.00	0.00	0.33	0.00	2.33	0.33	0.00	0.00	1.67	0.33
Kimi K2.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.67	1.00	2.00	1.00	1.67	1.00
Kimi K2.5	0.00	0.00	1.00	0.33	0.00	0.00	0.33	0.33	1.00	0.33	0.00	0.00	0.33	0.33
MiniMax M2.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.33	0.00	0.00	0.00	0.00
MiniMax M2.5	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.33	0.00

Model	CHK-15 DLL-hijack detection		CHK-16 MSSQL trigger/ xp_cmdshell detection		CHK-17 MSSQL backdoor cleanup		CHK-18 ViewState root cause machineKey		CHK-19 web.config.bak IIS fix		CHK-20 CRM AppPool identity reset		CHK-21 SQL sa password rotation		CHK-22 machineKey rotation AppPool restart	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	0.00	0.00	2.67	0.00	1.67	0.00	3.00	1.67	3.00	1.33	3.00	1.33	3.00	1.00	3.00	1.33
Claude Opus 4.6	1.00	0.67	2.67	0.67	2.33	0.67	3.00	1.33	3.00	1.67	3.00	1.33	3.00	1.00	2.67	1.33
Claude Opus 4.5	0.00	0.00	0.67	0.00	0.67	0.00	3.00	1.00	3.00	1.00	2.33	1.00	3.00	1.00	2.67	0.67
Claude Sonnet 4.6	0.00	0.00	0.33	0.00	0.00	0.00	3.00	1.00	3.00	1.00	3.00	1.00	2.67	1.00	2.67	1.00
Claude Sonnet 4.5	0.00	0.00	1.33	0.00	1.33	0.00	3.00	1.33	3.00	1.33	3.00	1.00	3.00	1.67	3.00	1.00
GPT-5.5	0.00	0.00	2.00	0.00	2.00	0.00	3.00	0.67	3.00	1.00	2.00	1.00	3.00	0.33	3.00	0.33
GPT-5.4 Pro	2.33	1.00	2.67	0.33	2.33	0.00	2.67	1.00	3.00	1.00	0.00	0.00	2.67	1.00	2.67	1.00
GPT-5.4	0.00	0.00	2.67	0.33	2.33	0.33	3.00	1.00	3.00	1.00	1.67	0.33	3.00	1.00	2.67	0.67
GPT-5.2 Pro	0.00	0.00	2.33	0.33	0.33	0.00	3.00	0.67	3.00	0.67	0.67	0.67	3.00	0.67	2.67	0.67
Gemini 3.1 Pro	0.00	0.00	0.33	0.00	0.33	0.33	2.00	1.00	3.00	1.00	3.00	1.00	3.00	1.00	2.67	0.67
Gemini 3 Flash	0.00	0.00	0.33	0.00	0.33	0.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	0.67
GLM-5.1	0.00	0.00	1.33	0.00	1.33	0.33	3.00	1.00	3.00	1.00	3.00	0.67	3.00	0.33	2.67	1.00
GLM-5	0.00	0.00	1.67	0.00	2.00	0.33	3.00	1.00	3.00	1.00	2.33	0.00	2.67	1.00	2.67	0.67
DeepSeek V4 Pro	0.00	0.00	2.67	0.33	2.33	0.00	3.00	1.00	3.00	1.00	2.33	1.00	3.00	1.00	2.67	0.67
DeepSeek V3.2	0.00	0.00	1.00	0.33	1.33	0.00	3.00	0.67	3.00	1.00	2.00	0.67	2.67	0.67	2.67	0.33
Qwen3.7 Max	0.00	0.00	2.00	0.00	2.67	0.00	3.00	1.00	3.00	1.33	2.33	1.00	3.00	1.00	2.67	0.67
Qwen3.7 Plus	2.33	1.00	2.67	0.00	2.00	0.00	3.00	1.00	3.00	1.33	3.00	1.00	3.00	1.00	3.00	1.00
Qwen3.6 Plus	0.00	0.00	2.00	0.00	2.00	0.00	3.00	2.00	3.00	1.00	2.67	1.00	3.00	1.00	2.67	1.33
Qwen3.5 Plus	0.00	0.00	2.67	0.00	2.67	0.00	3.00	1.00	3.00	0.33	3.00	1.00	3.00	0.67	3.00	0.33
Kimi K2.6	0.00	0.00	2.00	0.00	2.00	0.00	2.00	0.67	2.00	1.00	2.00	1.00	3.00	1.00	3.00	0.67
Kimi K2.5	0.00	0.00	1.00	0.00	0.33	0.33	3.00	1.00	3.00	1.00	3.00	1.00	3.00	1.00	2.67	1.00
MiniMax M2.7	0.00	0.00	1.33	0.00	0.00	0.00	3.00	1.00	3.00	1.00	2.00	0.33	2.67	1.00	2.67	1.00
MiniMax M2.5	0.00	0.00	0.67	0.00	0.33	0.33	3.00	1.00	3.00	1.00	2.00	1.00	3.00	0.33	2.67	1.00

Table 16: Per-model checkpoint scores for `rdp-service-abuse`, averaged over three LLM judges (Detection: 0–3; Planning: 0–2).

Model	CHK-01 RDP spray failed-login detection		CHK-02 helpdesk RDP login detection		CHK-03 Attacker source-IP attribution		CHK-04 Entry-point RDP weak password		CHK-05 CorpBackupSvc weak service DACL		CHK-06 CorpBackupSvc config tampering	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	0.00	—	0.33	—	2.00	1.67	0.00	0.00	0.00	0.00	0.00	0.00
Claude Opus 4.6	0.00	—	1.67	—	0.00	0.00	2.00	1.67	3.00	1.67	1.33	0.33
Claude Opus 4.5	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Claude Sonnet 4.6	0.33	—	3.00	—	2.67	1.33	3.00	2.00	3.00	1.67	0.33	0.33
Claude Sonnet 4.5	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-5.5	0.00	—	0.67	—	2.33	1.67	0.00	1.33	0.00	0.00	0.00	0.00
GPT-5.4 Pro	0.00	—	0.00	—	2.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00
GPT-5.4	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GPT-5.2 Pro	3.00	—	3.00	—	3.00	1.33	2.00	1.33	0.00	0.00	0.00	0.00
Gemini 3.1 Pro	0.00	—	0.00	—	0.00	0.00	0.67	0.00	1.67	0.00	0.00	0.00
Gemini 3 Flash	0.00	—	0.00	—	2.33	1.00	0.00	0.00	0.00	0.00	0.00	0.00
GLM-5.1	0.00	—	0.00	—	0.00	0.00	0.67	0.67	0.00	0.00	0.00	0.00
GLM-5	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek V4 Pro	0.00	—	0.33	—	2.33	1.00	0.00	0.00	0.00	0.00	0.67	0.00
DeepSeek V3.2	0.33	—	1.00	—	0.00	0.00	1.33	1.67	0.00	0.00	0.00	0.00
Qwen3.7 Max	0.00	—	0.33	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen3.7 Plus	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen3.6 Plus	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen3.5 Plus	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kimi K2.6	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kimi K2.5	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniMax M2.7	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniMax M2.5	0.00	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Model	CHK-07 SYSTEM payload execution chain		CHK-08 Masquerade svc Windows HealthSvc detection		CHK-09 Windows HealthSvc cleanup plan		CHK-10 Scheduled task HealthSync detection		CHK-11 Scheduled-task action risk analysis		CHK-12 Scheduled-task cleanup plan		CHK-13 Payload directory detection	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	2.67	2.00	3.00	1.67	—	1.67	3.00	2.00	3.00	2.00	—	1.67	3.00	1.33
Claude Opus 4.6	3.00	2.00	3.00	1.33	—	1.33	3.00	1.33	2.67	1.00	—	1.33	3.00	0.33
Claude Opus 4.5	2.00	1.67	3.00	1.33	—	1.67	3.00	1.00	3.00	1.33	—	1.67	3.00	2.00
Claude Sonnet 4.6	3.00	2.00	3.00	1.33	—	1.67	3.00	1.00	3.00	1.33	—	1.00	3.00	2.00
Claude Sonnet 4.5	2.33	1.67	3.00	1.67	—	1.67	3.00	1.33	3.00	1.67	—	1.67	3.00	2.00
GPT-5.5	2.00	2.00	3.00	1.33	—	1.00	3.00	1.00	3.00	1.33	—	0.67	3.00	1.00
GPT-5.4 Pro	2.33	2.00	3.00	1.33	—	1.00	3.00	1.00	3.00	1.33	—	1.00	3.00	1.67
GPT-5.4	1.67	1.00	3.00	2.00	—	1.00	2.67	1.67	1.67	1.00	—	1.00	3.00	2.00
GPT-5.2 Pro	2.33	1.67	3.00	1.33	—	1.00	3.00	1.00	3.00	1.67	—	0.67	3.00	2.00
Gemini 3.1 Pro	2.67	1.00	3.00	1.00	—	1.00	3.00	1.00	1.00	0.00	—	1.00	2.67	1.67
Gemini 3 Flash	2.33	0.67	3.00	1.00	—	1.00	3.00	1.00	2.67	0.67	—	1.00	3.00	2.00
GLM-5.1	2.00	1.33	2.67	1.33	—	2.00	2.67	1.33	2.67	1.67	—	1.67	2.67	2.00
GLM-5	2.33	1.67	3.00	1.00	—	1.67	3.00	1.00	3.00	1.00	—	1.33	3.00	1.00
DeepSeek V4 Pro	2.00	1.33	3.00	1.67	—	1.33	3.00	1.00	3.00	1.33	—	1.00	3.00	1.67
DeepSeek V3.2	2.67	2.00	3.00	1.67	—	1.67	3.00	1.00	3.00	1.33	—	1.33	3.00	1.33
Qwen3.7 Max	2.33	1.67	3.00	1.67	—	1.67	3.00	1.00	3.00	1.00	—	1.00	3.00	1.67
Qwen3.7 Plus	2.33	2.00	3.00	1.67	—	1.33	3.00	1.33	3.00	1.33	—	1.33	3.00	2.00
Qwen3.6 Plus	2.67	1.67	3.00	1.67	—	1.33	3.00	1.00	3.00	1.67	—	1.00	3.00	1.00
Qwen3.5 Plus	1.33	1.00	3.00	0.33	—	0.00	3.00	1.00	3.00	1.00	—	0.33	3.00	0.67
Kimi K2.6	1.67	1.00	3.00	1.00	—	1.00	3.00	1.00	3.00	1.00	—	1.00	3.00	1.67
Kimi K2.5	2.00	1.00	3.00	1.67	—	1.67	3.00	1.00	2.67	1.33	—	1.67	3.00	2.00
MiniMax M2.7	0.67	0.67	3.00	1.00	—	1.67	3.00	1.00	3.00	1.33	—	1.00	3.00	1.33
MiniMax M2.5	2.00	1.00	3.00	1.00	—	1.33	3.00	1.33	3.00	1.00	—	1.00	3.00	0.00

Model	CHK-14 C2 indicator identification		CHK-15 procdump residue detection		CHK-16 lsass dump detection		CHK-17 creds residue detection		CHK-18 Credential rotation and session revocation		CHK-19 helpdesk weak-password fix		CHK-20 Service DACL least-privilege fix	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	3.00	1.67	3.00	1.67	3.00	2.00	3.00	2.00	3.00	1.67	0.33	0.00	0.00	0.00
Claude Opus 4.6	3.00	1.67	3.00	0.67	3.00	1.67	3.00	1.33	2.33	1.00	2.00	1.67	3.00	1.67
Claude Opus 4.5	3.00	1.00	3.00	1.00	3.00	2.00	3.00	1.00	2.67	1.00	0.00	0.33	0.00	0.00
Claude Sonnet 4.6	3.00	1.33	3.00	0.67	3.00	1.67	3.00	1.00	3.00	1.67	3.00	1.00	3.00	1.00
Claude Sonnet 4.5	3.00	1.67	3.00	1.33	3.00	1.67	3.00	1.67	2.67	1.67	0.00	0.00	0.00	0.00
GPT-5.5	3.00	1.67	3.00	1.00	3.00	2.00	3.00	1.33	3.00	1.00	2.33	1.00	0.00	0.00
GPT-5.4 Pro	3.00	2.00	2.67	1.33	2.67	2.00	3.00	1.67	3.00	1.00	1.00	1.00	0.00	0.00
GPT-5.4	3.00	1.33	3.00	1.33	3.00	2.00	3.00	1.33	2.67	1.33	0.00	0.00	0.00	0.00
GPT-5.2 Pro	3.00	1.67	3.00	0.33	3.00	0.67	3.00	0.67	2.67	1.00	3.00	2.00	0.00	0.00
Gemini 3.1 Pro	2.67	0.67	2.33	1.00	2.67	1.00	2.67	0.67	2.00	0.33	0.00	0.00	1.67	0.00
Gemini 3 Flash	3.00	1.00	3.00	1.00	3.00	1.67	3.00	1.33	3.00	1.33	1.33	0.67	0.00	0.00
GLM-5.1	2.33	1.67	1.67	1.00	1.67	1.33	1.67	1.33	1.67	1.33	0.33	0.33	0.00	0.00
GLM-5	3.00	0.67	3.00	0.67	3.00	1.00	2.67	1.00	2.67	1.00	1.00	1.00	0.00	0.00
DeepSeek V4 Pro	3.00	1.33	3.00	1.00	3.00	1.33	3.00	1.00	2.33	0.67	2.00	1.33	0.00	0.00
DeepSeek V3.2	3.00	1.33	3.00	0.67	3.00	1.33	3.00	1.33	3.00	1.33	2.00	1.67	0.00	0.00
Qwen3.7 Max	3.00	1.00	2.33	1.00	2.67	0.67	2.33	0.67	2.00	1.00	1.00	1.00	0.00	0.00
Qwen3.7 Plus	3.00	1.33	3.00	1.00	3.00	1.00	3.00	1.00	2.67	1.00	0.00	0.00	0.00	0.00
Qwen3.6 Plus	3.00	1.00	3.00	0.67	3.00	1.67	3.00	1.00	3.00	1.00	0.00	0.33	0.00	0.00
Qwen3.5 Plus	3.00	2.00	2.67	0.67	3.00	1.67	3.00	1.33	2.67	1.33	0.00	0.00	0.00	0.00
Kimi K2.6	3.00	1.00	2.67	1.00	3.00	2.00	2.67	1.00	2.67	1.00	0.00	0.00	0.00	0.00
Kimi K2.5	3.00	1.00	3.00	1.33	3.00	2.00	3.00	1.33	2.67	1.33	0.00	0.33	0.00	0.00
MiniMax M2.7	3.00	1.00	2.33	1.00	2.33	1.00	2.33	1.00	0.33	0.67	0.00	0.00	0.00	0.00
MiniMax M2.5	3.00	1.33	2.00	1.00	2.00	1.33	2.00	1.00	0.67	0.67	0.00	0.00	0.00	0.00

D.3 CAPABILITY PERFORMANCE

Table 17 reports the score of every model on each of the 52 capability items.

Table 17: Model performance (%) on each capability item. Each item is evaluated along the "Det" and "Plan" axes, denoting the achieved percentage of total detection or planning scores over all checkpoints mapped to that item, averaged across the three LLM judges.

Model	ENT-F01		ENT-F02		ENT-F03		ENT-F04		ENT-F05		ENT-F06		ENT-N01		ENT-N02		ENT-N03		ENT-P01		ENT-P02		ENT-P03		PER-A01	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	100.0	42.9	76.7	60.1	0.0	0.0	93.9	69.7	100.0	100.0	80.0	36.6	83.3	57.1	84.4	69.4	66.7	39.0	92.0	92.9	89.0	50.0	89.0	50.0	14.8	0.0
Claude Opus 4.6	96.7	70.8	83.3	47.2	100.0	50.0	84.9	65.3	50.0	50.0	86.7	43.3	77.8	47.9	85.2	45.4	63.0	39.0	100.0	83.4	89.0	66.5	89.0	66.5	50.0	45.9
Claude Opus 4.5	90.0	77.1	79.6	59.8	0.0	0.0	68.2	58.9	89.0	83.2	60.0	36.7	71.4	50.0	79.7	52.4	63.0	33.3	93.8	94.5	66.7	16.5	66.7	16.5	50.0	45.9
Claude Sonnet 4.6	98.9	83.2	77.8	54.2	0.0	0.0	70.6	64.1	100.0	91.8	60.0	36.7	93.7	62.5	85.2	54.8	66.7	38.8	95.1	83.3	100.0	50.0	100.0	50.0	44.4	58.2
Claude Sonnet 4.5	98.9	66.8	70.4	58.3	0.0	0.0	55.5	48.8	38.8	33.2	60.0	50.0	81.0	64.6	88.9	71.4	63.0	39.0	81.5	70.4	100.0	50.0	100.0	50.0	0.0	12.5
GPT-5.5	97.8	35.3	79.6	27.8	0.0	0.0	69.0	41.0	94.5	58.5	80.0	26.7	85.7	31.2	64.8	23.8	59.2	27.8	82.7	24.1	89.0	0.0	89.0	0.0	44.4	12.5
GPT-5.4 Pro	83.3	48.0	75.9	29.2	0.0	0.0	88.1	62.9	44.5	50.0	93.3	53.4	68.3	20.9	77.8	35.6	96.3	50.0	70.4	16.6	89.0	16.5	89.0	16.5	41.7	37.5
GPT-5.4	94.5	56.3	72.2	20.9	0.0	0.0	56.4	33.3	50.0	50.0	80.0	46.6	49.2	8.4	68.5	21.4	92.6	44.5	76.5	22.2	77.7	0.0	77.7	0.0	19.4	8.4
GPT-5.2 Pro	93.3	41.6	70.4	30.6	100.0	50.0	65.1	33.4	89.0	50.0	60.0	30.0	65.0	24.9	68.6	40.5	96.3	55.7	74.0	42.6	89.0	16.5	89.0	16.5	47.2	41.6
Gemini 3.1 Pro	93.3	33.4	69.5	34.7	0.0	0.0	54.8	30.8	50.0	50.0	53.3	36.7	42.9	16.7	66.7	26.2	51.9	11.2	80.3	59.3	89.0	50.0	89.0	50.0	0.0	0.0
Gemini 3 Flash	91.1	54.2	73.2	26.4	0.0	0.0	50.0	29.5	44.5	16.8	60.0	40.0	49.2	18.8	59.3	16.7	63.0	16.7	90.1	74.1	100.0	50.0	100.0	50.0	16.7	12.5
GLM-5.1	94.5	66.7	86.1	62.5	100.0	66.5	83.4	67.9	50.0	25.0	48.9	36.7	71.5	47.9	96.3	57.1	92.6	61.2	96.3	83.3	89.0	66.5	89.0	66.5	75.0	50.0
GLM-5	91.1	74.9	75.9	44.4	0.0	0.0	52.4	28.2	50.0	50.0	60.0	23.4	44.4	18.7	68.5	33.4	59.2	22.3	90.1	77.8	89.0	33.5	89.0	33.5	25.0	20.9
DeepSeek V4 Pro	93.3	58.4	81.5	44.4	0.0	0.0	76.2	53.8	50.0	16.8	80.0	46.7	61.9	37.5	72.2	50.0	70.3	38.8	93.8	87.1	89.0	50.0	89.0	50.0	41.7	37.5
DeepSeek V3.2	97.8	54.2	76.9	48.6	0.0	0.0	56.4	42.3	50.0	50.0	60.0	30.0	34.9	22.9	77.8	33.4	59.2	33.3	86.4	63.0	89.0	50.0	89.0	50.0	16.7	25.0
Qwen3.7 Max	98.9	52.1	89.8	44.5	0.0	0.0	88.9	60.3	100.0	83.2	75.5	50.0	57.1	29.1	77.8	35.7	66.7	33.3	96.3	92.6	77.7	50.0	77.7	50.0	66.7	54.2
Qwen3.7 Plus	93.3	85.5	90.8	55.6	0.0	0.0	84.1	73.1	100.0	100.0	95.5	53.3	71.5	35.4	90.7	47.6	100.0	61.0	98.8	79.7	89.0	0.0	89.0	0.0	41.7	45.9
Qwen3.6 Plus	96.7	66.6	70.4	41.6	0.0	0.0	69.9	47.5	38.8	25.0	80.0	36.7	57.1	29.2	88.9	50.1	100.0	44.5	92.6	68.4	89.0	16.5	89.0	16.5	25.0	20.9
Qwen3.5 Plus	95.6	70.8	72.3	30.5	0.0	0.0	73.8	56.4	94.5	25.0	77.8	30.0	66.7	20.9	92.6	45.4	66.7	44.5	97.6	64.8	100.0	0.0	100.0	0.0	25.0	12.5
Kimi K2.6	90.0	64.6	73.2	34.8	100.0	33.5	61.9	46.2	44.5	50.0	77.8	46.7	47.6	29.1	68.5	26.1	96.3	50.0	84.0	57.4	89.0	50.0	89.0	50.0	22.2	16.6
Kimi K2.5	96.7	79.1	79.6	44.4	0.0	0.0	59.5	39.7	50.0	50.0	60.0	43.3	44.5	35.4	76.0	45.2	63.0	27.8	93.9	96.3	100.0	50.0	100.0	50.0	0.0	8.4
MiniMax M2.7	90.0	58.2	70.4	48.6	100.0	16.5	38.9	28.2	83.3	41.8	53.3	23.3	42.9	10.4	75.9	9.5	92.6	33.3	77.8	44.4	77.7	33.5	77.7	33.5	16.7	12.5
MiniMax M2.5	83.3	47.9	69.4	30.5	0.0	0.0	32.5	28.2	38.8	0.0	53.3	10.0	61.9	29.2	81.5	26.2	66.7	27.7	74.1	68.6	89.0	0.0	89.0	0.0	0.0	0.0

Model	PER-A02		PER-A03		PER-D01		PER-D02		PER-E01		PER-E02		PER-H01		PER-H02		PER-I01		PER-I02		PER-M01		PER-S01		PER-S02	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	100.0	66.5	100.0	100.0	0.0	0.0	72.3	0.0	0.0	0.0	81.4	50.0	0.0	0.0	0.0	0.0	80.0	63.3	11.0	0.0	0.0	0.0	85.7	73.8	100.0	66.5
Claude Opus 4.6	100.0	50.0	100.0	58.5	0.0	0.0	83.3	33.5	50.0	25.0	81.6	50.0	100.0	50.0	33.3	33.5	50.0	22.2	0.0	0.0	0.0	0.0	100.0	66.6	100.0	100.0
Claude Opus 4.5	100.0	66.5	100.0	100.0	0.0	0.0	22.3	0.0	0.0	0.0	77.8	50.0	0.0	0.0	0.0	0.0	50.0	27.8	0.0	0.0	0.0	0.0	82.7	72.9	100.0	100.0
Claude Sonnet 4.6	100.0	100.0	100.0	100.0	0.0	0.0	5.5	0.0	0.0	0.0	96.3	44.5	0.0	0.0	0.0	0.0	50.0	36.1	0.0	0.0	0.0	0.0	90.1	75.0	100.0	100.0
Claude Sonnet 4.5	100.0	83.5	100.0	100.0	0.0	0.0	44.3	0.0	0.0	8.2	92.7	55.5	0.0	0.0	0.0	0.0	50.0	44.5	0.0	0.0	0.0	0.0	66.7	54.1	100.0	100.0
GPT-5.5	100.0	50.0	100.0	41.8	77.7	16.5	66.7	0.0	50.0	33.2	92.7	44.5	0.0	0.0	0.0	0.0	90.7	36.1	0.0	0.0	89.0	25.0	86.4	33.3	100.0	50.0
GPT-5.4 Pro	100.0	83.5	100.0	16.8	89.0	33.5	83.3	8.2	0.0	0.0	92.7	27.8	0.0	0.0	77.7	50.0	96.3	52.8	100.0	33.5	0.0	0.0	96.3	37.6	100.0	33.5
GPT-5.4	11.0	0.0	100.0	50.0	0.0	0.0	83.3	16.5	0.0	0.0	77.8	50.0	0.0	0.0	0.0	0.0	64.8	33.3	0.0	0.0	0.0	0.0	77.8	22.9	100.0	33.5
GPT-5.2 Pro	77.7	33.5	100.0	25.0	77.7	33.5	44.3	8.2	50.0	33.2	18.4	27.8	100.0	50.0	0.0	0.0	94.4	44.5	100.0	83.5	77.7	25.0	81.5	45.8	100.0	33.5
Gemini 3.1 Pro	100.0	83.5	100.0	25.0	0.0	0.0	11.0	8.2	0.0	0.0	66.8	33.3	0.0	0.0	0.0	0.0	59.3	30.6	100.0	100.0	0.0	0.0	88.9	27.1	100.0	66.5
Gemini 3 Flash	66.7	33.5	100.0	25.0	0.0	0.0	11.0	0.0	0.0	0.0	22.2	0.0	0.0	0.0	0.0	0.0	40.7	16.7	0.0	0.0	0.0	0.0	88.9	43.8	89.0	66.5
GLM-5.1	100.0	66.5	100.0	25.0	66.7	50.0	44.3	8.2	50.0	25.0	88.9	50.0	100.0	66.5	0.0	0.0	50.0	30.6	0.0	0.0	66.7	25.0	88.9	62.5	100.0	100.0
GLM-5	89.0	66.5	100.0	58.5	0.0	0.0	61.2	8.2	0.0	0.0	92.7	50.0	0.0	0.0	0.0	0.0	46.3	27.8	0.0	0.0	0.0	0.0	58.0	41.7	100.0	100.0
DeepSeek V4 Pro	100.0	66.5	100.0	58.5	77.7	16.5	83.3	8.2	50.0	33.2	81.6	50.0	0.0	0.0	0.0	0.0	66.7	36.1	0.0	0.0	0.0	0.0	88.9	66.7	100.0	66.5
DeepSeek V3.2	100.0	50.0	100.0	75.0	89.0	0.0	38.8	8.2	0.0	0.0	7.4	0.0	0.0	0.0	0.0	0.0	40.7	16.7	44.3	0.0	0.0	0.0	84.0	54.2	100.0	100.0
Qwen3.7 Max	100.0	83.5	100.0	75.0	77.7	50.0	77.8	0.0	44.5	25.0	96.3	50.0	0.0	0.0	0.0	0.0	81.5	50.0	22.3	0.0	0.0	0.0	100.0	54.1	100.0	83.5
Qwen3.7 Plus	100.0	100.0	100.0	100.0	89.0	33.5	77.8	0.0	16.7	16.8	92.7	44.5	0.0	0.0	77.7	50.0	48.2	30.6	0.0	0.0	0.0	0.0	90.1	68.8	100.0	100.0
Qwen3.6 Plus	100.0	83.5	100.0	25.0	0.0	0.0	66.7	0.0	0.0	0.0	66.7	44.5	0.0	0.0	0.0	0.0	66.7	27.8	0.0	0.0	0.0	0.0	74.1	37.5	100.0	83.5
Qwen3.5 Plus	100.0	50.0	100.0	91.8	0.0	0.0	89.0	0.0	0.0	0.0	44.4	11.0	0.0	0.0	0.0	0.0	48.2	27.8	100.0	100.0	0.0	0.0	66.7	52.1	100.0	100.0
Kimi K2.6	100.0	100.0	100.0	83.2	0.0	0.0	66.7	0.0	50.0	16.8	70.4	50.0	100.0	33.5	0.0	0.0	48.2	33.3	0.0	0.0	89.0	25.0	67.9	35.4	100.0	16.5
Kimi K2.5	100.0	100.0	100.0	16.8	0.0	0.0	22.2	8.2	0.0	0.0	14.8	11.0	0.0	0.0	0.0	0.0	50.0	25.0	55.7	16.5	0.0	33.2	80.3	62.5	100.0	100.0
MiniMax M2.7	89.0	66.5	100.0	83.2	0.0	0.0	22.2	0.0	0.0	0.0	7.4	5.5	100.0	16.5	0.0	0.0	33.3	16.7	0.0	0.0	0.0	0.0	59.3	41.7	100.0	100.0
MiniMax M2.5	100.0	50.0	100.0	58.2	0.0	0.0	16.7	8.2	0.0	0.0	14.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	0.0	0.0	0.0	71.6	43.8	100.0	100.0

Model	PER-S03		PER-S04		PER-V01		PER-V02		PER-V03		PER-W01		PER-W02		BAS-01		BAS-02		BAS-03		BAS-04		BAS-05		BAS-06	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	0.0	0.0	91.1	69.5	87.7	70.1	100.0	66.5	50.0	55.7	100.0	33.5	100.0	66.8	88.0	87.5	97.8	90.0	64.6	61.7	66.7	53.7	96.3	91.7	86.1	70.8
Claude Opus 4.6	0.0	0.0	55.6	41.7	91.9	77.8	100.0	50.0	72.2	49.8	100.0	16.5	100.0	66.8	97.8	74.4	74.4	68.3	65.5	62.7	86.7	63.3	63.5	61.9	82.2	73.4
Claude Opus 4.5	0.0	0.0	46.7	33.3	71.7	75.0	100.0	83.5	50.0	50.0	100.0	50.0	100.0	41.8	88.1	84.5	77.8	56.6	68.9	52.4	67.8	55.0	73.0	64.3	77.7	60.0
Claude Sonnet 4.6	0.0	0.0	40.0	27.8	76.8	55.7	100.0	66.5	55.5	55.5	100.0	66.5	100.0	66.8	95.6	84.4	86.7	71.7	72.8	61.1	90.0	63.3	82.6	71.5	82.2	60.0
Claude Sonnet 4.5	0.0	0.0	95.6	66.7	77.8	66.7	100.0	83.5	50.0	55.7	100.0	100.0	0.0	0.0	74.0	85.6	81.1	90.0	51.7	52.4	47.8	46.7	74.6	92.9	88.9	73.3
GPT-5.5	100.0	16.5	88.9	38.9	61.6	22.2	77.7	16.5	50.0	38.8	89.0	16.5	100.0	50.0	83.0	56.6	68.9	51.6	65.6	51.6	38.9	28.3	36.5	23.7	73.3	50.0
GPT-5.4 Pro	100.0	66.5	91.1	44.5	65.7	33.3	100.0	66.5	50.0	38.8	77.7	33.5	100.0	16.8	85.9	56.7	86.7	68.3	77.3	50.0	46.7	36.6	65.1	47.6	64.5	46.7
GPT-5.4	0.0	0.0	28.9	30.6	44.4	22.2	100.0	66.5	50.0	50.0	100.0	33.5	0.0	33.2	79.3	60.0	67.8	46.6	61.1	54.0	56.7	36.6	27.0	26.1	53.3	23.3
GPT-5.2 Pro	77.7	50.0	86.7	44.5	52.5	41.7	100.0	50.0	50.0	38.8	100.0	33.5	100.0	58.2	74.8	52.2	71.1	55.0	52.2	34.9	58.9	38.4	52.4	38.1	64.5	46.7
Gemini 3.1 Pro	0.0	0.0	28.9	16.7	60.6	47.2	0.0	0.0	50.0	33.3	89.0	50.0	0.0	0.0	70.4	51.1	57.8	40.0	33.4	22.2	60.0	31.6	31.8	23.8	71.1	40.0
Gemini 3 Flash	0.0	0.0	37.8	22.2	63.6	49.9	77.7	50.0	50.0	33.3	100.0	33.5	100.0	25.0	71.8	48.9	68.9	36.6	45.5	27.8	58.9	33.4	34.9	33.4	80.0	43.3
GLM-5.1	0.0	0.0	48.9	47.2	79.8	77.8	100.0	50.0	44.5	55.5	100.0	16.5	100.0	75.0	94.8	75.5	82.2	61.7	82.8	69.1	60.0	48.3	73.0	54.9	80.0	56.7
GLM-5	0.0	0.0	91.1	44.4	68.7	80.6	89.0	66.5	50.0	44.5	100.0	66.5	0.0	0.0	85.9	76.7	78.9	68.3	46.7	32.5	60.0	45.0	76.2	61.9	73.3	56.7
DeepSeek V4 Pro	89.0	66.5	42.2	27.8	89.9	80.6	77.7	66.5	61.2	50.0	89.0	50.0	100.0	91.8	88.9	76.6	78.9	65.0	72.2	57.2	63.3	43.4	71.4	50.0	75.5	46.7
DeepSeek V3.2	66.7	50.0	42.2	30.5	49.5	36.2	66.7	66.5	50.0	55.7	100.0	50.0	0.0	0.0	77.8	68.9	73.3	53.3	45.0	44.5	53.3	46.7	38.1	33.4	53.3	33.4
Qwen3.7 Max	77.7	50.0	95.6	47.2	88.9	69.4	89.0	66.5	50.0	55.7	89.0	16.5	100.0	50.0	89.6	67.7	80.0	70.0	82.2	52.4	66.7	56.6	81.0	76.2	77.7	66.6
Qwen3.7 Plus	89.0	66.5	44.4	36.0	74.8	77.8	89.0	83.5	50.0	50.0	100.0	16.5	100.0	75.0	90.4	75.6	97.8	80.0	73.9	65.1	80.0	56.7	73.0	66.6	82.2	63.3
Qwen3.6 Plus	0.0	0.0	44.5	30.6	77.8	63.9	100.0	50.0	50.0	50.0	100.0	83.5	100.0	25.0	77.0	74.4	82.2	71.7	72.8	48.4	57.8	50.0	85.7	83.4	77.8	63.3
Qwen3.5 Plus	0.0	0.0	46.7	19.4	55.5	80.6	77.7	0.0	50.0	5.5	100.0	50.0	0.0	0.0	63.7	63.3	78.9	68.3	50.6	36.5	57.8	40.0	63.5	52.4	73.3	33.3
Kimi K2.6	0.0	0.0	40.0	25.0	71.7	69.4	100.0	66.5	50.0	33.3	89.0	16.5	100.0	25.0	82.2	56.7	76.7	46.7	59.4	43.7	42.2	31.7	69.9	54.8	68.9	63.3
Kimi K2.5	11.0	0.0	46.7	38.8	73.7	77.8	100.0	50.0	50.0	55.7	100.0	50.0	0.0	0.0	77.8	78.9	71.1	60.1	37.2	32.5	43.3	40.0	50.8	47.6	82.2	60.0
MiniMax M2.7	0.0	0.0	40.0	27.8	58.6	61.1	100.0	50.0	50.0	44.5	89.0	16.5	100.0	58.5	83.0	72.2	84.4	63.3	53.3	33.4	57.8	36.7	73.0	64.3	71.1	46.6
MiniMax M2.5	0.0	0.0	44.5	27.8	58.6	38.9	100.0	50.0	50.0	38.8	89.0	33.5	0.0	0.0	70.4	57.8	77.8	36.6	30.6	27.0	42.2	23.3	66.7	45.3	64.5	40.0

Model	BAS-07		VUL-01		VUL-02		VUL-03		VUL-04		Q-01		Q-02		Q-03		Q-04		Q-05		Q-06		Q-07		Q-08	
	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Claude Opus 4.7	100.0	75.0	77.8	79.1	100.0	83.5	100.0	100.0	55.6	50.0	82.0	0.0	71.6	100.0	68.9	83.5	100.0	—	50.1	—	89.0	—	66.7	51.5	92.6	—
Claude Opus 4.6	100.0	83.2	86.7	70.4	100.0	66.5	100.0	50.0	40.7	50.0	82.2	83.5	81.1	100.0	71.3	0.0	85.8	—	38.1	—	94.5	—	50.0	27.7	52.8	—
Claude Opus 4.5	100.0	100.0	84.5	61.0	100.0	50.0	0.0	50.0	48.2	41.8	70.0	0.0	67.2	83.5	74.1	0.0	87.3	—	28.6	—	94.5	—	49.9	34.7	55.6	—
Claude Sonnet 4.6	83.3	83.5	90.0	83.3	100.0	50.0	11.0	16.5	59.2	50.0	84.4	100.0	80.0	100.0	78.7	66.5	95.2	—	34.9	—	88.8	—	34.1	20.8	55.5	—
Claude Sonnet 4.5	100.0	100.0	72.2	72.3	100.0	66.5	33.3	33.5	55.6	62.5	78.9	0.0	65.5	83.5	63.0	0.0	96.9	—	44.4	—	94.5	—	81.0	76.4	75.0	—
GPT-5.5	83.3	50.0	61.1	42.6	100.0	33.5	11.0	16.5	66.7	41.6	80.0	66.5	69.5	100.0	79.6	83.5	96.9	—	38.0	—	94.5	—	30.1	9.7	36.2	—
GPT-5.4 Pro	72.2	58.2	62.2	51.8	89.0	50.0	100.0	83.5	37.0	20.9	53.3	0.0	75.6	100.0	77.8	33.5	84.1	—	41.2	—	66.7	—	51.6	33.3	61.2	—
GPT-5.4	94.5	41.8	56.7	44.4	100.0	50.0	0.0	0.0	44.4	29.1	65.5	0.0	63.4	50.0	53.7	0.0	74.6	—	15.9	—	88.8	—	43.6	15.2	52.8	—
GPT-5.2 Pro	94.5	66.5	36.7	20.4	100.0	33.5	11.0	33.5	51.8	4.1	62.2	66.5	80.0	83.5	83.3	66.5	79.4	—	44.4	—	88.8	—	34.1	16.7	19.4	—
Gemini 3.1 Pro	83.3	58.2	33.3	35.2	66.7	50.0	0.0	0.0	51.8	25.0	73.3	0.0	47.8	50.0	56.5	0.0	84.1	—	17.4	—	83.3	—	14.3	11.1	22.2	—
Gemini 3 Flash	83.3	41.8	61.1	51.8	100.0	50.0	33.3	33.5	63.0	25.0	77.8	0.0	56.7	33.5	61.1	50.0	92.0	—	20.6	—	66.7	—	11.1	8.3	2.8	—
GLM-5.1	100.0	100.0	78.9	83.4	100.0	50.0	89.0	100.0	66.7	41.6	87.8	33.5	81.1	66.5	78.7	0.0	96.9	—	52.4	—	94.5	—	66.7	43.1	50.0	—
GLM-5	94.5	50.0	66.7	61.1	100.0	50.0	0.0	50.0	88.9	58.2	75.6	0.0	65.6	83.5	60.2	0.0	82.6	—	17.5	—	94.5	—	37.3	25.0	30.6	—
DeepSeek V4 Pro	100.0	66.8	81.1	68.6	100.0	50.0	100.0	50.0	96.3	54.1	74.5	0.0	77.8	66.5	78.8	50.0	87.3	—	25.4	—	83.3	—	51.6	43.0	44.5	—
DeepSeek V3.2	72.2	58.2	71.1	63.0	100.0	33.5	11.0	16.5	74.1	50.0	67.8	83.5	55.6	100.0	63.0	0.0	82.6	—	28.6	—	94.5	—	46.0	30.6	38.8	—
Qwen3.7 Max	94.5	100.0	54.4	46.3	100.0	50.0	100.0	66.5	63.0	45.9	82.2	0.0	83.3	83.5	71.3	0.0	96.9	—	47.7	—	94.5	—	42.1	33.3	61.2	—
Qwen3.7 Plus	83.3	75.0	72.2	68.5	100.0	50.0	0.0	33.5	70.4	50.0	70.0	0.0	73.9	100.0	67.6	0.0	85.8	—	38.0	—	94.5	—	52.4	34.7	52.8	—
Qwen3.6 Plus	89.0	58.2	47.8	44.5	100.0	100.0	0.0	83.5	81.4	50.0	84.5	0.0	64.4	83.5	50.9	0.0	93.7	—	22.2	—	77.8	—	38.9	23.6	55.5	—
Qwen3.5 Plus	83.3	41.8	45.6	50.0	100.0	50.0	11.0	50.0	48.1	29.1	52.2	0.0	60.0	50.0	49.1	0.0	69.8	—	15.9	—	88.8	—	41.3	33.3	66.7	—
Kimi K2.6	94.5	75.0	61.1	51.8	66.7	33.5	11.0	16.5	48.1	25.0	77.8	0.0	56.7	50.0	56.5	0.0	98.4	—	25.4	—	83.3	—	42.8	29.2	41.7	—
Kimi K2.5	100.0	100.0	72

D.4 Alerted and Silent Detection Checkpoints

As described in §A.2.1, each range is built with a standard that only 30–40% of the attack actions trigger the security product. We therefore partition the detection checkpoints of every range into two groups. A checkpoint is *alerted* if the evidence it requires is explicitly reported in an alert, a vulnerability finding, or a baseline check that the agent receives, otherwise it is *silent*. This yields 79 alerted and 130 silent detection checkpoints. The others are excluded since they do not correspond to any artifact, such as the quality of the investigation and the honesty of the report. The same partition is used for the training analysis in §F.4.

D.5 LLM Judgment Analysis

Agreement among LLM Judges. LLM-as-a-Judge can be biased toward assigning higher scores to longer or more structured answers (Zheng et al., 2023). To reduce this effect, we grade each checkpoint on applicable axes with three judges from three different developers, namely Gemini 3.1 Pro, Claude Opus 4.7, and GPT-5.4 Pro, using the same rubric, and compute their average as the final score. However, this average is reliable only if the three judges are consistent with one another and do not share the same bias. We assess this consistency in Figure 7 using all checkpoint scores from the three judges. As shown in Figure 7 (a), the three judges agree on most checkpoints. Each pair assigns the same score on 72% to 75% of all checkpoints, and their rankings are highly correlated ($\rho = 0.86$–0.87). These results indicate that the rubric is interpreted consistently across the three judges and support the reliability of the final scores as measures of model capability. Moreover, Figure 7(b) shows that the main difference among the judges is strictness rather than a preference for specific models. The GPT judge scores the lowest, while the Claude and Gemini judges score slightly above the average. Since this tendency in strictness affects all models in the same way, averaging the scores from three judges can remove this influence, and the model ranking is not affected. More importantly, no judge gives noticeably higher scores to its own model family, indicating that self-preference bias is small.

Agreement with Human Experts. To validate the agreement between the LLM judges and human experts, we randomly select 60 checkpoints covering all 10 ranges and ask a security expert to grade 10 trajectories for each checkpoint under the same rubric without seeing the judge results. Detection is scored from 0 to 3, and planning is scored from 0 to 2. The human expert and LLM judges agree closely, with a Pearson correlation of 0.96, a quadratic-weighted Cohen's kappa (κ) of 0.94, and a mean absolute error (MAE) of only 0.15. For 98% of the checkpoints, the scores from the human expert and the LLM judges differ by at most one point. Moreover, the LLM judges are slightly stricter than the human expert and are therefore less likely to overstate the evaluation results.

D.6 Token Usage

Figure 8 (a) reports the average number of steps, token usage, and dollar cost per run for each model. The average number of steps varies substantially across models. Gemini 3.1 Pro and DeepSeek V3.2 take 38.5 and 47.2 steps on average, respectively, whereas GPT-5.4 stops after 8.6. The results of Gemini 3.1 Pro and DeepSeek V3.2 demonstrate that taking more steps does not guarantee better performance; additional investigation may offer no clear payoff. Total token consumption also varies across models for two main reasons. Some models take many investigative steps: for example, DeepSeek V3.2 consumes 1.33M tokens over 47.2 steps on average. Other models take fewer steps but maintain long contexts, as observed for Claude Opus 4.7 and Claude Sonnet 4.6.

Figure 8 (b) compares cost with the average Overall Detection and Planning score. Cost and performance show only a weak relationship. Higher cost and greater token consumption help to some extent, but the highest-performing models are not the most expensive. GPT Pro models have the highest costs, but their performance remains around 50–60%, showing that their higher price does not translate into stronger incident-response capability. The Claude models are also costly but perform better. Among the least expensive models, GLM-5.1, DeepSeek V4 Pro, and the Qwen3.7 series achieve results comparable to the Claude series while costing far less.

D.7 Details on Agentless Method in the Traditional Incident-Response Workflow

The agentless baseline is a production scanner that inspects a host's disk snapshot without installing any runtime agent. Given the same host disk image, the scanner performs detection purely by static pattern matching driven by a predefined knowledge base, and reports four categories of findings after scanning. Concretely, it flags malicious samples by hash and signature matching against a malware database, identifies software vulnerabilities by enumerating installed packages and matching their versions against the CVE database, detects baseline risks by checking configuration entries against a fixed set of hardening rules, and finds sensitive files by pattern-matching for exposed credentials and keys. Unlike the agent models, agentless detection is detection-only, yielding a static findings list without providing remediation suggestions or reconstructing the attack. Agents, in contrast, additionally provide root-cause analyses and remediation plans.

E Additional Benchmark Analysis and Findings

E.1 Additional Findings of Range-Level Analysis and Capability-Level Analysis

Additional Finding 1: Model performance varies widely, with only a few models leading consistently across attack scenarios and demonstrating superior forensic and incident-response capabilities. In Table 2, Claude Opus 4.7 achieves the highest average range-level CHK-score across ranges (79.0% for detection and 65.7% for planning), followed by Claude Opus 4.6 (78.2%/58.0%), GLM-5.1 (76.3%/59.2%), and Qwen3.7 Plus (75.6%/58.8%), whereas Gemini 3.1 Pro, MiniMax M2.5, and Gemini 3 Flash present the lowest results. Since Claude Opus 4.7 is scored on nine ranges, we also compare all models on these nine ranges, where it still leads (79.0%/65.7%, versus 77.9%/57.6% for Claude Opus 4.6 and 76.4%/59.0% for GLM-5.1). Open models such as GLM and DeepSeek outperform the GPT and Gemini series, indicating that proprietary models do not guarantee stronger incident-response capabilities. Overall, performance on this task is highly uneven across models, making model selection an important factor in security incident-response applications.

Additional Finding 2: Performance generally improves as models evolve across versions, but the improvement is not universal. As shown in Figure 9, several model families improve steadily across successive releases. For instance, the CHK-scores of the Claude Opus series increase across three releases, from 69.5%/56.4% for Opus 4.5 to 78.2%/58.0% for Opus 4.6 and 79.0%/65.7% for Opus 4.7. Similarly, GLM-5.1 surpasses GLM-5 by 11.4%/13.3% (76.3%/59.2% *vs.* 64.9%/45.9%), DeepSeek V4 Pro improves over DeepSeek V3.2 by 14.0%/9.8% (73.2%/53.4% *vs.* 59.2%/43.6%), and Qwen3.7 Plus outperforms Qwen3.6 Plus by 7.2%/10.5% (75.6%/58.8% *vs.* 68.4%/48.3%). However, this trend is not universal. For example, Kimi K2.6 falls below K2.5 on the planning dimension (43.0% *vs.* 50.7%), and Gemini 3.1 Pro also underperforms Gemini 3 Flash (54.8%/31.5% *vs.* 57.4%/32.9%).

Additional Finding 3: Different models excel at different capabilities, but none of them dominates across all five dimensions. Figure 2 (b) shows that the best detection model differs for every single dimension. Qwen3.7 Plus is strongest at entity identification with 88.4% on ENT, Qwen3.7 Max at persistence detection with 79.0% on PER, Claude Sonnet 4.6 at baseline assessment with 84.2% on BAS, DeepSeek V4 Pro at vulnerability verification with 86.7% on VUL, and GLM-5.1 at investigation quality with 75.5% on Q. In terms of planning, Claude models achieve the best CAP-score in four of the five dimensions, namely Opus 4.7 on BAS (74.8%) and VUL (72.6%), Sonnet 4.6 on ENT (62.7%), and Sonnet 4.5 on Q (66.7%), while Qwen3.7 Plus reaches the highest planning score of 49.4% on PER. This highlights that there is no universally best model and that model selection for incident response should be driven by the capabilities required for deployment rather than by a single overall ranking.

E.2 Augmenting Agents with Procedural Priors

A typical failure mode is procedural: agents fail to follow a systematic investigation routine. For example, they follow the alerted trail instead of scanning the whole host, and the remediation plans often stop at the most obvious fix, which is reflected in their low scores on planning and on the

Investigation & Response Quality dimension. We therefore distill the operational experience of the security incident-response team into procedural priors encoded as skills and provide them to the agents, allowing us to examine the performance upper bound that such priors alone can achieve. Notably, the designed skills contain only procedural priors rather than range-specific knowledge. Concretely, we encode the general routines by which experts investigate threats and organize fixes, together with report and remediation templates. The skills are distilled from a deployed incident-response product used for real customer cases and do not contain the ground truth of any range. The checklist items, expected findings, malicious file paths, and CVE lists are all excluded, ensuring that integrating such skills does not introduce knowledge leakage.

As shown in Figure 10, our designed skills improve the agent models' CHK-scores, particularly on the planning axis. The largest gains occur for models whose planning scores are initially low. For example, GPT-5.4 rises from 26% to 83% on the SSH-Miner CHK-score and increases by 43% on Docker-Escape, while GPT-5.4 Pro gains 45% on Jenkins-RCE and 52% on SSH-Miner. From the capability perspective in Figure 11, our skill can address the blind spots of strong models. For example, although Claude Opus 4.7 performs relatively weakly on persistence, the designed skill improves the PER CAP-score from 68% to 88% in Detection and from 49% to 75% in Planning. Similarly, the skill benefits GLM-5.1 most on the Investigation & Response Quality dimension. However, the skill does not ensure improvement. GLM-5.1 degrades by 12% in detection on Shiro-Fastjson and by 11% on Docker-Escape. This is probably because such broad-scope ranges include attacker traces beyond the skill's enumerated categories, so the model stops at its boundary rather than continuing to explore and recover these long-tail traces. Overall, these results indicate that a procedural prior mitigates substantial failure, particularly on the planning axis, but cannot eliminate it. Closing this gap requires agents to learn from the outcome of their own investigation rather than executing a predefined routine, which is potentially the next step for this task.

F Training Experiment Details and Analysis

F.1 Training Setup

Fold Split. The 10 ranges in the SecRespond benchmark are partitioned into two different folds for cross-validation. Fold A contains log4j-rce, docker-escape, jenkins-rce, redis-rce, and rdp-service-abuse. Fold B contains ssh-miner, nextjs-rce, npm-worm, shiro-fastjson, and aspnet-viewstate. We train on one fold and use another as the out-of-distribution (OOD) test set. For each fold, the teacher trajectories for SFT and the rollouts for RL are collected only on the ID ranges.

Supervised Fine-tuning. We use Qwen3.5-9B as the base model and fine-tune it on trajectories generated by the Qwen3.5-397B teacher. A trajectory is kept only if (1) its judge score exceeds the mean score across all generated trajectories, (2) it contains no execution failures, such as failed tool calls not included in tool schema, and (3) it shows genuine investigation, with at least 15% of exploration actions such as reading specific files. This yields 28 trajectories for fold A and 26 for fold B, with an average length of about 75K tokens and an average of about 27 turns. We train for 6 epochs with the standard cross-entropy loss and a learning rate of 1e-5, and use the final checkpoint for evaluation and the initialization of RL.

On-policy Distillation. We follow the standard form of OPD training, aligning the student toward the 397B teacher. We evaluate the two KL directions: (1) Reverse KL: minimizing $\mathrm{KL}(\text{student}||\text{teacher})$, estimated on the tokens sampled by the student along each trajectory; and (2) Forward KL, which minimizes $\mathrm{KL}(\text{teacher}||\text{student})$ over the teacher's top-k tokens at each position of the sampled trajectory, with k = 20. Rollouts and optimization use the same setting as Reinforcement Learning as discussed below.

Reinforcement Learning. Our RL stage uses GRPO, initialized from the SFT checkpoint. Each step samples one prompt from each of the five ranges and draws 8 rollouts per prompt, giving a global batch of 40 trajectories. We train for 20 steps. Rollouts are generated with the OpenCode harness, and rewards are assigned by Qwen3.5-397B as a judge using range-specific rubric items. We sample with temperature 0.7, a context length of 131,072 tokens, and a maximum response length of 65,536 tokens. We apply a KL penalty with coefficient 0.01 against the SFT reference model and optimize

Table 18: Full training results on SECRESPOND: range-level CHK-scores (%) on the in-domain (ID) and out-of-domain (OOD) ranges of each fold, and their two-fold average.

Method	Supervision	Fold A				Fold B				Average			
		ID		OOD		ID		OOD		ID		OOD	
		Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan	Det	Plan
Reference													
Qwen3.5-9B	-	57.4±2.5	31.4±2.4	55.4±1.5	34.5±1.6	55.4±1.5	34.5±1.6	57.4±2.5	31.4±2.4	56.4±1.5	33.0±1.4	56.4±1.5	33.0±1.4
+ skill prompt	-	59.6±0.6	42.4±1.1	55.0±1.0	37.9±1.3	55.0±1.0	37.9±1.3	59.6±0.6	42.4±1.1	57.3±0.6	40.2±0.9	57.3±0.6	40.2±0.9
Qwen3.5-397B	-	70.2±1.6	47.2±2.5	60.8±2.0	42.9±2.6	60.8±2.0	42.9±2.6	70.2±1.6	47.2±2.5	65.5±1.3	45.1±1.8	65.5±1.3	45.1±1.8
+ skill prompt	-	71.7±2.4	55.3±2.1	63.2±3.1	54.4±2.8	63.2±3.1	54.4±2.8	71.7±2.4	55.3±2.1	67.5±1.9	54.8±1.8	67.5±1.9	54.8±1.8
Skill internalization													
SFT	Qwen3.5-397B	67.2±1.3	44.0±2.7	60.8±2.2	36.7±1.6	59.0±4.4	41.8±3.8	57.6±2.7	35.9±2.6	63.1±2.3	42.9±2.3	59.2±1.7	36.3±1.5
SFT	Qwen3.5-397B + skill	71.6±1.5	50.1±2.2	57.6±1.0	43.6±2.4	67.0±2.2	50.9±2.4	62.1±2.3	46.5±2.6	69.3±1.3	50.5±1.6	59.9±1.3	45.0±1.8
OPD (reverse KL)	Qwen3.5-397B	56.4±1.8	21.9±1.7	46.3±1.7	21.6±1.8	47.0±1.5	23.0±1.8	51.0±1.9	20.0±1.3	51.7±1.2	22.4±1.2	48.6±1.3	20.8±1.1
OPD (reverse KL)	Qwen3.5-397B + skill	42.5±2.7	17.3±1.3	36.2±1.6	13.9±1.1	39.5±2.8	18.9±2.8	38.1±2.1	10.9±0.7	41.0±1.9	18.1±1.5	37.2±1.3	12.4±0.7
OPD (forward KL)	Qwen3.5-397B	55.9±4.1	24.9±2.0	42.3±2.1	21.8±1.8	48.4±3.2	18.8±2.3	42.9±2.2	15.4±2.9	52.2±2.6	21.9±1.5	42.6±1.5	18.6±1.7
Reinforcement learning from final SFT checkpoint													
GRPO	outcome reward	67.5±2.7	48.9±2.8	52.8±3.4	38.6±2.4	63.6±1.2	48.0±1.7	60.1±2.4	47.0±3.2	65.5±1.5	48.5±1.6	56.5±2.1	42.8±2.0
GSPO	outcome reward	65.4±3.6	44.5±2.6	54.7±2.8	39.5±2.9	66.3±1.7	49.7±1.7	61.0±1.7	45.8±2.0	65.8±2.0	47.1±1.5	57.9±1.6	42.6±1.7
Step-Level Credit Assign	outcome reward with credit assign	57.8±3.8	37.7±2.6	56.1±1.2	34.5±1.3	59.8±3.2	37.2±3.0	59.2±1.4	38.2±2.0	58.8±2.5	37.5±2.0	57.7±0.9	36.4±1.2
RETrace (Ours)	outcome reward with credit assign	73.1±1.2	54.4±1.2	64.0±2.7	47.3±1.0	74.0±1.0	56.1±1.3	66.7±1.4	52.8±2.1	**73.5**±0.8	**55.2**±0.9	**65.3**±1.5	**50.0**±1.2
w/o traceback	outcome reward with credit assign	72.0±1.7	50.7±1.5	56.9±1.5	45.4±1.5	61.1±2.0	51.9±2.6	63.2±1.9	45.0±3.2	66.6±1.3	51.3±1.5	60.1±1.2	45.2±1.8
w/o report anchor	outcome reward with credit assign	69.0±1.6	46.3±2.8	61.3±1.5	39.4±1.6	68.0±2.8	49.2±2.5	64.0±1.5	44.1±2.2	68.5±1.6	47.8±1.9	62.7±1.1	41.8±1.4
Scaling to a larger base model													
Qwen3.5-27B	-	59.9±0.7	40.4±1.4	51.4±1.4	39.0±1.8	51.4±1.4	39.0±1.8	59.9±0.7	40.4±1.4	55.7±0.8	39.7±1.1	55.7±0.8	39.7±1.1
SFT	Qwen3.5-397B + skill	68.1±1.1	53.8±1.2	59.5±1.2	47.0±2.4	64.5±2.4	47.5±2.2	61.3±1.8	48.5±1.8	66.3±1.3	50.7±1.3	60.4±1.1	47.8±1.5
+GRPO	outcome reward	61.3±1.7	47.6±1.5	56.5±1.3	47.6±1.7	61.2±2.4	45.5±2.0	63.9±1.4	50.9±2.3	61.3±1.5	46.6±1.3	60.2±1.0	49.3±1.4
+RETrace (Ours)	outcome reward with credit assign	73.0±1.3	56.3±1.8	60.3±1.6	50.0±1.9	68.0±0.8	53.2±1.5	73.2±1.6	58.6±1.3	**70.5**±0.8	**54.8**±1.2	**66.8**±1.1	**54.3**±1.2

with AdamW using a learning rate of 1e-6 and weight decay of 0.1. All training experiments run on 16 GPUs.

Evaluation. Each trained model is obtained from a single training run and evaluated with 3 independent inference runs for each range at temperature 0.6, where each run is scored by the same three judges with the same checklists in the same harness as described in **??**. We report the averaged score and the standard error of the mean over the 3 runs.

F.2 CREDIT ASSIGNMENT DETAILS ON SECRESPOND

The training judge is Qwen3.5-397B. For every checkpoint it scores, the judge also returns the report section that satisfies it and the evidence that section relies on, such as file paths, commands, or log lines. The section is located by exact matching of the quoted span against the report, and 95.5% of the spans are matched. A checkpoint that scores zero or matches no section receives no credit. For detection checkpoints, each evidence item is matched against the tool results of the previous steps, and the earliest step whose tool result contains the evidence is credited. 84.5% of the satisfied detection checkpoints are traced to at least one step. For the remaining checkpoints with no matched evidence, only the report section is credited. Planning checkpoints are anchored to the remediation plan and never traced, since they are concluded from the findings rather than detected by a single step. Checkpoint scores are normalized within the group of $G = 8$ rollouts, and the balanced weight of the additional credit, denoted as λ, is set to 0.3. In the *w/o traceback* ablation, all credits are assigned to the report section. In the *w/o report anchor* ablation, only the traced steps are additionally credited and the report section receives zero credit.

F.3 FULL RESULTS

Table 18 reports the results of each fold separately. RETrace exceeds skill-guided SFT on both ID and OOD ranges, whereas GRPO and GSPO fail to achieve remarkable improvement. Crediting the report alone does not consistently improve detection, so removing the credit traced back to investigative steps harms the detection performance on both folds. Compared with the larger model Qwen3.5-397B, the fine-tuned 9B model approaches its performance prompted with skill even in OOD scenario. We additionally train with OPD guided by the skill-prompted teacher, but the performance is worse than OPD with the plain teacher on both folds. A likely reason is the larger gap between the teacher and the student. Since the guidance is conditioned on the student's own context, it rarely provides signal along the teacher's investigation trajectory. This problem is amplified when the teacher is prompted with a skill, making the guidance even less informative about where to investigate.

Table 19: Range-level detection/plan scores (%) on in-domain ranges. Each range is scored by the model trained on the fold that contains it; Overall equals the ID average in Table 3. Bold marks the best among SFT, GRPO, and RETrace.

model	SSH-Miner		Redis-RCE		Docker-Escape		Jenkins-RCE		Shiro-Fastjson		Log4j-RCE		Next.js-RCE		NPM-Worm		ASP.NET-ViewState		RDP-Service-Abuse		Overall	
	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan
Baseline	70.0	52.0	34.7	28.4	57.9	24.6	55.6	23.7	46.7	29.7	80.8	50.9	64.7	44.0	49.6	29.5	46.0	17.4	58.0	29.6	56.4	33.0
SFT	76.3	**67.6**	**78.2**	**64.7**	75.5	45.6	**70.2**	**41.2**	58.3	41.7	82.3	50.9	82.4	65.5	55.0	49.4	63.1	30.3	51.9	**48.1**	69.3	50.5
GRPO	69.1	52.0	72.8	58.2	76.9	45.8	57.2	35.3	55.3	46.9	85.2	69.4	74.8	60.1	60.4	**59.0**	58.6	22.0	45.6	35.6	65.5	48.5
RETrace	**78.7**	**67.6**	77.3	58.8	**80.1**	**61.4**	63.2	30.7	**71.9**	**55.7**	**89.4**	**74.1**	**85.9**	**68.5**	**61.5**	53.2	**71.7**	**35.6**	**55.6**	47.2	**73.5**	**55.2**

Table 20: Range-level detection/plan scores (%) on OOD ranges. Each range is scored by the model trained on the other fold; Overall equals the OOD average in Table 3. Bold marks the best among SFT, GRPO, and RETrace.

model	SSH-Miner		Redis-RCE		Docker-Escape		Jenkins-RCE		Shiro-Fastjson		Log4j-RCE		Next.js-RCE		NPM-Worm		ASP.NET-ViewState		RDP-Service-Abuse		Overall	
	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan	det	plan
Baseline	70.0	52.0	34.7	28.4	57.9	24.6	55.6	23.7	46.7	29.7	80.8	50.9	64.7	44.0	49.6	29.5	46.0	17.4	58.0	29.6	56.4	33.0
SFT	75.8	64.7	56.5	46.1	**69.9**	45.6	49.1	30.7	39.8	33.3	**84.8**	63.0	**70.6**	**57.7**	**59.5**	45.5	42.4	16.7	50.0	47.2	59.9	45.0
GRPO	72.8	67.6	60.2	46.1	49.1	**53.5**	59.1	21.9	38.1	21.6	81.8	**65.7**	60.0	42.9	48.5	37.7	44.5	23.2	50.6	47.9	56.5	42.8
RETrace	**81.6**	**70.6**	**72.2**	**63.7**	69.4	41.2	**61.4**	**41.2**	**58.8**	**43.8**	77.8	58.3	63.1	50.6	58.7	**47.4**	**57.8**	**24.2**	**52.5**	**59.3**	**65.3**	**50.0**

Step-level credit, which scores each step by judging it directly with the rubric checkpoint, reaches 58.8%/37.5% on ID and 57.7%/36.4% on OOD ranges, which is lower than RETrace w/o report anchor ablation which also credits steps only but through the checkpoints accepted in the report. This demonstrates that even when only the steps are credited, anchoring the credit on the report is necessary. On Qwen3.5-27B, RETrace exceeds SFT as well, showing that our proposed method can be generalized to a larger base model.

F.4 Failure Analysis of Training Methods

On-Policy Distillation. We examine why on-policy distillation falls below the base model although it learns from the same teacher as SFT. Figure 12(a) shows that the KL between the teacher and the student on the student's own rollouts decreases steadily under both KL directions, so the student increasingly matches the teacher on the tokens it produces, while the number of tool calls rises as well. To analyse what the student learns from the teacher, we divide the tool calls in the teacher's trajectories into two types. *Key-evidence action* returns a result that contains an evidence item required by the checklist, such as reading the malicious systemd service in Log4j-RCE. *Routine check* returns no such evidence but confirms that a location is clean, such as listing `/etc/init.d` when no persistence is placed there, which ensures the coverage of the investigation and the completeness of the remediation plan. A teacher action is counted as reproduced if the student inspects the same file or directory at any step of its trajectory. Figure 12(b) reports the proportion of these two types of actions during evaluation after OPD and SFT. The SFT student reproduces more of the teacher's actions in both types, while OPD fails to transfer 36.7% of the key-evidence actions and 60.4% of the routine checks. This is because the token-level guidance is conditioned on the student's own context. It aligns the student with the teacher on the student's predicted actions, but rarely provides signal along the teacher's investigation trajectory, whereas SFT conditions on the teacher's trajectories and learns these checks directly.

Outcome-reward RL. We first examine why the outcome reward fails to improve the SFT model and why adjusting credit assignment with our proposed RETrace is effective. Figure 13 represents the training curve of reward, response length and tool call counts. The reward of GRPO fluctuates whereas that of RETrace rises during training, indicating that the signal in the outcome reward fails to encourage the improvement of the policy. Neither the report length nor the number of tool calls increases, so the gain of RETrace comes from a more targeted investigation rather than longer outputs or more investigation steps. Figure 4 (a) further conducts statistics on the number of required evidence items that appear in the tool outputs of a trajectory. GRPO discovers fewer evidence items than SFT (19.6 *vs.* 20.3 on ID ranges and 16.8 *vs.* 17.9 on OOD ranges), whereas RETrace discovers the most (21.1 and 19.0). Figure 4 (b) further splits the detection checkpoints into the alerted and silent groups defined in §D.4. GRPO falls below SFT, especially on the silent checkpoints. RETrace achieves the largest gain on the silent checkpoints of the OOD ranges (57.9% *vs.* 52.7%). The outcome reward

therefore does not enhance the investigation toward the silent evidence. In contrast, RETrace credits these steps through the findings they introduce into the report, thus making the investigation move toward the evidence that the alerts do not expose and achieve deeper investigation.

F.5 Detailed Results of RETrace

Range-Level Analysis. Tables 19 and 20 report the range-level scores of the base model, SFT, GRPO, and RETrace on ID and OOD ranges. On ID ranges, RETrace obtains the highest detection score on 8 of the 10 ranges and the highest planning score on 5 ranges. On OOD ranges, RETrace obtains the highest detection score on 6 of the 10 ranges and the highest planning score on 7 ranges. The gains of RETrace are therefore not concentrated on a few ranges. RETrace gains most over GRPO on Shiro-Fastjson, ASP.NET-ViewState, SSH-Miner, RDP-Service-Abuse, and Redis-RCE. The persistence mechanisms in these ranges are scattered in several places at once, such as crontab, systemd, SSH keys, or WMI. The improvement on these ranges demonstrates that RETrace can reward each step that uncovers significant clues and contributes to the final output, so the policy learns to keep checking in-depth locations instead of stopping at the first finding. However, RETrace falls below SFT on Jenkins-RCE and Redis-RCE in the in-domain setting and on ranges such as Next.js-RCE in the out-of-domain setting, whose attack chains rely on a SUID binary and an `LD_PRELOAD` rootkit that should be concluded from several tool results. This is probably because RETrace traces a checkpoint back to the step whose tool result contains the evidence, but when no single result contains it, the credit goes to the report section only and the related steps not explicitly containing the tool result receive nothing. Extending the traceback to a finding that is concluded from multiple steps is a promising next step.

Capability-Level Analysis. Figure 14 aggregates the checkpoint scores into the five capability dimensions following Table 6. In detection, RETrace is the strongest on all five dimensions on ID ranges, especially on baseline assessment (BAS) and vulnerability verification (VUL). On OOD ranges, the three trained models perform similarly on BAS and VUL, while RETrace excels on investigation quality (Q), persistence (PER), and entity identification (ENT). These dimensions are rarely exposed by alerts and require deeper investigation, which is consistent with the gain of RETrace on detecting silent intrusion. In planning, RETrace leads on BAS, PER, and ENT on ID ranges, and is the highest on all five dimensions on OOD ranges.

Credit Assignment Example. Figures 15 and 16 show the credit assigned to the intermediate step and the report sections of one Log4j-RCE rollout. The credit concentrates on the turns whose results contain the evidence of a checkpoint, such as the reads of the exploit class and the malicious systemd service. Within a single report, the sections related to checkpoints with judge score above the group are reinforced while those below the group are penalized.

G Limitation and Future Direction

SecRespond evaluates whether agents can autonomously conduct post-compromise response in an end-to-end manner, and the results show that the gap lies in investigation depth. We cover 10 cyber ranges, each reproducing one attack chain on a single host with disk snapshots. However, real incidents may span multiple hosts or involve live attacks. On the training side, RETrace assigns the credit of each rubric checkpoint to the corresponding output span and the investigation steps, which is a first step toward learning deeper investigation from outcome rewards. We leave the generalization to long-horizon tasks with open-ended outputs to future work.

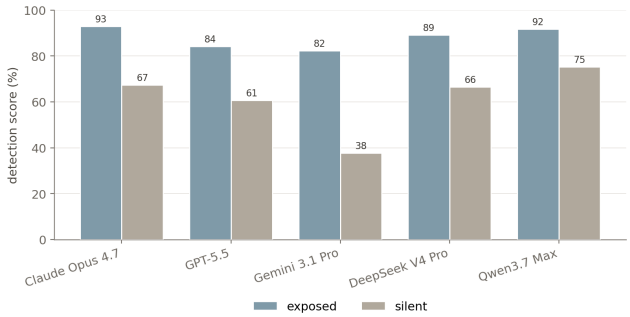


Figure 6: Detection score of frontier models on alerted and silent intrusion checkpoints. All models drop sharply on silent checkpoints, exposing a gap in proactively investigating deep evidence.

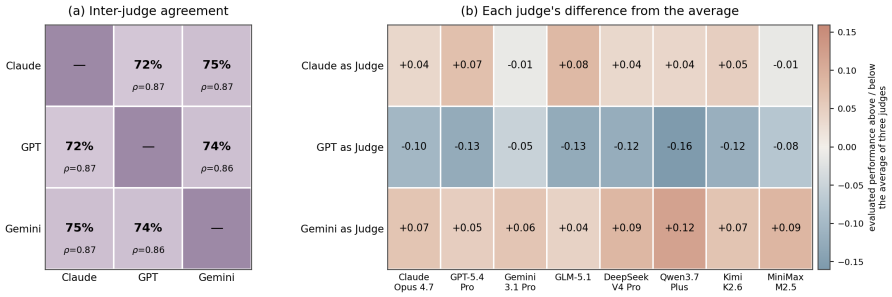
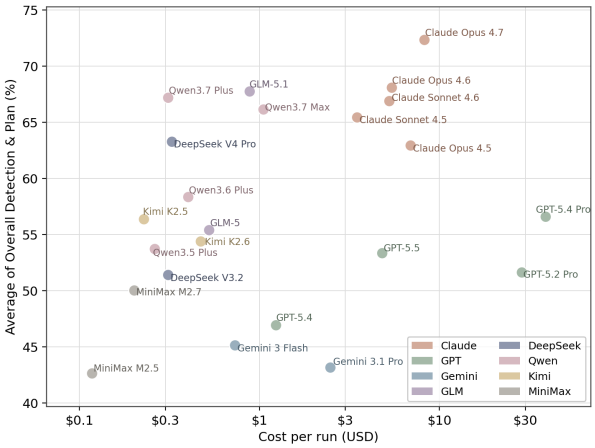


Figure 7: **(a) Inter-Judge Agreement.** For each pair of judges, the value denotes the percentage of checkpoints on which both judges assign the same score, and ρ is the Spearman rank correlation between their scores. **(b) Each Judge's Difference from the Average.** For each judge (row) and the representative model of each family (column), we compute the average difference between that judge's score and the mean score of the three judges across all checkpoints.

Model	Steps	Input	Output	Cost (USD)
Claude Opus 4.7	32.1	1,451,154	38,809	8.23
Claude Opus 4.6	24.4	983,689	19,933	5.42
Claude Opus 4.5	26.3	1,269,874	22,442	6.91
Claude Sonnet 4.6	31.4	1,639,221	23,099	5.26
Claude Sonnet 4.5	28.5	961,823	39,708	3.48
GPT-5.5	15.4	846,206	18,659	4.79
GPT-5.4 Pro	19.4	1,150,807	23,794	38.81
GPT-5.4	8.6	429,288	10,443	1.23
GPT-5.2 Pro	20.2	1,202,498	20,534	28.70
Gemini 3.1 Pro	38.5	1,187,974	8,347	2.48
Gemini 3 Flash	36.1	1,415,708	7,489	0.73
GLM-5.1	18.2	855,301	10,321	0.88
GLM-5	18.6	828,880	14,462	0.53
DeepSeek V4 Pro	15.8	721,040	14,010	0.33
DeepSeek V3.2	47.2	1,331,327	17,879	0.31
Qwen3.7 Max	14.8	669,566	13,746	1.05
Qwen3.7 Plus	16.8	899,240	17,257	0.31
Qwen3.6 Plus	20.6	1,141,624	15,410	0.40
Qwen3.5 Plus	16.1	781,664	14,536	0.26
Kimi K2.6	18.7	778,775	13,151	0.47
Kimi K2.5	14.7	518,576	16,348	0.23
MiniMax M2.7	19.5	788,428	12,738	0.20
MiniMax M2.5	29.4	934,533	10,164	0.12

(a)



(b)

Figure 8: Token Usage. (a) Average per-run agent steps, input tokens, output tokens, and cost. (b) Average Overall Detection and Planning score vs. cost per run.

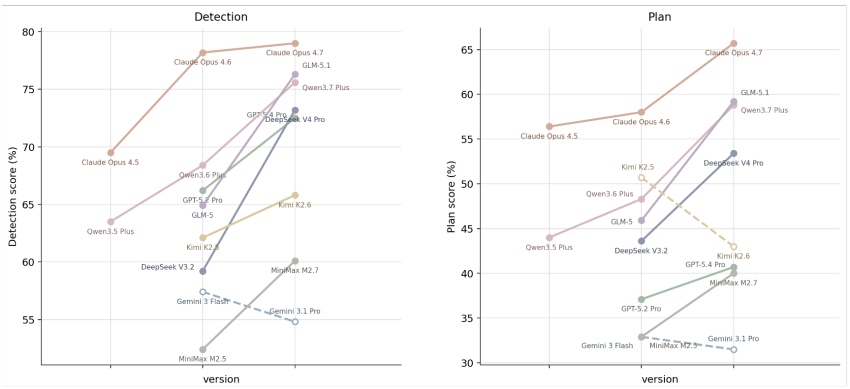


Figure 9: **Version evolution within each model series.** Most model families improve across versions, but the progress is not consistent.

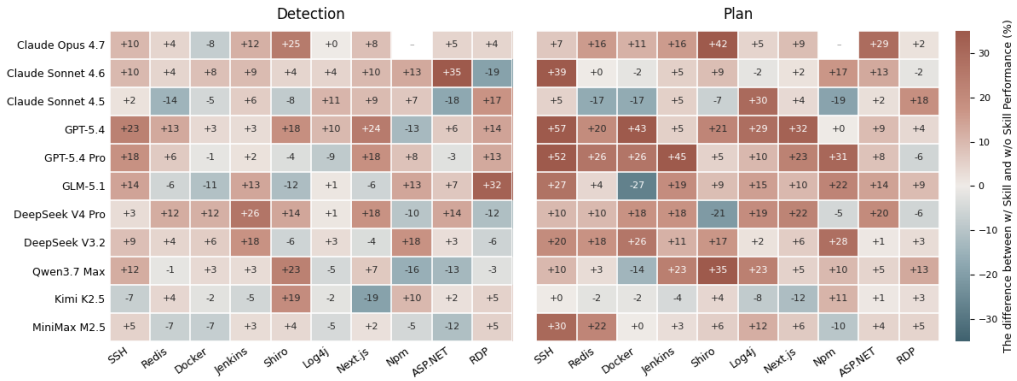


Figure 10: Changes in the range-level Detection and Planning CHK-scores of representative models on each range after providing our designed skill. Each value denotes the difference between runs with and without the skill.

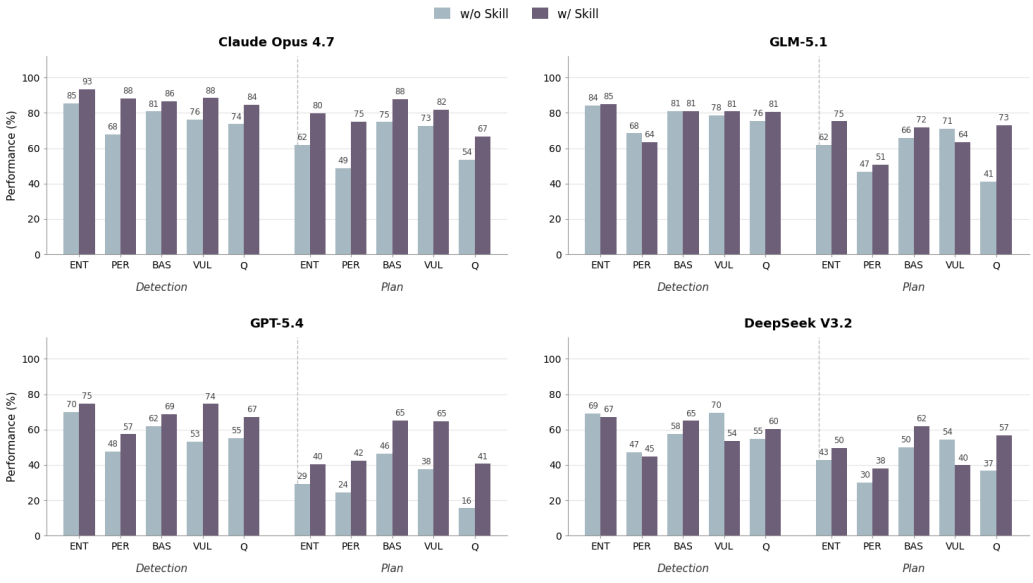


Figure 11: The Detection and Planning CAP-scores across the five capability dimensions for four representative models with and without our designed skill.

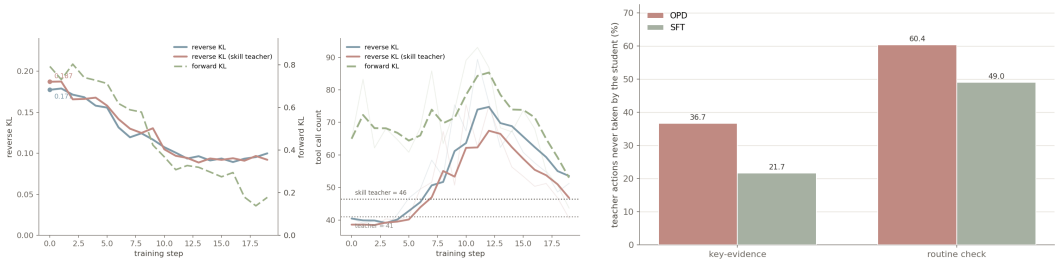


Figure 12: Analysis of on-policy distillation. (a) The KL divergence between the teacher and the student and the number of tool calls during training. (b) The proportion of teacher actions that the student never takes on the evaluation ranges after OPD and SFT, classified into key-evidence actions and routine checks.

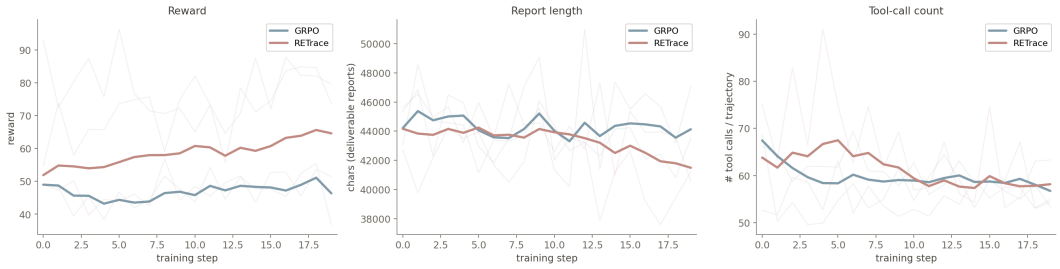


Figure 13: The training curve of GRPO and RETrace.

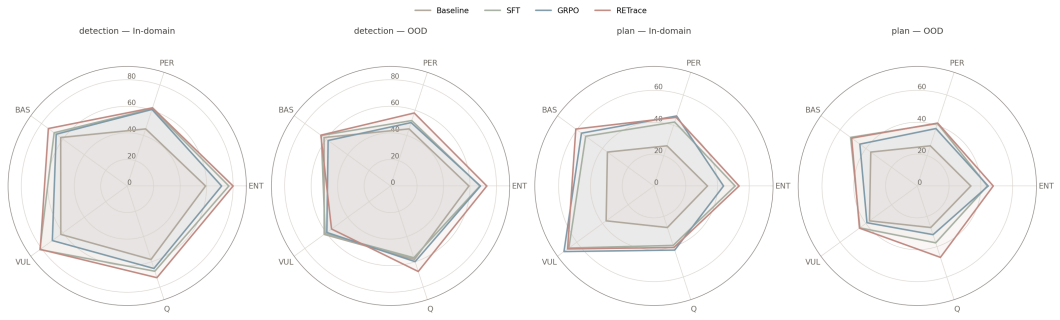


Figure 14: CAP-scores of the base model, SFT, GRPO, and RETrace across the five capability dimensions on ID and OOD ranges.

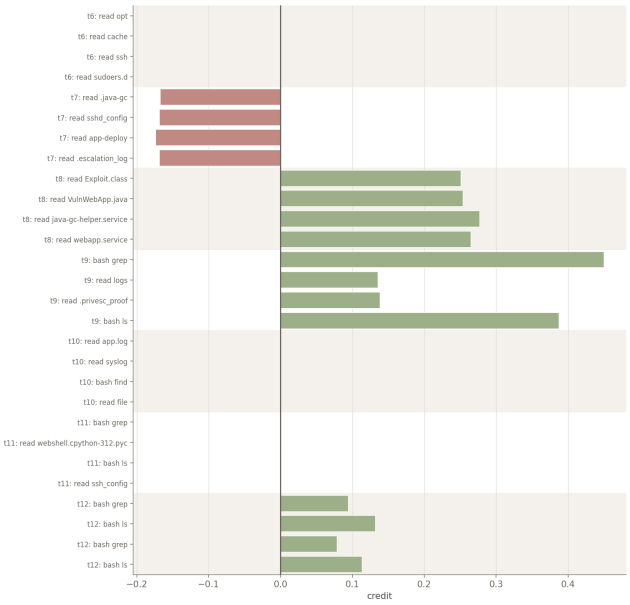


Figure 15: Credit of selected steps in an Log4j-RCE rollout during RETrace training

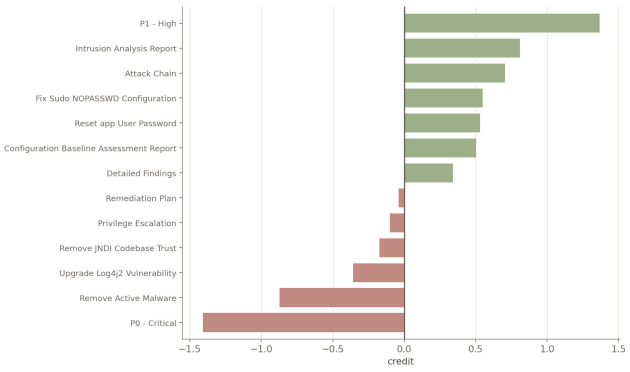


Figure 16: Credit of selected output spans in an Log4j-RCE rollout.