EARNED REDUNDANCY: AUDITING LLM SKILL-PROMOTION QUORUMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Promotion gates combine language-model reports to decide whether reusable agent skills can be admitted. Different model names, prompts, or endpoints do not establish that agreement reduces malicious admission while retaining benign skills. We introduce an intervention-conditioned audit that separates observer identity from acquired evidence, measures joint failure, and tests decision benefit against a declared reference under an explicit missing-report policy. On an identity-disjoint panel of 120 malicious packages, the sole preregistered dependence contrast yields $C = 0.665$ (95% interval $[.390, 1.026]$, one-sided $p = 0.00010$). Yet the least dependent pair jointly clears 14/120 packages, versus 13/120 for the more dependent prompt pair, because their error margins differ. In calibrated simulation, target clones reverse the sign of proxy-adjusted dependence. A paired evidence-view audit changes joint clearance from 16/28 to 0/28 on an outcome-selected cohort. A benign-inclusive Claude–Qwen extension finds lower raw cross-family correlation but no resolved balanced-accuracy gain for the calibration-selected rule, with the dependence advantage sensitive to marginal-rate normalization. Redundancy credit therefore requires a declared intervention, evidence view, and measured risk–utility improvement against the reference.

1 INTRODUCTION

Language-model agents package experience as reusable skills and workflow memory (Wang et al., 2025a;b; Zheng et al., 2025). Promotion gates decide which skills later agents may install and follow. Backdoors, poisoned trajectories, and security drift can outlast the originating run (Yang et al., 2024; Badhe & Tiwari, 2026; Chen et al., 2026a; Luo et al., 2026; Ying et al., 2026; Mao et al., 2026; Li & Zhu, 2026), making registries supply-chain targets. Hundreds of malicious skills were found in one public marketplace within weeks of launch (Seetharam et al., 2026).

Agent-security scanners use model-based inspection (Cisco AI Defense, 2026; Snyk, 2026; Beurer-Kellner & Fischer, 2025), alongside research on malicious-package detection and skill auditing (Zahan et al., 2025; Wang et al., 2026b; Gao et al., 2026; Shang et al., 2026) and benchmarked malicious/benign artifacts (Wang et al., 2026c). Reports can be combined by voting (Verga et al., 2024; Hu & Qu, 2026), monitor ensembles (Koran et al., 2026), or flagging whenever any scanner does (Hu et al., 2026b). Threshold aggregation has precedents in anti-malware pipelines (Zhu et al., 2020; Kantchelian et al., 2015) and replicated systems (Gifford, 1979; Malkhi & Reiter, 1998). When admission requires at least k of n clear reports, how much assurance does a second clear add?

Identity alone cannot answer this question. Independently developed programs share failures more often than independence predicts (Knight & Leveson, 1986; Eckhardt & Lee, 1985; Littlewood & Miller, 1989), and redundancy protects insofar as components fail differently (Littlewood & Strigini, 2004). Models from different providers also share many errors, particularly the largest models (Kim et al., 2025). A second signature identifies a report's source without establishing independent failure. Correlation is insufficient too. Joint false-clear risk combines marginal error rates and covariance, so a less correlated pair with a weaker member can accept more malicious skills. Difficulty adjustment can also reverse residual dependence when its vote-derived proxy contains target-observer clones (Section 3).

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.

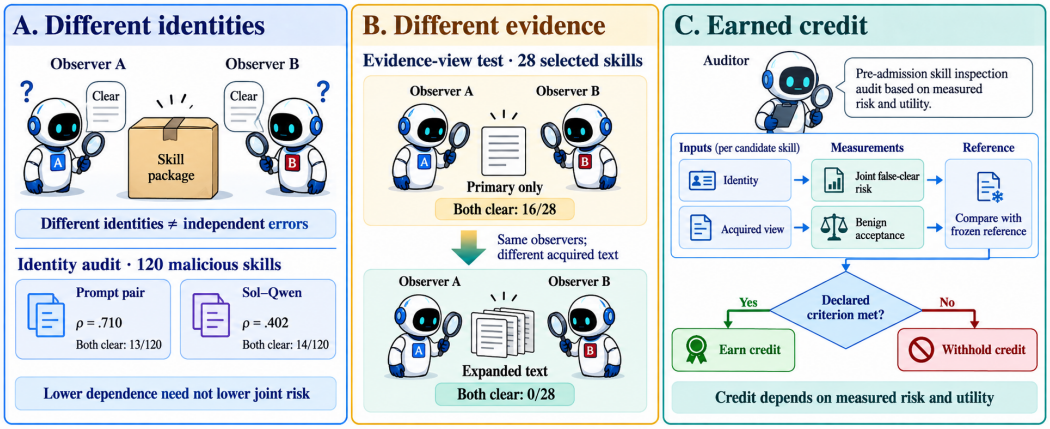


Figure 1: Auditing redundancy credit. (A) Different identities do not establish independent errors, and lower dependence need not lower joint false-clear risk. (B) Expanded text changes fixed observers' verdicts on a separate selected cohort. (C) Joint risk and benign acceptance are measured across labeled skills under the declared identity and evidence view, then compared with a frozen reference to determine credit.

We audit assurance through interventions on observer configuration and evidence. Redundancy credit is the assurance attributable to a second report (Littlewood & Strigini, 2004). Reports bind identity, supplied text, and labeled failure. Configuration or evidence changes while other inputs stay fixed. Paired packets reveal text beyond the primary file under fixed manifests and observers (Section 3). Given a roster, view, malicious/benign panel, and independently calibrated frozen reference, the audit measures joint false-clear risk and balanced-accuracy gain, accounting for blocked benign skills. Credit requires improvement under the declared risk–utility criterion and deployed view (Figure 1C).

MaliciousSkillBench supplies exploratory and identity-disjoint preregistered panels spanning provider, prompt, temperature, and model family, including GPT-5.6 (Sol) and Qwen3.6-27B (Qwen). A sealed cross-corpus panel tests portability (OpenA2A, 2026; Anthropic, 2026). A balanced Claude Sonnet 4.6–Qwen3.8-27B extension selects operating points on disjoint calibration labels.

The results separate three ways an apparent redundancy benefit can fail. The preregistered contrast confirms higher same-model dependence, but the least dependent cross-family pair jointly clears slightly more malicious packages because its error margins differ (Section 5). In a separate selected failure cohort, expanded text changes joint clearance from 16 of 28 to 0 of 28 for two fixed GPT-5.6 observers. These cohorts do not compare identity and acquisition interventions directly. Target clones reverse residual dependence in simulation despite conditional independence. On benign- inclusive panels, neither exploratory quorums nor the calibration-selected Claude–Qwen rule establish a BA gain over the relevant best or selected member. The Claude–Qwen reference's malicious false clears are fully shared by its paired observer.

Our contribution is a decision-level account of redundancy credit. The audit separates provenance, acquired evidence, and joint failure, then relates dependence to malicious risk and benign acceptance (Sections 2.2 and 3). Identity-disjoint confirmation and paired hide/reveal packets test these relationships with sealed protocols (Section 4, Appendix B). The results connect dependence, acquisition, and missing-report handling to admission decisions (Section 5). We release the protocol, sealed observations, and analysis code.

2 Background and Problem

2.1 Related Work

Skill-promotion quorums link artifact security, ensemble reliability, and risk-aware evaluation through report identity, acquired evidence, and decision benefit.

Agent artifacts and their security lifecycle. Tool-use methods and benchmarks study interactive decisions (Yao et al., 2023; Qin et al., 2024; Huang et al., 2024; Liu et al., 2024; Ruan et al., 2024). Workflow memory and induced skills enable reuse (Wang et al., 2025b;a; Zheng et al., 2025). Contamination, backdoors, poisoning, and security drift can persist beyond source trajectories (Yang et al., 2024; Badhe & Tiwari, 2026; Chen et al., 2026a; Luo et al., 2026; Ying et al., 2026; Mao et al., 2026; Li & Zhu, 2026). SkillAudit and pre-commit gates inspect admission candidates (Gao et al., 2026; Shang et al., 2026). MaliciousSkillBench provides malicious and benign artifacts (Wang et al., 2026c). We audit inspection-report aggregation.

LLM judges and diversity. Diverse juries can improve evaluation (Verga et al., 2024); evaluators can favor their own outputs (Panickssery et al., 2024). Judge accuracy limits label savings (Dorner et al., 2025), and judicial benchmarks expose bias (Hu et al., 2026a). Process supervision and semantic uncertainty support reasoning and confidence (Lightman et al., 2024; Kuhn et al., 2023).

Correlated failure and ensemble benefit. Multiversion studies demonstrate coincident failures and model shared difficulty (Knight & Leveson, 1986; Eckhardt & Lee, 1985; Littlewood & Miller, 1989). Ensemble analysis relates benefit to member strength and error dependence (Breiman, 2001; Kuncheva & Whitaker, 2003). Shared decisions have welfare consequences (Kleinberg & Raghavan, 2021). In skill promotion, target-like observers can contaminate difficulty adjustment, and invalid reports change deployed decisions.

Security evaluation and evidence views. Agent-security benchmarks test harmful tool use and prompt injection (Andriushchenko et al., 2025; Zhan et al., 2024; Debenedetti et al., 2024; Zou et al., 2025). MCP, managerial-decision, and software-security benchmarks extend coverage (Zong et al., 2026; Simhi et al., 2026; Lee et al., 2025). Sequential monitoring and runtime policies (Chen et al., 2026b; Wang et al., 2026a), safe task composition (Leahy et al., 2024), and program verification (Wu et al., 2024) extend assurance beyond text inspection. Evaluation awareness can affect behavior (Hua et al., 2026). Paired hide/reveal isolates acquired content for fixed observers.

Provenance and decision guarantees. SLSA, SCITT, and diverse-context signing authenticate artifacts and claims (Supply-chain Levels for Software Artifacts, 2025; Birkholz et al., 2026; Okafor et al., 2025). Proof-carrying and runtime systems check proofs or enforce policies (Necula, 1997; Ball et al., 2024; Wang et al., 2026a), without establishing semantic independence. Selective prediction, conformal methods, and prompt risk control link coverage or selection decisions to risk under explicit assumptions (Geifman & El-Yaniv, 2019; Angelopoulos et al., 2024; Quach et al., 2024; Zollo et al., 2024). Small panels can destabilize rankings (Yauney et al., 2026). Our intervention-conditioned audit reports risk, benign utility, and uncertainty without inheriting a distribution-free deployment guarantee.

2.2 Problem formulation

A promotion service admits reusable skills to an agent library using observer reports about fixed artifacts. Later execution is evaluated separately. Let C be a candidate skill and $Y \in \{0,1\}$ its evaluation label, with $Y = 1$ denoting malicious content that should not be promoted. Observer i receives a registered view $X_i(C)$ and reports clear, protected, or insufficient. Let $V_i = 1$ denote passing schema, cardinality, identifier, and binding checks. Complete-clear requires $V_i = 1$, complete coverage, untruncated input, and a clear verdict. The formal fail-closed map sets $A_i(C) = 1$ for complete-clear and 0 otherwise. Write $A = (A_1, \ldots, A_n)$. For malicious candidates define

$$E_i = \mathbb{1}\{\text{observer } i \text{ returns complete-clear}\}. \tag{1}$$

Pairwise all-clear error is $E_i = E_j = 1$. A promotion rule $g : \{0,1\}^n \to \{0,1\}$ has malicious-promotion risk, benign acceptance, and balanced accuracy

$$\begin{aligned} &R_M(g) = \Pr(g(A) = 1 \mid Y = 1), \qquad U_B(g) = \Pr(g(A) = 1 \mid Y = 0), \\ &\mathrm{BA}(g) = \tfrac{1}{2}\{1 - R_M(g) + U_B(g)\}. \end{aligned} \tag{2}$$

Benign utility, unlike failure-conditioned dependence, requires benign items.

The audit separates identity, view, and failure. Identity records the key, endpoint, provider, process, prompt, and call. View records the primary text, package files, provenance, registry state, or trace

supplied to an observer. Failure concerns class-conditional joint errors and the decision rule. Identity checks and evidence inventories establish provenance and view, but not the joint-error distribution.

For a pair (i, j) the registered full-denominator dependence estimand is

$$\rho_{ij} = \mathrm{Corr}(E_i, E_j \mid Y = 1), \tag{3}$$

reported with marginal and joint false-clear rates under the frozen report map. Secondary paired-valid correlation additionally conditions on $V_i = V_j = 1$. Deployed rules retain full denominators under explicit missing-report policies.

Shared item difficulty can induce correlation. Adjusting for it through surrogates of the tested outcomes can distort dependence. Let $F(i)$ be observer i's model family and let $\mathcal{O}_{-F(i,j)}$ contain only observers outside both target families. On a fixed item panel $\mathcal{P}$, we define the clone-safe proxy

$$D_{-F(i,j)}(C) = \frac{1}{|\mathcal{O}_{-F(i,j)}|} \sum_{k \in \mathcal{O}_{-F(i,j)}} E_k(C). \tag{4}$$

The all-six-valid panel requires valid normalized observations from every observer. Excluding target-family copies on this fixed panel prevents changing proxy composition. Package covariates and simulation test sensitivity to the remaining proxy error.

The exploratory audit specifies one confirmatory contrast. Let ρ_P compare Sol under two prompts, ρ_T compare Qwen under two temperatures, and ρ_X compare Sol with deterministic Qwen. With $z(r) = \mathrm{atanh}(r)$, define

$$C = \tfrac{1}{2}\{z(\rho_P) + z(\rho_T)\} - z(\rho_X). \tag{5}$$

The sole directional hypothesis is $C > 0$ on a new fixed roster. Invalid or missing reports are non-clear. A 20,000-draw item bootstrap must give both a positive 95% lower bound and one-sided $p < .05$.

Research questions. RQ1 tests the registered dependence contrast on disjoint identities and its relation to joint risk under unequal margins. RQ2 tests sensitivity to margins, clones, and proxy construction. Secondary audits assess utility, acquisition, missingness, and runtime behavior under separate protocols.

3 Clone-Safe Three-Layer Audit

Identity and evidence view Reports bind the candidate digest, provider, endpoint, prompt, sampling setting, call identifier, validity, and coverage. Evidence is recorded separately. Let $\mathcal{M}(C)$ be the text-file manifest available to the gate, $\mathcal{S}_i(C) \subseteq \mathcal{M}(C)$ the files actually supplied to observer i, and $\mathcal{P}_i(C) \subseteq \mathcal{S}_i(C)$ the exact paths attributed in a protected report. Thus $\mathcal{P}_i \subseteq \mathcal{S}_i \subseteq \mathcal{M}$ distinguishes attribution, acquisition, and availability, recorded through validated paths, packets, and manifest geometry, respectively. Hide/reveal fixes observer, prompt, item, and manifest while changing only $\mathcal{S}_i$, making it an acquisition intervention rather than geometric correlation.

Registered contrast and dependence audit The identity-disjoint panel and report map are frozen before calls. A fixed-seed bootstrap (Efron, 1979) jointly resamples four-observer rows, recomputing all three 2×2 tables, correlations, and C while preserving margins and covariance among overlapping estimates. Both registered criteria must pass. Complete cases are secondary, and inference concerns the declared perturbation axis rather than random effects over families. We report each pair's margins, joint error, independence product, Pearson correlation, attainable ceiling, and paired-bootstrap interval. For paired cells $(n_{11}, n_{10}, n_{01}, n_{00})$, correlation is

$$\widehat{\rho} = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}}}. \tag{6}$$

We enumerate leave-pair-out proxy members, then apply Equation 4 on the complete panel. For member i we fit $\widehat{p}_i(C) = \sigma(\alpha_i + \beta_i D_{-F(i,j)}(C))$ and correlate Pearson residuals $(E_i - \widehat{p}_i)/\sqrt{\widehat{p}_i(1-\widehat{p}_i)}$, refitting both models in each bootstrap draw. Holm correction covers all 15 contrasts among six prespecified pair correlations (Holm, 1979). Target-independent package covariates provide another sensitivity check. Simulation under known conditional independence compares zero-, one-, and two-clone vote proxies, calibrating clone agreement from the sealed matrix. It isolates estimator distortion rather than corpus bias.

Decision and missingness audit Let $K(C)$ and $M(C)$ count complete-clear and missing/invalid reports. Equation 9 encodes the latter as non-clear ($m = 0$) or clear ($m = 1$). For every q, we report Equation 2 with full-denominator bootstrap uncertainty at both endpoints, comparing each quorum with every individual and separating calibration from evaluation. Cross-corpus reuse is descriptive because it was registered after matrix inspection.

Endpoint differences locate decisions affected by report handling rather than establishing semantic detection.

A reference must be declared in advance or calibrated on representative labels before disjoint testing. A retrospective best measures reference regret. Without calibration labels, no empirical credit is awarded.

View intervention and runtime diagnostic The view experiment creates two opaque packets per selected item. Let $s_0(C)$ denote the primary instruction file and $\mathcal{M}_{\text{text}}(C) \subseteq \mathcal{M}(C)$ the eligible UTF-8 files. The control has $\mathcal{S}_i^{(0)} = \{s_0(C)\}$ and the expanded packet $\mathcal{S}_i^{(1)} = \mathcal{M}_{\text{text}}(C)$. Packet order is hash-randomized, observers remain fixed, and protected verdicts name exact manifest paths. If $B_c^{(v)} = \prod_{i\in\{\text{Sol,Terra}\}} A_i(C; v)$ is joint clear, the paired estimand is $\widehat{\tau}_{\text{view}} = |\mathcal{H}|^{-1} \sum_{c\in\mathcal{H}} \{B_c^{(1)} - B_c^{(0)}\}$. Exact McNemar and paired bootstrap quantify acquisition effects on the frozen failure cohort $\mathcal{H}$, rather than prevalence.

The runtime diagnostic supplies installed packages and instrumented tools that record and block protected actions. Exact tables use functionally complete runs. Conditional power assesses whether sparse margins make non-rejection uninformative. Appendix B specifies contracts, bootstrap algorithms, and instrumentation.

Risk, credit, and estimator implications. Let $\epsilon_i = \Pr(E_i = 1 \mid Y = 1)$ under a fixed report map and $s_{ij} = \sqrt{\epsilon_i(1-\epsilon_i)\epsilon_j(1-\epsilon_j)}$.

Proposition 1 (Dependence is one component of joint risk)**.** *For non-degenerate Bernoulli errors with correlation ρ_{ij},*

$$\begin{aligned} J_{ij} &:= \Pr(E_i = E_j = 1) = \epsilon_i\epsilon_j + \rho_{ij}s_{ij}, \\ J_a - J_b &= (\epsilon_{a1}\epsilon_{a2} - \epsilon_{b1}\epsilon_{b2}) + (\rho_a s_a - \rho_b s_b). \end{aligned} \tag{7}$$

Unequal margins constrain Pearson correlation to $\rho_{\max} = \{\min(\epsilon_i, \epsilon_j) - \epsilon_i\epsilon_j\}/s_{ij}$, motivating the post-registered normalized sensitivity check.

Proposition 2 (Reference-relative redundancy credit)**.** *For a rule g and frozen reference observer r, define malicious safety gain $G_M = \epsilon_r - R_M(g)$ and benign-acceptance loss $L_B = U_B(r) - U_B(g)$. Then*

$$\text{BA}(g) - \text{BA}(r) = \tfrac{1}{2}(G_M - L_B). \tag{8}$$

Thus g improves BA exactly when its safety gain exceeds its utility loss. For the two-member all-clear rule relative to member 1, $G_M = \epsilon_1(1-\epsilon_2) - \rho_{12}s_{12}$.

This makes "beyond the strongest member" a labeled, reference-conditioned decision claim rather than a consequence of lower ρ.

Missingness boundary. With K observed clears and M missing/invalid reports, define $g_q^{(m)} = \mathbb{1}\{K + mM \geq q\}$. Then

$$g_q^{(1)}(C) - g_q^{(0)}(C) = \mathbb{1}\{K(C) < q \leq K(C) + M(C)\}. \tag{9}$$

Hence the endpoint gap is exactly the class-conditional mass at this boundary, not evidence about why reports are missing.

Finally, for an item variable D, total covariance separates within-D dependence from covariance in conditional error rates. This motivates difficulty adjustment but does not validate an estimated D. If independent fair Bernoulli errors satisfy $D^\star = E_1 + E_2$, their residuals after conditioning on $D^\star$ have correlation -1. We therefore retain direct joint-risk accounting and treat adjusted correlation as sensitivity analysis. Appendix A gives the full derivations, constructive example, and the calibrated simulation.

4 EXPERIMENTAL DESIGN

Exploratory audit and sealed reanalysis A fixed salted digest selects 10+10 calibration and 60+60 formal malicious/benign items from MaliciousSkillBench's archived, source-balanced random split, excluding development identities. Labels remain hidden until raw and normalized reports are hash-sealed. Six observers vary provider, prompt, temperature, and family under identical primary text and output contracts. Formal calls are separate, with no retries after label opening. Validity ranges from 80/120 to 120/120.

Clone-safe reanalysis reconstructs proxy membership and evaluates every pair on 37 all-six-valid malicious items, excluding target families without new calls or labels. This matrix supplies the estimator, simulation, missingness, and individual comparisons in Section 3. A post-registered 37-versus-23 selection audit uses the same covariates. Algorithms and call accounting appear in Appendices B–C.

Preregistered identity-disjoint confirmation The 12 malicious calibration and 120 formal malicious identities exclude all 900 earlier identities, with distinct formal structural families. GPT-5.6 Sol Prompts A/B and Qwen3.6-27B temperatures 0/.7 implement Equation 5 under frozen providers, routes, prompts, schemas, reasoning, batching, and no-retry rules. A symmetric calibration amendment truncates non-decision rationales while preserving raw responses and hashes, after which all observers pass. One wrong identifier invalidates a ten-item Prompt B batch, counted as non-clear. Outputs are sealed before analysis. This panel estimates dependence and joint risk, with benign utility on balanced panels. Margin-constrained correlation and log odds ratio are post-registered diagnostics.

Benign-inclusive cross-family extension This panel tests whether dependence separation translates into benefit when benign utility is included. Claude Sonnet 4.6 and Qwen3.8-27B use unchanged Prompts A/B, temperature zero, and identical primary-text packets on 20+20 calibration and 120+120 formal MaliciousSkillBench items. Maximizing calibration BA with lexicographic ties over individuals and prompt-matched cross-family pairs selects Claude A and Claude A–Qwen A. The historical seal binds calibration validity, not these reconstructed choices. Roster and output-contract amendments also distinguish this extension from preregistered confirmation. Invalids remain non-clear. We use 20,000 label-stratified paired bootstrap draws, with provider differences, selection provenance, multiplicity, batch resampling, and missingness sensitivities in Appendix B.5.

Package-view intervention Geometry covers all 60 exploratory malicious packages. The frozen intervention selects 28 historical multi-file false-clear packages meeting pre-call archive, path-safety, text-coverage, size, and exact primary-text criteria. Opaque paired packets expose primary or all eligible text under fixed manifests and observers. All 124 calibration and formal assessments are valid.

Cross-corpus decision and runtime diagnostics The sealed 36-malicious/36-benign MaliciousSkillBench, OASB, and Anthropic Skills panel (OpenA2A, 2026; Anthropic, 2026) supports post-registered Sol/Qwen portability analysis. A separate instrumented sandbox records protected actions and blocks external effects. Fixed trace adjudication yields 69/72 functionally complete episodes. Both machine and adjudicated endpoints, protocols, and power analysis appear in Appendix B.8.

5 RESULTS

5.1 RQ1: DEPENDENCE ORDERING DIFFERS FROM JOINT-RISK ORDERING

We test whether lower error dependence also lowers joint false-clear risk (Figure 2).

The registered contrast passes both criteria. Correlations $\rho_P = .710$ for prompts, $\rho_T = .860$ for temperatures, and $\rho_X = .402$ for Sol–Qwen give $C = .665$ (bootstrap 95% interval $[.390, 1.026]$, one-sided $p = .00010$). Complete cases agree (Appendix C). Replication uses disjoint identities and one item per structural family, supporting the declared contrast rather than a population-level family ranking.

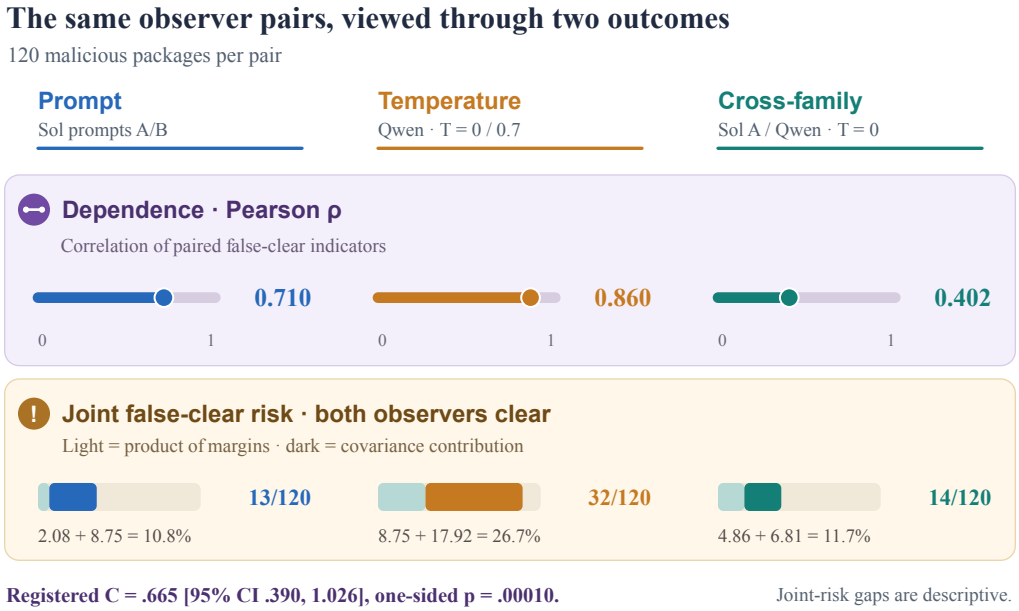


Figure 2: Three observer pairs on 120 malicious packages. The upper band shows Pearson error correlation; the lower band splits joint false-clear risk into the marginal-rate product (light) and covariance (dark). Only the Fisher-z contrast is preregistered; the risk decomposition and 14 versus 13 joint clears are descriptive (Table 2).

Prompt, temperature, and cross-family pairs jointly clear 13/120 (10.8%), 32/120 (26.7%), and 14/120 (11.7%) malicious packages. All joint rates exceed their independence products. Qwen's 35/120 marginal false-clear rate versus Sol Prompt B's 15/120 outweighs the smaller correlation term (Equation 7). This malicious-only panel cannot estimate benign acceptance, BA, or deployment superiority.

Relative to the prompt pair, cross-family margins add 2.78 percentage points to the independence product, while covariance removes 1.94, leaving $+0.83$ points, the observed 14/120 versus 13/120 gap. Marginal-ceiling normalization and log odds ratios preserve the registered direction. These post-registered diagnostics neither change absolute risk nor replace the registered statistic (Table 2). Lower dependence reduces both-clear risk at fixed margins, but replacing an observer changes both. Deployment comparisons therefore require the resulting joint rate as well as failure association.

5.2 RQ2: TARGET CLONES CAN BIAS THE ADJUSTMENT

All six exploratory raw correlations are positive (.276–.877), and every joint rate exceeds its independence product. The original leave-pair-out proxy contains target-family surrogates and changes composition under missingness. On the fixed 37-item leave-family-out panel, Sol–Qwen residual correlation is .315 $[.000, .560]$ versus the original $-.217$ (Table 4).

The 37-versus-23 package-covariate audit finds no detectable shift, weakening a geometry explanation without establishing missing-at-random (Table 5). Target-independent covariates leave residual correlations near their raw values (Appendix C). Only vote-derived adjustment changes the sign among tested proxies.

In calibrated simulation, mean adjusted correlation is $-.0002$ with known latent difficulty, .135 with no clones, and $-.199$ with two, a paired shift of $-.334$. One-clone results and all paired intervals appear in Appendix C. Proxy membership and item-varying composition are therefore necessary to interpret adjusted correlation alongside joint risk. Reanalysis changes both the item panel and proxy membership, so its sign change cannot be attributed to clone removal alone. Simulation isolates membership under known conditional independence, demonstrating distortion without estimating real-matrix bias. None of the 15 residual-correlation contrasts survives Holm correction (minimum $p = .210$, Table 15), leaving pair ordering unresolved.

Table 1: Calibration-selected operating points on a disjoint 120-malicious/120-benign Claude–Qwen3.8 evaluation panel. Mal. and Ben. are accepted counts. Both-clear uses Prompt A for both models. BA shows paired, label-stratified bootstrap 95% intervals. Invalid reports remain non-clear.

Operating point	Mal. /120	Ben. /120	BA % [95% CI]
Claude A (reference)	11	107	90.00 [85.83,93.75]
Claude A–Qwen3.8 A (both-clear)	11	105	89.17 [85.00,92.92]

5.3 Secondary audit: decision benefit is a separate estimand

Claude A/B and both prompt-matched cross-family pairs tie at 97.5% calibration BA, selecting A and A/A by the declared rule. Both clear the same 11 malicious items. The pair clears two fewer benign items (Table 1). Thus $\widehat{G}_M = 0$ and $\widehat{L}_B = 2/120$ in Equation 8: added selectivity yields no observed error recovery. The BA difference is -0.83 percentage points (paired bootstrap 95% interval $[-2.08, 0.00]$, exact two-sided McNemar $p = .50$). A boundary-aware conservative interval, accounting for unobserved malicious recovery, is $[-3.30, 1.72]$ points. Neither gain nor equivalence is established.

Mean same-family raw correlation exceeds mean cross-family correlation by .274 ($[.132, .439]$, nine-test Holm $p = .0119$). Batch resampling agrees ($[.124, .450]$, Holm $p = .0182$). Yet the $\rho/\rho_{\max}$ difference is $-.024$ ($[-.135, .131]$). The selected pair's $\rho = .670$ reaches its marginal ceiling: Qwen A clears every observed Claude A malicious false clear. Invalid-as-clear coding reverses the BA difference to $+0.83$ points ($[-1.25, 2.92]$), leaving gain unresolved. Exhaustive completion of 17 invalid assessments gives fixed-panel identification bounds $[-0.83, +0.83]$ points, not confidence intervals (Appendix C.5).

On the exploratory fail-closed panel, 4-of-6 has the highest quorum BA, 82.5% (95% interval $[75.8, 89.2]$), below Sol Prompt B's 84.2% $[77.5, 90.0]$. The observed difference is -1.67 percentage points. Its paired-bootstrap interval is $[-5.8, 2.5]$. No quorum exceeds the best member's point estimate, but the difference remains unresolved (Table 8). Calibration does not uniquely select q=4, so this optimum is descriptive.

Equation 8 explains q=4's difference from Sol Prompt B through $G_M = -3/60$ and $L_B = -1/60$, yielding -1.67 BA points. Missingness endpoints leave q=4's BA unchanged because one added malicious and one added benign acceptance cancel. At q=6, one malicious and ten benign boundary items change (Equation 9), raising BA from 75.8% to 83.3% (Table 8). This is report-handling sensitivity, not evidence that missingness is benign.

On the sealed cross-corpus panel, Qwen has 87.5% BA (6/36 malicious false clears, 33/36 benign accepts), versus Sol–Qwen's 62.5% (1/36 and 10/36), a descriptive -25.0-point difference $[-34.7, -15.3]$. The gain of $5/36$ in malicious blocking is outweighed by $23/36$ lost benign acceptance. Stricter selectivity alone therefore earns no credit. Both source compositions preserve this direction (Table 9), motivating the post-registered hypothesis that conservative aggregation can lose more utility than it gains in malicious blocking. The leading observed member varies with composition, motivating a label-conditioned reference (Appendix C). BA weights both classes equally. Other loss ratios or hard risk constraints may rank these rules differently.

5.4 Secondary audit: acquired content changes fixed observers

We test whether supplying more package text changes the verdicts of fixed observers (Figure 3).

Package geometry alone cannot establish what changes a verdict (Table 6). The intervention includes 28 eligible packages from a separate historical cohort of 45 joint false clears, 37 of them multi-file. Historical clearance defines eligibility, while the control is contemporaneous. Sol falls from 16/28 to 1/28 and Terra from 21/28 to 2/28, with no reverse discordances. Joint clear falls from 16/28 to 0/28, also with no reverse discordance, giving $\widehat{\tau}_{\text{view}} = -.571$ (95% interval $[-.750, -.393]$, one-sided exact McNemar $p = 1.53 \times 10^{-5}$). Expanded reports cite exact non-primary manifest paths on 27/28 packages, but these attributions have no independent adjudication as locations of malicious logic. The common manifest fixes declared scope, while additional text changes acquired evidence and the fixed observers' decisions. This effect concerns the outcome-selected cohort, not population prevalence, file-level ground truth, or benign utility. Runtime diagnostics appear in Appendix C.10.

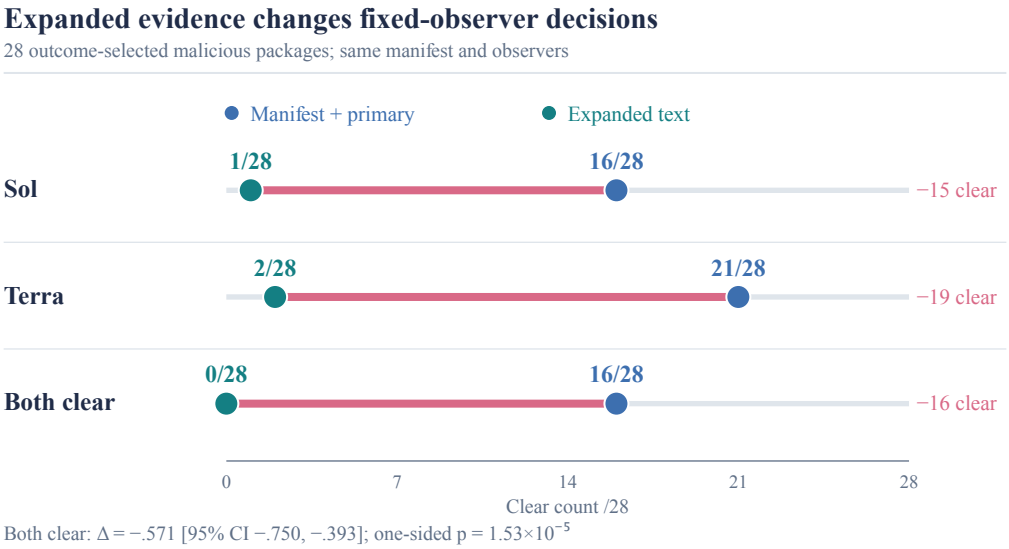


Figure 3: Paired audit on 28 outcome-selected malicious packages. Both views share a manifest; dot positions give contemporaneous primary (blue) and expanded (teal) clear counts on a common axis. No rule has a reverse discordance. Joint uncertainty uses a paired bootstrap and one-sided exact McNemar test (Table 14).

6 Discussion

Redundancy is a decision comparison. The malicious-only confirmation separates dependence from joint risk, while the balanced Claude–Qwen extension tests decision benefit. Its raw separation survives Holm correction and batch resampling, yet Qwen A repeats the reference's observed malicious clears while blocking two additional benign items. Marginal normalization leaves dependence ordering unresolved. Equation 8 therefore measures added value through recovered malicious errors and lost benign acceptance, rather than family identity or correlation rank.

The paired view intervention expands acquired package text while holding observers and manifests fixed. Its outcome-selected cohort cannot rank acquisition against the separate identity intervention. A promotion gate can declare its view, then test its rule on disjoint malicious and benign labels against a prospectively chosen reference.

Report handling belongs in the same comparison. Invalid-as-clear coding reverses the BA point difference without resolving a gain. Reindexing 15 affected assessments raises acceptance without changing that difference, distinguishing contract failure from semantic detection. Separately reporting safety gains and benign losses exposes offsets hidden by BA. Authentication supports replay. Measured improvement under the declared audit contract earns credit.

The measurement contract spans fixed-roster confirmation, outcome-selected view effects, and post-registered cross-corpus reuse. The balanced extension's calibration choices were reconstructed after formal outputs existed. Prospective selection and another data source are needed to establish a deployment benefit. The comparisons do not establish permanent family rankings or universal effects of aggregation.

7 Conclusion

Redundancy credit depends on the acquired evidence view, joint errors, and admission rule. The identity-disjoint confirmation shows that lower dependence can coexist with higher joint false-clear risk. The balanced Claude–Qwen extension finds no resolved BA gain, and report handling reverses its BA point difference. The paired view audit changes fixed observers' decisions on a selected failure cohort. These tests make credit conditional on a declared reference and risk–utility criterion.

STATEMENTS

Ethics. This work studies admission controls for reusable agent skills. The same stress recipes could help an attacker reason about gaps in a weak verifier, so the released language uses abstract operations and an effect-safe local runtime that records network attempts, subprocess requests, protected writes, credential requests, and external messages without executing them. No experiment targets an external service or user data. Static-audit requests disable provider-side storage. Runtime requests transmit only public benchmark or public Skill text; the package-view intervention also excludes sanitized credential-bearing package records before transmission. Exposed runtime credentials are synthetic canaries. The complete experimental record preserves infrastructure repairs, costs, adjudications, and observer-dependence analyses. Production use would additionally require organizational governance, privacy review, revocation, incident response, and post-promotion monitoring.

Reproducibility. The supplementary artifact contains corpus builders, excluded identities, prompts, JSON Schemas, raw provider records, compiled observations, pre-label seals, runtime traces, completion adjudications, and SHA-256 manifests. The confirmatory materials contain the frozen single-test protocol, 900-identity exclusion ledger, 12+120 blinded packets, provider-route snapshot, calibration amendment, immutable raw responses, normalization ledgers, output seal, 20,000-draw bootstrap, and machine-readable result. The clone-safe extension contains the exact 37-item panel; per-pair family exclusions; proxy-composition audit; Holm correction; external-proxy sensitivity; calibrated zero/one/two- clone simulation; both missingness curves; single-observer comparison; package-view manifest; the post-registered margin-risk and simulation-truth diagnostic; exact runtime power calculation; and unit tests. The view-intervention materials include the frozen hide/reveal protocol, official archive hashes, mechanical eligibility ledger, opaque paired packets, no-retry raw responses, file-bound observations, and paired exact test. Separate materials document the cross-corpus transfer calculation. The complete-panel selection materials include the frozen 37/23 audit, report, summary, and tests. The balanced Claude–Qwen3.8 extension retains its original and amended protocols, provider records, uniform parser replay, calibration-rule reconstruction, all pair outcomes, fixed-seed paired and batch analyses, nine-test Holm correction, and exhaustive invalid-vote bounds. Its analysis records distinguish historical validity seals from reconstructed selections. Public package contents are referenced by hash rather than bundled where third-party terms apply. All archived decisions and aggregates replay offline. Analytic details appear in Appendix A; exact observer and sandbox protocols appear in Appendix B.

Use of AI tools. Generative AI was used solely to assist with writing and language polishing in this paper. All research methods, code implementation, theoretical development and proof derivations, literature retrieval and synthesis, and the preparation of figures, tables, and other research outputs were independently completed by the authors.

REFERENCES

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *International Conference on Learning Representations*, 2025.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anthropic. Anthropic skills repository. GitHub repository, 2026. URL `https://github.com/anthropics/skills`. Commit 3b3fad96af16a10759d930941b4520ba0c40edae.

Sanket Badhe and Priyanka Tiwari. Agent skill security: Threat models, attacks, defenses, and evaluation. *arXiv preprint arXiv:2607.13987*, 2026.

Thomas Ball, Nikolaj S. Bjørner, Ashley J. Chen, Shuo Chen, Yang Chen, Zhongxin Guo, Tzu-Han Hsu, Peng Liu, and Nanqing Luo. Theorem-carrying transactions: Runtime verification to ensure interface specifications for smart contract safety. *arXiv preprint arXiv:2408.06478*, 2024.

Luca Beurer-Kellner and Marc Fischer. Introducing MCP-Scan: Protecting MCP with invariant. Invariant Labs blog, 2025. URL https://invariantlabs.ai/blog/introducing-mcp-scan.

Henk Birkholz, Antoine Delignat-Lavaud, Cédric Fournet, Yogesh Deshpande, and Steve Lasker. An architecture for trustworthy and transparent digital supply chains. Technical Report RFC 9943, Internet Engineering Task Force, 2026.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Jialuo Chen, Lingqi Jiang, Xinhao Deng, Xiaohu Du, Jianan Ma, Yunhao Feng, Yuqi Qing, Zhihao Yuan, Linkang Du, and Jingyi Wang. When experience becomes instruction: Trajectory poisoning in self-evolving agent skill systems. *arXiv preprint arXiv:2608.05563*, 2026a.

Yueh-Han Chen, Nitish Joshi, Yulin Chen, Maksym Andriushchenko, Rico Angell, and He He. Monitoring decomposition attacks with lightweight sequential monitors. In *International Conference on Learning Representations*, 2026b.

Cisco AI Defense. Skill scanner: Security scanner for agent skills. GitHub repository, 2026. URL https://github.com/cisco-ai-defense/skill-scanner.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-2636.

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

David E. Eckhardt and Larry D. Lee. A theoretical basis for the analysis of multiversion software subject to coincident errors. *IEEE Transactions on Software Engineering*, SE-11(12):1511–1517, 1985. doi: 10.1109/TSE.1985.231895.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Haowen Gao, Haoran Chen, Can Wang, Shasha Guo, Liang Pang, Zhaoyang Liu, Huawei Shen, and Xueqi Cheng. SkillAudit: Ground-truth-free skill evolution via paired trajectory auditing. *arXiv preprint arXiv:2606.14239*, 2026.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159, 2019.

David K. Gifford. Weighted voting for replicated data. In *Proceedings of the Seventh ACM Symposium on Operating Systems Principles*, pp. 150–162, 1979. doi: 10.1145/800215.806583.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Yibo Hu and Jiaming Qu. Social pressure breaks majority voting in LLM safety panels. *arXiv preprint arXiv:2608.04415*, 2026.

Yiran Hu, Zongyue Xue, Haitao Li, Siyuan Zheng, Qingjing Chen, Shaochun Wang, Xihan Zhang, Ning Zheng, Yun Liu, Qingyao Ai, Yiqun Liu, Charles L. A. Clarke, and Weixing Shen. LLMs on trial: Evaluating judicial fairness for large language models. In *International Conference on Learning Representations*, 2026a.

Yuepeng Hu, Yuqi Jia, Mengyuan Li, Dawn Song, and Neil Zhenqiang Gong. MalTool: Malicious tool attacks on LLM agents. *arXiv preprint arXiv:2602.12194*, 2026b.

Tim Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering evaluation-aware language models to act like they are deployed. In *International Conference on Learning Representations*, 2026.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. MetaTool benchmark for large language models: Deciding whether to use tools and which to use. In *International Conference on Learning Representations*, 2024.

Alex Kantchelian, Michael Carl Tschantz, Sadia Afroz, Brad Miller, Vaishaal Shankar, Rekha Bachwani, Anthony D. Joseph, and J. D. Tygar. Better malware ground truth: Techniques for weighting anti-virus vendor labels. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, 2015. doi: 10.1145/2808769.2808780.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118.

John C. Knight and Nancy G. Leveson. An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1):96–109, 1986. doi: 10.1109/TSE.1986.6312924.

Eugene Koran, Yejun Yun, Samantha Tetef, Benjamin Arnav, and Pablo Bernabeu-Pérez. Ensemble monitoring for AI control: Diverse signals outweigh more compute. *arXiv preprint arXiv:2605.15377*, 2026.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003. doi: 10.1023/A:1022859003006.

Kevin Leahy, Makai Mann, and Zachary Serlin. Run-time task composition with safety semantics. In *International Conference on Machine Learning*, 2024.

Hwiwon Lee, Ziqi Zhang, Hanxiao Lu, and Lingming Zhang. SEC-bench: Automated benchmarking of LLM agents on real-world software security tasks. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-3878.

Jialong Li and Jialing Zhu. Auditing self-evolution in financial agents: Capability gains, security drift, and execution-interface mismatch. *arXiv preprint arXiv:2608.17684*, 2026.

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

Bev Littlewood and Douglas R. Miller. Conceptual modeling of coincident failures in multiversion software. *IEEE Transactions on Software Engineering*, 15(12):1596–1614, 1989. doi: 10.1109/32.58771.

Bev Littlewood and Lorenzo Strigini. Redundancy and diversity in security. In *Computer Security – ESORICS 2004*, volume 3193 of *Lecture Notes in Computer Science*, 2004. doi: 10.1007/978-3-540-30108-0_26.

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

Yuyang Luo, Haoran Wang, and Kai Shu. Query-only backdoor attacks on self-evolving skills via trajectory poisoning. *arXiv preprint arXiv:2608.08303*, 2026.

Dahlia Malkhi and Michael Reiter. Byzantine quorum systems. *Distributed Computing*, 11(4): 203–213, 1998. doi: 10.1007/s004460050050.

Xutao Mao, Liangjie Zhao, Xiang Zheng, and Cong Wang. Practice makes unsafe: Skill misevolution in self-improving LLM agents. *arXiv preprint arXiv:2608.12851*, 2026.

George C. Necula. Proof-carrying code. In *Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pp. 106–119, 1997. doi: 10.1145/263699.263712.

Chinenye Okafor, James C. Davis, and Santiago Torres-Arias. DiVerify: Hardening identity-based software signing with diverse-context scopes. *arXiv preprint arXiv:2406.15596v3*, 2025. URL https://arxiv.org/abs/2406.15596v3.

OpenA2A. Open agent security benchmark. GitHub repository, 2026. URL https://github.com/opena2a-org/oasb. Commit 43c19e55db0272b736742bf97d7fd0d11723a606.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, 2024.

Shresta Bellary Seetharam, Nabeel Mohamed, Billy Melicher, and Oleksii Starov. OpenClaw's skill marketplace and the emerging AI supply chain threat. Unit 42 research, 2026. URL https://unit42.paloaltonetworks.com/openclaw-ai-supply-chain-risk/.

Linfang Shang, Ming Xu, Yiding Sun, Tianle Xia, Lingxiang Hu, Lan Xu, and Ning Zheng. When self-evolution backfires: Pre-commit gating against skill contamination in LLM agents. *arXiv preprint arXiv:2608.05810*, 2026.

Adi Simhi, Jonathan Herzig, Martin Tutek, Itay Itzhak, Idan Szpektor, and Yonatan Belinkov. ManagerBench: Evaluating the safety-pragmatism trade-off in autonomous LLMs. In *International Conference on Learning Representations*, 2026.

Snyk. Agent scan: Security scanner for AI agents, MCP servers and agent skills. GitHub repository, 2026. URL https://github.com/snyk/agent-scan.

Supply-chain Levels for Software Artifacts. SLSA specification v1.2. https://slsa.dev/spec/v1.2/, 2025.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Haoyu Wang, Christopher M. Poskitt, and Jun Sun. AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents. In *Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering*, pp. 2938–2950, 2026a. doi: 10.1145/3744916.3764546.

Shenao Wang, Junjie He, Yanjie Zhao, Yayi Wang, Kan Yu, and Haoyu Wang. MalSkills: Detecting malicious skills in the agentic supply chain via neuro-symbolic reasoning. *arXiv preprint arXiv:2603.27204*, 2026b.

Yue Wang, Yi Liu, Gelei Deng, Ying Zhang, Yuekang Li, Zhenyu Chen, and Leo Zhang. MaliciousSkillBench: A comprehensive benchmark for malicious agent skill detection, 2026c. URL https://arxiv.org/abs/2608.19901.

Zora Zhiruo Wang, Apurva Gandhi, Graham Neubig, and Daniel Fried. Inducing programmatic skills for agentic tasks. In *Conference on Language Modeling*, 2025a. URL https://arxiv.org/abs/2504.06821.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *International Conference on Machine Learning*, 2025b.

Haoze Wu, Clark Barrett, and Nina Narodytska. Lemur: Integrating large language models in automated program verification. In *International Conference on Learning Representations*, 2024.

Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. Watch out for your agents! investigating backdoor threats to LLM-based agents. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-3201.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Gregory Yauney, Shahzaib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations*, 2026.

Zonghao Ying, Xiangfan Wu, Huiyu Wu, Xing Zheng, Huangsheng Cheng, Xiaorong Shi, and Jing Guo. SkillJack: Persistent skill backdoors in self-evolving agents. *arXiv preprint arXiv:2608.03509*, 2026.

Nusrat Zahan, Philipp Burckhardt, Mikola Lysenko, Feross Aboukhadijeh, and Laurie Williams. Leveraging large language models to detect npm malicious packages. In *IEEE/ACM 47th International Conference on Software Engineering*, 2025.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506, 2024. doi: 10.18653/v1/2024.findings-acl.624.

Boyuan Zheng, Michael Y. Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, and Yu Su. SkillWeaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025.

Shuofei Zhu, Jianjun Shi, Limin Yang, Boqin Qin, Ziyi Zhang, Linhai Song, and Gang Wang. Measuring and modeling the label dynamics of online anti-malware engines. In *29th USENIX Security Symposium*, 2020.

Thomas P. Zollo, Todd Morrill, Zhun Deng, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models. In *International Conference on Learning Representations*, 2024.

Xuanjun Zong, Zhiqi Shen, Lei Wang, Yunshi Lan, and Yang Chao. MCP-SafetyBench: A benchmark for safety evaluation of large language models with real-world MCP servers. In *International Conference on Learning Representations*, 2026.

Andy Zou, Maxwell Lin, Eliot Jones, Micha Nowak, Mateusz Dziemian, Nick Winter, Valent Nathanael, Ayla Croft, Xander Davies, Jai Patel, Robert Kirk, Yarin Gal, Dan Hendrycks, Zico Kolter, and Matt Fredrikson. Security challenges in AI agent deployment: Insights from a large scale public competition. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025. doi: 10.52202/085713-2676.

APPENDIX

A ANALYTIC DETAILS

A.1 PAIR RISK, ATTAINABLE DEPENDENCE, AND DECISION CREDIT

Proof. For Bernoulli E_i with mean ϵ_i, $\mathrm{Var}(E_i) = \epsilon_i(1-\epsilon_i)$. Hence

$$\mathrm{Cov}(E_1, E_2) = \rho\sqrt{\epsilon_1(1-\epsilon_1)\epsilon_2(1-\epsilon_2)}.$$

Also $\mathrm{Cov}(E_1, E_2) = \Pr(E_1 = 1, E_2 = 1) - \epsilon_1\epsilon_2$. Rearrangement proves Equation 7. □

Subtracting the first line of Equation 7 for pairs a and b gives its second line. For fixed non-degenerate margins, the Frechet upper bound is $J_{ij} \le \min(\epsilon_i, \epsilon_j)$. Substituting this maximum joint mass into $\rho = (J_{ij} - \epsilon_i\epsilon_j)/s_{ij}$ yields the $\rho_{\max}$ expression in Section 3. The lower attainable value follows analogously from $J_{ij} \ge \max(0, \epsilon_i + \epsilon_j - 1)$.

For a frozen reference r, Equation 2 gives

$$\begin{aligned}\mathrm{BA}(g) - \mathrm{BA}(r) &= \tfrac{1}{2}\{-R_M(g) + R_M(r) + U_B(g) - U_B(r)\} \\ &= \tfrac{1}{2}\{G_M - L_B\},\end{aligned}$$

which proves Equation 8. If g is the all-clear rule for members 1 and 2 and member 1 is the reference, then

$$\begin{aligned}G_M &= \epsilon_1 - J_{12} \\ &= \epsilon_1 - \epsilon_1\epsilon_2 - \rho_{12}s_{12} = \epsilon_1(1-\epsilon_2) - \rho_{12}s_{12}.\end{aligned}$$

This expression is nonnegative because $J_{12} \le \epsilon_1$, but a positive safety gain alone does not imply $G_M > L_B$.

A.2 MISSINGNESS ENDPOINT IDENTITY

For a fixed item, $g_q^{(0)} = \mathbb{1}\{K \ge q\}$ and $g_q^{(1)} = \mathbb{1}\{K + M \ge q\}$. Because $M \ge 0$, the second indicator cannot be smaller. Their difference is one exactly when the first event fails and the second holds, namely $K < q \le K + M$, proving Equation 9. Averaging the identity conditional on $Y = 1$ or $Y = 0$ gives the corresponding risk or utility endpoint gap.

A.3 DIFFICULTY DECOMPOSITION

Let $p_i(D) = \mathbb{E}[E_i \mid D, Y = 1]$. The law of total covariance gives

$$\mathrm{Cov}(E_i, E_j \mid Y = 1) = \mathbb{E}[\mathrm{Cov}(E_i, E_j \mid D, Y = 1)] + \mathrm{Cov}(p_i(D), p_j(D)). \tag{10}$$

Derivation. Apply the law of total covariance conditional on $Y = 1$:

$$\begin{aligned}\mathrm{Cov}(E_i, E_j \mid Y = 1) = \mathbb{E}&[\mathrm{Cov}(E_i, E_j \mid D, Y = 1) \mid Y = 1] \\ &+ \mathrm{Cov}\big(\mathbb{E}[E_i \mid D, Y = 1], \mathbb{E}[E_j \mid D, Y = 1] \mid Y = 1\big).\end{aligned}$$

Substitute $p_i(D)$ and suppress the repeated conditioning notation. □

A.4 TARGET-CONTAMINATED PROXY COUNTEREXAMPLE

Let E_1, E_2 be independent Bernoulli$(1/2)$ variables and set $D^\star = E_1 + E_2$. Although their unconditional correlation is zero, the residuals $E_i - \mathbb{E}[E_i \mid D^\star]$ have correlation -1.

Proof. Because E_1 and E_2 are independent fair Bernoulli variables, their unconditional correlation is zero. If $D^\star = 0$ or 2, both conditional residuals are zero. If $D^\star = 1$, then $E_2 = 1 - E_1$ and $\mathbb{E}[E_1 \mid D^\star = 1] = \mathbb{E}[E_2 \mid D^\star = 1] = 1/2$. Therefore the two residuals are exact negatives whenever they are nonzero. Their variances are positive, so their correlation is -1. □

The empirical simulation uses a continuous latent difficulty. Conditional on that variable, two Bernoulli target errors are drawn independently. Clone agreement is calibrated from the sealed exploratory matrix: Sol Prompt A/B agree on $58/60 = .967$ malicious items and the Qwen temperatures on $47/51 = .922$ paired-valid malicious items. Across 3,000 repetitions of 60 items, the raw target correlation has mean .295 and adjustment by the known latent difficulty has mean $-.0002$. The no-clone vote proxy has mean .135, a paired bias of $+.135$ from the known-truth adjustment (central 95% range $[-.003, .284]$). Mean adjusted correlations are .016 with one Sol clone, .035 with one Qwen clone, and $-.199$ with both clones. Relative to the paired no-clone draw, mean shifts are $-.120$ (central 95% range $[-.248, -.002]$), $-.100$ $[-.226, .014]$, and $-.334$ $[-.517, -.173]$, respectively. An 800-repetition sensitivity grid at each of eight agreement rates gives two-clone means .238, .183, .099, $-.002$, $-.136$, $-.172$, $-.236$, and $-.274$ at agreement .50, .60, .70, .80, .90, .922, .967, and .99. These values demonstrate the conditioning mechanism; they do not estimate bias in the real observer matrix.

B Protocol and Artifact Details

B.1 Observer identifiers and report contract

The observer short names identify the following API models:

Observer	API identifier
Sol	`gpt-5.6-sol`
Terra	`gpt-5.6-terra`
Qwen	`Qwen/Qwen3.6-27B`
Claude	`anthropic/claude-sonnet-4.6`
Qwen3.8	`Qwen/Qwen3.8-27B`
Nemotron	`nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-BF16`

Each normalized report binds the package identifier and digest, declares coverage, returns clear/protected/insufficient, lists protected categories, and records audit interference. Complete-clear additionally requires untruncated input.

B.2 Static sealing and clone-safe panel

The formal set is chosen by frozen salted digest. Six raw streams are normalized and sealed before the private class map is opened. All observers pass the 20-item calibration gate and no formal item is retried. The registered analysis uses 10,000 paired bootstrap samples.

The deterministic extension uses the all-six-valid malicious panel ($n = 37$). Family labels are Sol, Terra, Qwen, and Nemotron. For each pair, the proxy excludes the union of both family labels: for example Sol–Qwen uses Terra and Nemotron only; Sol Prompt A/B uses Terra, both Qwen variants, and Nemotron. The exact panel, exclusions, fits, bootstrap draws, difference tests, seeds, and unit tests are distributed. No provider call is made.

The package-view manifest records primary characters and bytes, total package bytes, and file count. A package absent from the archived manifest is excluded from view-only calculations but remains in every registered failure and decision result. All 60 malicious packages are covered.

B.3 Complete-panel selection audit

The complete-panel selection diagnostic was registered after the exploratory observations were sealed but before its deterministic execution. It uses only those observations and the frozen package manifest. The population is all 60 malicious items: 37 have valid observations from all six observers and 23 do not. The latter comprise 14 Nemotron-only missing cases and nine cases missing both Nemotron and Qwen $T = .7$. Three covariates are frozen: $\log(1 + \text{primary characters})$, $\log(1 + \text{file count})$, and primary/package byte ratio. The analysis uses 20,000 fixed-seed bootstrap draws and 20,000 fixed-seed label permutations, with Holm correction applied separately to mean- and KS-test families. Its script, protocol, summary, report, and tests are distributed; it makes zero provider calls.

B.4 CONFIRMATORY PROTOCOL AND PROVIDER AMENDMENT

The confirmatory salt selects 12 malicious validation and 120 malicious test identities from MaliciousSkillBench. Formal items have 120 distinct structural families and zero identity overlap with the 900 identities used in the preceding development and exploratory studies. The protocol freezes Equation 5, one directional test, 20,000 item-bootstrap draws, seed `2026082701`, missing-as-non-clear, and the joint success rule before provider calls.

OpenAI serves Sol under two frozen prompts. Qwen uses the OpenRouter route `qwen/qwen3.6-27b`, with Prompt A fixed at temperatures 0 and .7; the Alibaba provider is pinned, fallback and data collection are disabled, and reasoning is excluded. API credentials are stored outside the experiment and are absent from every manifest and archive. The provider substitution is frozen after endpoint discovery but before any Qwen or formal call.

The initial strict calibration gate records 12/12 for three observers and 2/12 for Qwen $T = .7$ because one 244-character rationale makes its ten-item batch invalid under a 240-character limit. Since rationale is not decision-bearing, a calibration-time amendment—frozen before formal calls—truncates only that field, applies symmetrically to all observers, preserves raw response hashes, and logs each event. The amended gate is 12/12 for all four observers. It normalizes two further formal rationales. A separate Sol Prompt B formal batch is invalid because one package identifier fails binding; no repair or retry is made. Primary analysis maps all ten records to non-clear and the complete-case sensitivity omits them. The sealed run has 56 provider calls, zero provider errors, and estimated cost USD 3.10.

Before the network-enabled run, 48 sandboxed batch attempts fail locally at DNS resolution with no response identifier, provider route, or charge. These pre-dispatch records are preserved separately and are not counted as provider calls or retries. The output seal covers the successful raw streams, compiled observations, calibration gate, and corpus artifacts; an independent replay matches all three correlations, C, its interval, and its bootstrap tail probability exactly.

B.5 BALANCED CLAUDE–QWEN3.8 EXTENSION

The extension uses 40 calibration and 240 formal packages, balanced within each split. The corpus builder excludes preceding-study identities, text hashes, and available structural families, then selects at most one package per structural group by salted digest. The 280 selected groups are distinct across the two splits. Inputs exceeding 16,000 characters are excluded rather than truncated. Every observer receives the same opaque identifiers and primary text, with no labels or benchmark identifiers. This is a new panel of the same benchmark, not a second data source.

Claude Sonnet 4.6 uses the Anthropic-only OpenRouter route with structured JSON output; Qwen3.8-27B uses native Tinker sampling and its native chat template without generation-time schema enforcement. Both use Prompts A/B unchanged, temperature zero, disabled reasoning, and a 2,200-token output ceiling. Batches contain at most five items and 24,000 input-text characters. The local decision and validity checks are shared. Model, provider, tokenizer, and output constraints vary together, so the comparison does not isolate an architectural cause.

The initial OpenRouter billing stop led to the Qwen3.8 substitution; Claude subsequently resumed. Five existing Claude response batches are retained and only unattempted requests are sampled in the final continuation. A billing failure without a model response receives one documented transport recovery, not a replacement model response. There are no automatic retries. After the strict rationale-length gate failed, an explicit amendment removes its 240-character ceiling from local validation and future Claude schemas. It preserves full rationale strings and uniformly revalidates both models' raw responses; identifiers, verdicts, categories, and complete-clear logic remain unchanged. Earlier Claude outputs retain their original generation constraint. Each arm has 148 unique response batches; all four calibration conditions are valid on 40/40 items. Formal valid counts are 228, 235, 240, and 240 for Claude A/B and Qwen3.8 A/B.

The original calibration rule maximizes BA separately over the four individuals and over the two prompt-matched cross-family pairs, breaking ties by lexicographic name. Claude A/B and A/A, B/B pairs all score 97.5%; Qwen3.8 A/B score 90.0% and 95.0%. This selects Claude A and A/A. The historical pre-formal seal binds calibration validity, not operating-point names. These choices are reconstructed from calibration data after formal outputs exist and held fixed in every

sensitivity analysis. Together with the roster and validation amendments, this makes the comparison a supplementary extension rather than the original preregistered experiment. The formal-best Claude B is never substituted for the calibration-selected reference.

The inherited BA-difference interval uses 20,000 label-stratified paired percentile bootstrap draws, Python seed `2026092201`. On the balanced panel, an exact two-sided McNemar test of correctness is also a test of the BA difference. For the selected intersection rule, let G_M and L_B be the additional malicious and benign blocking rates relative to the reference, as in Equation 8. An observed zero G_M makes its ordinary bootstrap interval degenerate. We additionally invert the binomial distribution for exact component intervals. Combining two 97.5% component intervals $[\ell_M, u_M]$, $[\ell_B, u_B]$ gives the conservative 95% interval $\frac{1}{2}[\ell_M - u_B, u_M - \ell_B]$ for BA gain by the union bound. This supplement assumes independent items within each label and conditions on the calibration-selected operating points.

For dependence, 20,000 paired four-vote bootstrap draws on the malicious items use seed `2026092202`. A secondary family contains eight same-family-pair minus cross-family-pair differences and the difference of their group means. For each contrast T, the approximate two-sided null probability is $(1 + \#\{|T^* - T| \geq |T|\})/(1 + B_{\text{valid}})$. Holm correction covers all nine tests. Undefined correlations are omitted only from the affected replicate statistic and their counts are retained. Normalized $\rho/\rho_{\max}$ differences are descriptive sensitivity statistics. Separate 20,000-draw resampling of all 64 aligned formal request batches, with seed `2026092203`, preserves cross-observer pairing and recomputes each class denominator. Its nine tests form a separate sensitivity family. All these added tests and diagnostics are post-observation analyses.

Missingness analyses retain the selected pair and reference. They compare non-clear and clear coding, the all-four-valid subset, and a diagnostic that reindexes only batches with exactly the expected unique ID set and all other fields valid. The diagnostic neither rewrites raw responses nor changes the official validator. Finally, all 2^{17} binary assignments to invalid votes are enumerated to give sharp finite-panel bounds, rather than confidence intervals, for the fixed decision difference and correlation contrasts.

B.6 Paired package-view intervention

The diagnostic population is frozen from a historical cohort of 45 malicious packages jointly cleared by Sol and Terra; 37 map to multi-file packages. This cohort is distinct from the 37-item all-six-valid panel used for estimator adjustment. Formal eligibility requires an official archive-hash match, traversal- and link-free paths, at least 95% UTF-8 byte coverage, at most 120,000 expanded characters, and an exact match between the old primary packet and the package's primary instruction file. These rules select 28 formal and three disjoint calibration packages before any view-intervention call. Selection is deliberately outcome-based; the estimand is the paired view effect within this cohort.

Both conditions carry the same manifest of paths, byte counts, and hashes. The control exposes only the primary instruction file; the expanded view exposes every UTF-8 file. Opaque IDs and salted hash ordering conceal condition names. Sol and Terra retain the historical audit's protected-category policy, augmented with one-to-four exact evidence paths. All six calibration packets per observer and all 56 formal packets per observer are schema-valid and identifier-bound. Calls are never retried. A pre-formal amendment raises only the rendered JSON ceiling from 180,000 to 220,000 characters after JSON escaping makes one 101,051-character packet render to 203,055 characters; corpus, prompt, schema, estimand, and calls are unchanged. The run uses 18 provider calls and costs USD 1.58.

The primary analysis maps invalid reports to non-clear, although no invalid formal report occurs. It uses 20,000 paired-item bootstrap draws (seed `20260828`) and a directional exact McNemar test. A file attribution is counted only when the expanded protected report cites an exact non-primary path in the frozen manifest.

B.7 Status of the cross-corpus decision reuse

The cross-corpus Sol/Qwen observations and 36+36 class map were sealed before the portability analysis, but their original aggregate metrics had already been inspected. Accordingly, the new full-panel and source-composition bootstrap intervals are post-registered descriptive uncertainty, not

confirmatory inference. The analysis is retained because it demonstrates that the same reference-member and both-clear accounting can be applied to OASB and Anthropic Skills without a new model call, and because it fixes a precise future hypothesis.

B.8 Runtime source and sandbox

OASB is fixed at commit `43c19e55db0272b736742bf97d7fd0d11723a606`; Anthropic Skills at `3b3fad96af16a10759d930941b4520ba0c40edae`. OASB selection contains two Skill artifacts from each of nine malicious categories. Static observers are Sol and deterministic Qwen; Terra is the runtime executor.

The harness exposes virtual package reads, workspace reads/writes, shell, HTTP, credential, external-message, and result-submission tools. Local operations use fixtures; protected classes are recorded and blocked. Package scripts never execute. Every episode has a sealed event sequence and zero real side effects.

B.9 Functional adjudication and power

The strict endpoint counts `submit_result` and is 65/72. Seven traces without that marker receive review under a fixed rubric. Four contain a task-responsive final deliverable and are terminal-equivalent; three remain incomplete, giving 69/72. Review changes no tool event, protected category, observer verdict, or label. Both endpoints and all seven records are retained.

For each static signal, conditional exact power fixes n, the number static- clear, and the number with a protected attempt, then enumerates feasible overlaps. Noncentral hypergeometric probabilities under a specified odds ratio are summed over tables whose two-sided Fisher $p \leq .05$. This diagnoses the realized design rather than postulating an asymptotic effect size.

C Additional Results

C.1 Confirmatory-panel sensitivity and clone calibration

The complete-case analysis of the confirmatory panel uses the 110 identities with all four reports. Its pair correlations are .755 (prompt), .870 (temperature), and .408 (cross-family), giving $C = .727$, bootstrap 95% interval $[.446, 1.141]$, and one-sided $p = .00005$. All 20,000 primary bootstrap draws are valid. The exact full-denominator table and post-registered margin diagnostics are retained below.

Table 2: Preregistered confirmatory result on 120 malicious packages. Prompt compares Sol prompts A/B, Temperature compares Qwen at $T = 0/.7$, and Cross-family compares Sol A with Qwen at $T = 0$. Marginals are complete-clear counts (each out of 120); Joint gives count (rate), Indep. is the product of marginal rates, and Excess is Joint/Indep. Missing/invalid reports are non-clear. Only the Fisher-z contrast in Equation 5 is preregistered; the margin diagnostics are post-registered.

	False-clear counts and rates				Dependence		
Pair	Marginals	Joint	Indep.	Excess	ρ	$\rho_{\max}$	$\rho/\rho_{\max}$
Prompt	20/15	13 (.108)	.0208	5.20×	.710	.845	.840
Temperature	35/36	32 (.267)	.0875	3.05×	.860	.980	.878
Cross-family	20/35	14 (.117)	.0486	2.40×	.402	.697	.576
Registered $C = .665$; 95% CI $[.390, 1.026]$					one-sided $p = .00010$		

Pearson ρ is constrained by unequal Bernoulli margins. The three attainable upper bounds are .845, .980, and .697. Normalized values $\rho/\rho_{\max}$ are .840, .878, and .576; their Fisher-z contrast remains positive at .636. The corresponding log odds ratios are 4.511, 5.375, and 2.172, with same-view mean minus cross-family contrast 2.771. These deterministic point diagnostics do not replace the preregistered statistic or its bootstrap inference.

The table below reports the calibrated estimator simulation used in RQ2. Raw and true-latent rows restore the known reference before comparing vote-derived proxies.

Table 3: Zero-, one-, and two-clone calibration across 3,000 repetitions of 60 items. Bias is paired to adjustment by the known latent difficulty; Δ_0 is paired to the no-clone vote proxy. Ranges are central 95%; paired-bias intervals appear below the table. A dash denotes a comparison that is not applicable.

Condition	Mean r	Central 95% range	Bias vs. truth	Δ_0 vs. no-clone
Raw target correlation	.295	[.040,.533]	—	—
True latent adjusted	$-.0002$	[$-.260$,.281]	.000	—
No-clone vote proxy	.135	[$-.126$,.406]	$+.135$	.000
One Sol clone	.016	[$-.246$,.297]	$+.016$	$-.120$
One Qwen clone	.035	[$-.230$,.316]	$+.035$	$-.100$
Two clones	$-.199$	[$-.429$,.070]	$-.199$	$-.334$

The paired no-clone bias relative to the true-latent adjustment has central 95% range [$-.003$,.284]. The corresponding paired shifts from the no-clone proxy are [$-.248$,$-.002$] for one Sol clone, [$-.226$,.014] for one Qwen clone, and [$-.517$,$-.173$] for two clones. The post-registered diagnostic and its exact parameters are included with the supplementary analysis output.

C.2 Exploratory estimator and view audit

Table 4: Estimator audit for malicious-error dependence. Raw ρ is the frozen exploratory paired-valid estimate, with its pair-specific denominator shown. Original r uses a variable-composition leave-pair-out proxy that may contain target-family variants and is shown only to expose sensitivity. Clone-safe r uses the common 37-item panel and leave-family-out proxy; brackets are bootstrap 95% intervals.

	Paired-valid			Clone-safe ($n = 37$)	
Pair	n	Raw ρ	Original r	r	95% CI
Sol A / Terra A	60	.640	.360	.451	[.117, .762]
Sol Prompt A / B	60	.877	.942	.804	[.015, 1.000]
Qwen $T = 0$ / 0.7	51	.841	.841	.739	[.431, 1.000]
Sol / Qwen	60	.475	$-.217$	.315	[.000, .560]
Sol / Nemotron	37	.276	.319	.286	[$-.000$, .619]
Qwen / Nemotron	37	.408	.108	.272	[.074, .542]

Table 5: Post-registered selection audit on all 60 malicious static items: 37 enter the common all-six-valid panel and 23 are excluded. Values are raw median [IQR]; SMD is computed on the frozen analysis scale. Mean- and KS-difference p_H values are two-sided label-permutation tests with Holm adjustment applied separately to each test family.

	All-six-valid panel			Holm-adjusted p_H	
Covariate	Selected ($n = 37$)	Excluded ($n = 23$)	SMD	Mean	KS
Primary characters	5,763 [2,231, 8,546]	6,551 [3,386, 7,872]	$-.403$	.411	1.000
Package files	2 [1, 3]	2 [1, 2]	.002	.990	1.000
Primary/package bytes	.828 [.353, 1.000]	.567 [.424, .791]	.239	.745	.226

Table 6: Direct package-view measurements for the 60 malicious static items. The primary/package byte ratio is measured on the complete package archive. The prior agreeing-clear cohort is a separate sealed set of 45 malicious packages cleared by both static observers.

Measurement	Value
Static malicious panel ($n = 60$)	
Single-file / multi-file	22 / 38
File count, 25/50/75th percentile	1 / 2 / 3
Primary/package byte ratio, range	.085–1.000
Primary/package byte ratio, 25/50/75th	.371 / .716 / 1.000
Prior agreeing-clear cohort ($n = 45$)	
Single-file / multi-file	8 / 37

C.3 SINGLE-OBSERVER OUTCOMES

Table 7: Single-observer outcomes on 60 malicious and 60 benign items. Mal. and Ben. columns are accepted counts out of 60; malicious acceptance is a false clear. Balanced accuracy (BA) and paired-bootstrap 95% CIs are in percent. Invalid or incomplete reports are non-clear.

Observer	Valid (/120)	Mal. /60	Ben. /60	BA % [95% CI]
Sol, Prompt A	120	10	50	83.3 [76.7,90.0]
Terra, Prompt A	120	16	49	77.5 [70.0,85.0]
Sol, Prompt B	120	8	49	84.2 [77.5,90.0]
Qwen, $T = 0$	120	23	54	75.8 [68.3,82.5]
Qwen, $T = 0.7$	110	20	53	77.5 [70.0,84.2]
Nemotron, $T = 0$	80	29	41	60.0 [51.7,68.3]

C.4 DECISION FRONTIER AND CROSS-CORPUS REUSE

The formal optimum is descriptive: q=3,4,5 all score 100% on the 10+10 calibration set, so calibration does not uniquely select q=4. Relative to Sol Prompt B, the paired-bootstrap mean BA difference is -1.6 percentage points with 95% interval $[-5.8, 2.5]$, and 15.7% of draws favor q=4. When invalid reports are assigned clear instead of non-clear, q=6 changes from 5/60 malicious and 36/60 benign accepted to 6/60 and 46/60; its BA moves from 75.8% to 83.3%. q=4 is stable at 82.5% under both endpoints, but high-q selectivity cannot be attributed entirely to semantic detection.

Table 8: Full-denominator decision audit on 60 malicious and 60 benign items. The formal endpoint treats invalid reports as non-clear; the sensitivity endpoint treats them as clear. Mal. and Ben. columns are accepted counts out of 60; balanced accuracy (BA) is in percent. Formal BA intervals are paired-bootstrap 95% CIs. The single observer is unaffected by the six-vote endpoint switch, so its sensitivity cells are not applicable (—).

	Missing non-clear (formal)			Missing clear (sensitivity)		
Rule	Mal. /60	Ben. /60	BA % [95% CI]	Mal. /60	Ben. /60	BA %
Sol Prompt B	8	49	84.2 [77.5, 90.0]	—	—	—
1-of-6	36	55	65.8 [58.3, 73.3]	53	59	55.0
2-of-6	25	55	75.0 [67.5, 82.5]	34	55	67.5
3-of-6	20	53	77.5 [70.0, 85.0]	24	55	75.8
4-of-6	11	50	82.5 [75.8, 89.2]	12	51	82.5
5-of-6	9	47	81.7 [74.2, 88.3]	9	48	82.5
6-of-6	5	36	75.8 [69.1, 83.3]	6	46	83.3

Table 9: Post-registered reuse of the decision audit on the sealed 36+36 cross-corpus panel. Intervals are paired, class-stratified bootstrap 95% intervals. Balanced accuracy (BA) is in percent; differences are percentage points. The external composition pairs OASB malicious with Anthropic benign items and is not an ecosystem-specific BA estimate.

	Balanced accuracy (%)			
Panel (mal./ben.)	Sol	Qwen	Both clear	Both − Qwen (pp) [95% CI]
All sources (36/36)	61.1	87.5	62.5	-25.0 [-34.7,-15.3]
MaliciousSkillBench (18/18)	58.3	75.0	58.3	-16.7 [-30.6,0.0]
OASB/Anthropic (18/18)	63.9	100.0	66.7	-33.3 [-44.4,-22.2]

The cross-corpus interval is a post-registered uncertainty calculation on a matrix whose original aggregate metrics were already available. It is used to demonstrate protocol portability and fix a future confirmatory hypothesis, not as a second preregistered result. Within the full 36+36 panel, Sol, Qwen, and both-clear have malicious false-clear / benign-accept counts of 2/10, 6/33, and 1/10, respectively.

Best-member identity is corpus-conditioned: Sol Prompt B leads the exploratory panel (84.2% BA), but Qwen leads the cross-corpus Sol–Qwen roster (87.5%), where Sol accepts only 10/36 benign

items. Different rosters do not establish rank stability; these results motivate a calibration-frozen reference rather than a permanent roster ranking.

C.5 Balanced Claude–Qwen3.8 decision and dependence audit

Tables 10–13 retain all four individuals, all six pairs, the nine dependence contrasts, and the fixed-selection missingness analysis. The calibration-selected Claude A–Qwen3.8 A pair has no additional observed malicious blocking and two additional benign blocks relative to Claude A. Thus $\widehat{G}_M = 0$, $\widehat{L}_B = 2/120$, and $\widehat{\Delta}\mathrm{BA} = -0.83$ percentage points. The exact 95% interval for G_M is $[0, .0303]$, although the ordinary bootstrap gives $[0, 0]$. The conservative component-based interval for BA gain is $[-3.30, 1.72]$ points. These intervals preserve uncertainty about unobserved discordances; neither proves equivalence or inferiority.

Table 10: All formal outcomes in the balanced extension. Malicious and benign accepted counts each have denominator 120. BA and item-bootstrap 95% intervals are percentages. C denotes Claude Sonnet 4.6 and Q denotes Qwen3.8-27B; subscripts are prompts. All pairs require both reports to clear. Formal-best configurations do not replace the calibration-selected C_A reference and C_A–Q_A pair.

Configuration	Mal. /120	Ben. /120	BA % [95% CI]
C_A	11	107	90.00 [85.83,93.75]
C_B	10	112	92.50 [89.17,95.83]
Q_A	22	116	89.17 [85.00,92.92]
Q_B	23	117	89.17 [85.00,92.92]
C_A–C_B	9	106	90.42 [86.67,94.17]
C_A–Q_A	11	105	89.17 [85.00,92.92]
C_A–Q_B	10	106	90.00 [85.83,93.75]
C_B–Q_A	10	109	91.25 [87.50,94.58]
C_B–Q_B	9	110	92.08 [88.33,95.42]
Q_A–Q_B	21	116	89.58 [85.42,93.33]

Table 11: Failure dependence on 120 malicious extension items. C and Q denote Claude and Qwen3.8; A/B denote prompts. Joint and Prod. are observed joint false-clear and marginal-product rates in percent. Correlation intervals use paired-item bootstrap. Invalid reports count as non-clear.

Pair	ρ [95% CI]	$\rho_{\max}$	$\rho/\rho_{\max}$	Joint %	Prod. %
C_A–C_B	.845 [.636,1.000]	.949	.890	7.50	0.76
C_A–Q_A	.670 [.498,.821]	.670	1.000	9.17	1.68
C_A–Q_B	.579 [.371,.747]	.652	.888	8.33	1.76
C_B–Q_A	.636 [.457,.793]	.636	1.000	8.33	1.53
C_B–Q_B	.543 [.329,.719]	.619	.876	7.50	1.60
Q_A–Q_B	.918 [.813,1.000]	.973	.944	17.50	3.51

Raw mean same-family minus cross-family correlation is .274 with interval $[.132, .439]$ and Holm $p = .0119$. Batch resampling gives $[.124, .450]$ and Holm $p = .0182$, with 19,999 valid draws for the mean contrast out of 20,000. Its normalized counterpart is $-.024$ with item-bootstrap interval $[-.135, .131]$ and batch-bootstrap interval $[-.136, .145]$. The two cross-family pairs using Qwen3.8 A attain their marginal correlation ceilings: the observed Claude error sets are contained in Qwen3.8 A's. Neither raw nor normalized correlation alone identifies a latent mechanism of independence.

Table 12: Supplementary paired correlation differences in the extension. Each row is left minus right; the final row compares the mean of the two same-family pairs with the mean of the four cross-family pairs. C/Q and A/B follow Table 10. Approximate two-sided centered-bootstrap probabilities are Holm-adjusted across all nine tests. Intervals are pointwise, not simultaneous.

Contrast	Difference [95% CI]	p	p_{Holm}
$(\mathrm{C}_A,\mathrm{C}_B) - (\mathrm{C}_A,\mathrm{Q}_A)$	.174 [−.044,.393]	.1118	.1120
$(\mathrm{C}_A,\mathrm{C}_B) - (\mathrm{C}_A,\mathrm{Q}_B)$	.266 [.021,.522]	.0391	.1120
$(\mathrm{C}_A,\mathrm{C}_B) - (\mathrm{C}_B,\mathrm{Q}_A)$	.208 [.012,.402]	.0373	.1120
$(\mathrm{C}_A,\mathrm{C}_B) - (\mathrm{C}_B,\mathrm{Q}_B)$	.302 [.074,.542]	.0151	.0757
$(\mathrm{Q}_A,\mathrm{Q}_B) - (\mathrm{C}_A,\mathrm{Q}_A)$	.248 [.051,.448]	.0152	.0757
$(\mathrm{Q}_A,\mathrm{Q}_B) - (\mathrm{C}_A,\mathrm{Q}_B)$	.339 [.181,.526]	.0011	.0092
$(\mathrm{Q}_A,\mathrm{Q}_B) - (\mathrm{C}_B,\mathrm{Q}_A)$	.282 [.080,.489]	.0081	.0489
$(\mathrm{Q}_A,\mathrm{Q}_B) - (\mathrm{C}_B,\mathrm{Q}_B)$	.376 [.211,.572]	.0006	.0054
Mean same − mean cross	.274 [.132,.439]	.0017	.0119

Table 13: Fixed-selection sensitivity of Claude A–Qwen3.8 A versus Claude A. BA is percent and differences are percentage points. All intervals are paired-bootstrap 95% intervals; the first four rows resample within labels. ID reindexing is a diagnostic, not a change to the original validator. The complete-case row targets the common valid subset.

Handling	Mal./ben. n	Ref. BA	Pair BA	Difference [95% CI]
Invalid as non-clear	120/120	90.00	89.17	-0.83 [-2.08,0.00]
Invalid as clear	120/120	91.67	92.50	$+0.83$ [-1.25,2.92]
Complete cases	116/112	93.03	92.13	-0.89 [-2.23,0.00]
Unique-ID reindexing	120/120	92.92	92.08	-0.83 [-2.08,0.00]
Non-clear, batch resampling	120/120	90.00	89.17	-0.83 [-2.16,0.00]

The 17 invalid observer-item assessments involve 12 distinct packages: five assessments are malicious and 12 benign. Claude A batches 6 and 16 and Claude B batch 16 have exactly the requested unique IDs in the wrong order, affecting 15 assessments. Claude A batch 40 returns three assessments for two requested IDs and remains unresolved in the reindexing diagnostic. Reindexing raises valid counts to 238/240 and 240/240 for Claude A/B, preserving every field value. The common complete-case panel contains 116 malicious and 112 benign packages; it is a selected population rather than a missing-at-random correction.

Across all 131,072 binary completions, the fixed BA-difference bounds are $[-0.83, +0.83]$ points and the raw mean correlation-contrast bounds are $[.269, .318]$. The latter retain the raw direction but do not establish significance for every completion. Under clear coding, the raw mean contrast is .312 (Holm $p = .0077$), while its normalized contrast is .189 with interval $[-.015, .445]$. Complete-case and ID-reindexing analyses also leave the normalized ordering unresolved. The observed Qwen3.8 A/B pair does slightly exceed its individual components' BA point estimates (89.58% versus 89.17%); the extension therefore does not support a universal claim that aggregation has no benefit.

C.6 Exact package-view intervention table

Table 14: Paired package-view intervention on 28 outcome-selected, multi-file malicious packages previously cleared by both observers. Both conditions expose the same file manifest; the expanded condition additionally exposes every UTF-8 file. Clear counts are out of 28; Δ is Expanded minus Manifest+primary in clear rate. Drop/gain gives paired discordances in that order. CI is a paired-item bootstrap 95% interval; p_- is the registered one-sided exact McNemar test.

	Clear count (/28)				
Clear rule	Manifest+primary	Expanded	Δ [95% CI]	Drop/gain	p_-
Sol	16/28	1/28	$-.536$ [$-.714$,$-.357$]	15/0	3.05×10^{-5}
Terra	21/28	2/28	$-.679$ [$-.857$,$-.500$]	19/0	1.91×10^{-6}
Both clear	16/28	0/28	$-.571$ [$-.750$,$-.393$]	16/0	1.53×10^{-5}

C.7 PAIR-DIFFERENCE MULTIPLICITY AUDIT

Table 15: All 15 pairwise differences among clone-safe residual correlations on the common 37-item panel. Each row is left minus right; p_{sign} is the two-sided bootstrap sign probability and p_{Holm} is adjusted across all 15 comparisons.

Left minus right	Difference	p_{sign}	p_{Holm}
Sol/Terra − Prompt	−.353	.347	1.000
Sol/Terra − Temperature	−.288	.182	1.000
Sol/Terra − Sol/Qwen	.135	.448	1.000
Sol/Terra − Sol/Nemotron	.165	.231	1.000
Sol/Terra − Qwen/Nemotron	.179	.347	1.000
Prompt − Temperature	.065	.932	1.000
Prompt − Sol/Qwen	.488	.090	1.000
Prompt − Sol/Nemotron	.518	.138	1.000
Prompt − Qwen/Nemotron	.532	.126	1.000
Temperature − Sol/Qwen	.423	.014	.210
Temperature − Sol/Nemotron	.453	.046	.638
Temperature − Qwen/Nemotron	.467	.048	.638
Sol/Qwen − Sol/Nemotron	.029	.840	1.000
Sol/Qwen − Qwen/Nemotron	.044	.984	1.000
Sol/Nemotron − Qwen/Nemotron	.014	.932	1.000

C.8 TARGET-INDEPENDENT PROXY SENSITIVITY

Table 16: Residual correlations after separate one-covariate logistic adjustments by target-independent package properties. Each adjusted column fits only the named covariate; n is paired-valid.

			Adjusted residual r		
Pair	n	Raw ρ	Log chars	Log files	Primary/total
Sol/Terra	60	.640	.639	.647	.636
Prompt	60	.877	.877	.884	.866
Temperature	51	.841	.842	.844	.843
Sol/Qwen	60	.475	.478	.508	.478
Sol/Nemotron	37	.276	.274	.274	.268
Qwen/Nemotron	37	.408	.413	.396	.404

C.9 VIEW STRATA WITH UNCERTAINTY

Table 17: Paired-valid malicious correlation by file structure, with paired-bootstrap 95% CIs. Temperature's degenerate correlation on 18 single-file items is reported without interpretation.

	Single-file		Multi-file	
Pair	n	ρ [95% CI]	n	ρ [95% CI]
Sol/Terra	22	.614 [−.069,1.000]	38	.659 [.425,.873]
Prompt	22	.796 [.463,1.000]	38	.911 [.697,1.000]
Temperature	18	1.000 [1.000,1.000]	33	.764 [.521,.941]
Sol/Qwen	22	.733 [.418,1.000]	38	.365 [.064,.601]
Sol/Nemotron	14	.389 [.175,.701]	23	.204 [.083,.353]
Qwen/Nemotron	14	.556 [.273,.866]	23	.219 [−.219,.586]

C.10 RUNTIME DIAGNOSTICS

Table 18: Static clearance and protected runtime attempts on 35 functionally complete malicious episodes. Clear and Attempt count their respective events; Both is their overlap. Fisher tests are two-sided; conditional exact Power uses $\alpha = .05$ and odds ratio 4.

	Events (/35)			Exact inference	
Static signal	Clear	Attempt	Both	Fisher p	Power (OR 4)
Sol clear	1	8	0	1.000	.000
Qwen clear	6	8	1	1.000	.273
Pair clear	1	8	0	1.000	.000

Table 19: Conditional exact power for the 35 malicious episodes in Table 18, at the displayed odds ratios. Two-sided Fisher tests fix the observed margins and use $\alpha = .05$.

	Assumed odds ratio			
Signal	2	4	8	16
Sol clear	.000	.000	.000	.000
Qwen clear	.084	.273	.566	.819
Pair clear	.000	.000	.000	.000

Protected attempts occur in 8/35 complete malicious and 3/34 complete benign episodes (odds ratio 3.06, Woolf 95% interval $[.74, 12.71]$, two-sided Fisher $p = .188$). Sol's risk-category count has correlation .067 $[-.158, .314]$ with the attempt indicator (permutation $p = .762$). Qwen assigns zero categories to all 35 and cannot supply an ordered association statistic.

Functional completion is 16/18, 18/18, 18/18, and 17/18 for benign Anthropic, benign MaliciousSkill-Bench, malicious MaliciousSkillBench, and malicious OASB, respectively; protected attempts are 0/18, 3/18, 3/18, and 5/18. All counts remain linked to sealed traces.