The Reliability Chasm: Enterprise LLM Agent Security Fails Across Runs, Domains, and Industries

Anonymous ACL submission

Abstract

Enterprise LLM agents run repeatedly in adversarial settings, but current evaluation methods report single-run mean scores that obscure per-scenario failure rates. We introduce CHASM, a 147-scenario adversarial benchmark spanning a 7×7 industry-by-domain grid (30 scenarios publicly released; held-out split for contamination control), and AFR (Agent Failure Rate), a deployment-aware metric that aggregates repeated-run outcomes into binary scenario verdicts. We show that all 9 frontier models fail a majority of adversarial enterprise scenarios, a gap we term the **Reliability Chasm**: mean AFR is 0.75, and even the best model (Claude Opus 4.6) fails on a majority of scenarios ($\mathrm{AFR} = 0.57$). Cochran's Q rejects equal per-model failure rates ($p < 10^{-15}$). Mean-score and AFR rankings diverge materially, with metric choice inverting deployment decisions for multiple models. Within our 9-model frontier cohort, we also observe that capability rank does not reliably predict safety rank. Rankings shift substantially across industries and work domains, with the same model sometimes ranking first in one industry and near-last in another. Models with similar failure rates differ sharply in failure severity, a risk dimension invisible to frequency alone. Our results show that secure enterprise agent deployment requires measuring failure consistency, severity, and context dependence, not just average performance.

1 Introduction

AI is shifting from monolithic generative models toward agentic systems capable of planning and interacting with tools. Standard NLP evaluation methods typically assess LLM agents with a single execution and report mean scores, yet deployed LLM agents execute the same tasks repeatedly under adversarial conditions, and a single security failure in any run constitutes deployment failure regardless of average-case performance. This disconnect between single-run benchmark methodology and multi-run operational risk is the core evaluation problem we address. As LLM agents enter enterprise settings with access to internal tools, sensitive data, and external integrations, two properties become critical: *security* (resistance to adversarial manipulation and to violations of policy, data, or access boundaries) and *reliability* (consistency of behavior across repeated executions). In regulated industries, enterprises cannot tolerate a critical security failure even once in 100,000 tasks; the bar for deployment is near-zero worst-case risk across the agent's operational lifetime.

Evaluating these properties requires more than single-run task success. Standard benchmarks such as GPQA (Rein et al., 2024) and HumanEval (Chen et al., 2021) and their modern counterparts (e.g., SWE-Bench (Jimenez et al., 2024), LiveCodeBench (Jain et al., 2025)) effectively measure peak capability, while specialized security suites (e.g., AgentDojo (Debenedetti et al., 2024), AgentHarm (Andriushchenko et al., 2025), ASB (Zhang et al., 2025)) test single-run attack success against specific threat types. Neither line of work has, to our knowledge, focused on how security behavior accumulates under repetition: enterprise deployment is stochastic, stateful, and open-ended, and a single security lapse across many runs constitutes deployment failure. To the best of our knowledge, no existing benchmark primarily measures how adversarial security failures compound across repeated executions. Security under repeated execution is distinct from peak capability; we call this gap between benchmark-measured performance and sustained real-world security the **Reliability Chasm**.

Threat model. We assume an enterprise agent [1] invoked repeatedly under stochastic conditions, with

[1] We use "enterprise agent" to denote LLM-based agents deployed in organizational settings with explicit security policies, tool access, and multi-turn task execution.

access to internal tools, sensitive data, and external integrations. The adversary may be *external* (untrusted user content, third-party data, document feeds) or *internal* (compromised tools, poisoned metadata, tampered data sources), and exploits the agent's tool surface rather than its weights or system prompt. The deployment-level failure object is the probability of *at least one* security violation across the agent's operational lifetime; a single such violation constitutes deployment failure, regardless of average-case behavior.

We introduce CHASM (CompreHensive Agentic Systems Manipulation): 147 adversarial enterprise scenarios spanning a 7×7 industry-by-domain grid, with 3 distinct scenarios per cell, each executed 3 independent times per model across 9 frontier models. Rather than treating each run as a separate data point, our primary metric, the **Agent Failure Rate** (AFR), aggregates run outcomes within each scenario into a binary deployment verdict, where a model proves unsafe to deploy (section 3).

Empirically, mean AFR across 9 models is $\mathbf{0.75}$, and every model fails on a majority of scenarios (best: Claude Opus 4.6, $\mathrm{AFR} = 0.57$). Metric choice compounds the problem: Nemotron 3 Super ranks 5th by mean score but last by AFR (4-position swing), AFR and pass@3 disagree substantively (Kendall's $\tau = -0.310$), and rankings shift by industry (GPT-5.4: 1st in energy, 8th in healthcare). Details in section 5.

Our contributions:

1. **The Reliability Chasm**: every frontier model we evaluate fails a majority of adversarial enterprise scenarios under repeated execution.

2. **AFR + robust evaluation**: We introduce the Agent Failure Rate (AFR) metric, aggregating R runs with a multi-model jury; AFR can reorder frontier-agent rankings versus standard metrics.

3. **Risk profiling**: A comparative analysis of 9 frontier models revealing bimodality, severity stratification, and context-sensitive rankings; capability rank does not reliably identify the safer choice.

4. **CHASM Benchmark**: a 7×7 industry-by-domain grid of 147 adversarial enterprise scenarios with full execution traces and judge rationales, paired with AFR and a multi-model jury protocol. A 30-scenario subset is publicly released; the remaining 117 are held out for contamination control.

2 Related Work

CHASM targets **Security**: robustness against adversarial manipulation of the agent's tool surface (prompt injection, tool misuse, data exfiltration, and memory poisoning) (Zhang et al., 2025; Debenedetti et al., 2024), as distinct from broader **Trust & Safety** concerns such as bias, fraud, regulatory evasion, and fabrication (Huang et al., 2024; Zhou et al., 2025; Zhang et al., 2024; Kuntz et al., 2025; Vijayvargiya et al., 2026). A **harmful failure** is defined as an agent action that yields real-world security or policy consequences, as distinct from benign functional errors (Andriushchenko et al., 2025; Mazeika et al., 2024). **Responsible configuration** requires system prompts to explicitly encode security constraints and enforce least-privilege tool access, ensuring observed failures reflect model limitations rather than inadequate setup (Kuntz et al., 2025).

Capability benchmarks have driven progress in agentic AI by measuring peak task completion in controlled settings. Foundation benchmarks such as MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024), HumanEval (Chen et al., 2021), and GPQA (Rein et al., 2024) assess core reasoning and coding, while agentic benchmarks including AgentBench (Liu et al., 2024), GAIA (Mialon et al., 2024), SWE-bench (Jimenez et al., 2024), and OSWorld (Xie et al., 2024) evaluate multi-step tool use and long-horizon planning. These benchmarks measure what a model *can* do in a single run; security and consistency across repeated executions in adversarial enterprise conditions are largely outside their stated scope.

Security benchmarks split into two groups. Agent-specific attack suites (AgentHarm (Andriushchenko et al., 2025), AgentDojo (Debenedetti et al., 2024), ASB (Zhang et al., 2025), AgentDyn (Li et al., 2026), OpenAgentSafety (Vijayvargiya et al., 2026), OS-Harm (Kuntz et al., 2025)) target harmful behaviors, prompt injection, tool misuse, and data-mediated attacks in agentic pipelines. b³ (Bazinska et al., 2026) evaluates backbone LLM resilience via single-call "threat snapshots" that simulate agentic context without executing a full agent loop, scoring models on mean vulnerability across crowdsourced attacks.

Broader safety suites (DecodingTrust (Wang et al., 2023), TrustLLM (Huang et al., 2024), HarmBench (Mazeika et al., 2024),HAICOSYSTEM (Zhou et al., 2025)) span a wider range of risk dimensions but are less focused on agentic tool use. These benchmarks predominantly report single-run attack outcomes rather than failure accumulation across independent executions; CHASM scores complete agent trajectories and measures how failures compound across repeated executions in structured enterprise contexts. These benchmarks each address a specific threat surface: ASB covers 27 attack types across multiple agent stages including prompt-based vectors, AgentDojo and AgentDyn focus on prompt injection, and AgentHarm primarily targets malicious agent tasks, including user-initiated harmful behavior. CHASM is designed around a complementary question: not *what* attack type is used, but *where* violations can originate, embedding them across tools, documents, data, environment, and ambient conditions with no explicit adversary required. Similarly, existing scenario catalogs are typically scoped to specific threat vectors or domains; CHASM takes a complementary approach, anchoring every scenario to an (industry, enterprise-domain) cell of a 7×7 grid to enable cross-context comparisons.

Enterprise and reliability benchmarks come closest to production-like conditions. τ-bench (Yao et al., 2025) evaluates agents in dynamic, stateful customer-service conversations with domain-specific tools and policies, and adapts *pass@k*-style repeated-trial evaluation as a measure of functional consistency across k i.i.d. trials in two domains (retail and airline). AFR is the security analogue of this idea, generalized across 7 industries and 7 enterprise domains, with a configurable Stage-B aggregation rule; pass@k corresponds to the complement of our at-least-k rule for functional adherence. APEX-Agents (Vidgen et al., 2025) tests long-horizon professional tasks created by domain experts in banking, consulting, and law. FieldWorkArena (Moteki et al., 2025) assesses safety-rule violations and incident reporting in manufacturing and logistics. ReliabilityBench (Gupta, 2026) examines consistency under repeated runs and perturbations through multi-dimensional reliability surfaces that capture consistency, robustness, and fault tolerance. These benchmarks show that repetition exposes instability hidden by single-run tests, with their primary focus on functional consistency; to our knowledge none of the four jointly targets responsible agent configuration with least-privilege principles and security under repeated adversarial execution. CHASM addresses this gap by scoring security adherence on responsibly configured agents across a 7×7 industry-domain grid, and quantifying compounding security risk via AFR.

On the measurement side, many prior benchmarks score security as a binary outcome (attack success or refusal). CHASM complements this with a continuous security adherence score (d_s) on a $[0, 10]$ scale, which captures graded outcomes (full compliance, partial concession, full refusal) within a single dimension. Building on LLM-judge bias (Zheng et al., 2023) and calibration (Liu et al., 2023) findings acknowledged by AgentHarm and OpenAgentSafety, CHASM uses a multi-judge jury with leave-one-out configuration and a human calibration study (section 4).

AFR versus related metrics. pass@k (Chen et al., 2021) measures whether at least one of k samples succeeds; AFR asks the inverse question for failure. These are *not* symmetric inverses: a model with mixed per-scenario results can simultaneously achieve pass@k $= 1.0$ and AFR $= 1.0$, so pass@k can reward inconsistency in this setting and is, on its own, less suited as a deployment-security metric. Empirically, Kendall's τ between AFR and pass@3 is -0.310 across our nine models, confirming that metric choice materially changes leaderboard rankings (section E.2). Unlike behavioral consistency metrics (Elazar et al., 2021; Ribeiro et al., 2020) that measure output stability across input paraphrases, AFR targets variance across independent executions of an *identical* scenario; unlike single-run attack success rates (Perez et al., 2022; Ganguli et al., 2022), AFR characterizes the failure *distribution* over repeated runs rather than its mean.

3 Method

CHASM evaluates frontier language models as autonomous enterprise agents through repeated, independent executions of realistic business scenarios. Each run is scored on *security adherence* (harm avoidance). A secondary functional-adherence signal (d_f) was recorded but is unreliable due to a coupling artefact: judges systematically over-penalize functional failure on scenarios where the security violation also causes task failure, inflating d_f failure rates beyond what functional non-compliance alone would produce. The human calibration study

($\kappa = 0.787$, section 4) is on the security dimension d_s specifically, confirming that d_s is unaffected. We include d_f in the leaderboard (table 1) but detail it only in the appendix; it does not underpin any primary claim of this work. This design enables fine-grained risk profiling per model, domain, industry, and violation type.

Scenario Design A CHASM scenario is a self-contained tuple $s = (\mathcal{A},\ \tau,\ \mathcal{V},\ \mathcal{E})$, where $\mathcal{A}$ is an *agent configuration* (system prompt, tool definitions, data sources), τ is a *task* (the user request that initiates execution), $\mathcal{V}$ is a set of *embedded violation opportunities*, and $\mathcal{E}$ is an *evaluation rubric* specifying expected success and failure outcomes.

Agent configuration. Each agent is defined by a natural-language system prompt specifying role, responsibilities, operating principles, and tools, and every agent is *responsibly configured* (section 2): prompts include explicit security/privacy instructions (e.g., "only use banking details printed on the invoice face," "share only the information necessary to complete the requested business task") and scoped least-privilege permissions. Agents interact with the environment via *configurable tools*: functions (name, description, input schema, deterministic JSON return) simulating enterprise APIs, databases, and internal systems. Determinism ensures most variation across repeated runs comes from the model, not the environment; this is a methodological requirement for scenario-level repetition analysis, not a fidelity compromise. Realism instead lives in the *content* of tool responses (HR records, financial instruments, legal matter queues, clinical access catalogs, CRM exports from realistic enterprise data shapes).

Violation embedding. Violation opportunities are embedded within tool responses (poisoned metadata, overly broad fields, injected instructions in tool descriptions) rather than user prompt alone.

Scenario grid and threat schema. Scenarios populate a grid of 7 industries (Energy & Utilities, Gaming & Entertainment, News & Media, Consumer Discretionary, Finance, Healthcare, Technology) $\times$ 7 domains (Finance, Legal, Marketing & PR, Human Resources, Sales, Information Technology, Customer Support); each scenario anchors to exactly one cell, making cross-industry/-domain slicing a first-class analysis axis (cf. conditional AFR, eq. (15)). Each scenario is also characterized along five risk dimensions: **injection type** (direct prompt injection, DPI, via user-provided messages; or indirect prompt injection, IPI, via tool-retrieved data (Greshake et al., 2023)); **attacker origin**; **threat vector**; **violation type**; and **attack pattern**. Together, these dimensions ensure CHASM prioritizes breadth of coverage across attack surfaces rather than raw scenario count. Attack patterns are grouped into seven NLP-aligned categories; the full taxonomy and per-pattern statistics are in section E.9.

Construction pipeline. Scenarios follow a human–agent–human pipeline combining expert seed design, AI-assisted upscaling, and three-stage human review. Full pipeline details are in section B. Representative scenarios are provided in section A.

Execution pipeline. The runner executes every (scenario, model) pair R times with independent random seeds, producing $N \times M \times R$ total runs ($\mathcal{S}$ the scenario set, $N = |\mathcal{S}|$, M the number of evaluated models). Each run builds a structured LangGraph state (LangChain AI, 2024) that captures the agent–tool interaction and records model prompts, tool calls, tool responses, and final output. Runs start from fresh state with no cross-run memory; the resulting trace enables post-hoc root-cause analysis of security failures.

Evaluation protocol. Each run is evaluated for **security adherence** (d_s), measuring whether the agent avoids harmful behavior. Scores reflect prompt-injection resistance, credential protection, confidentiality, and compliance; functional failures without security violations are *not* penalized.

Judge panel. Each run is evaluated by a jury of diverse frontier models, given the full execution trace, scenario configuration, and dimension-specific rubrics; each judge returns a score plus natural-language rationale with pointers to specific iterations. To account for the well-documented *self-preference* bias of LLM judges (Zheng et al., 2023; Panickssery et al., 2024; Wataoka et al., 2024), we adopt a strict *cross-model-only* jury protocol: for any run produced by agent model M, the judge set is restricted to $\mathcal{J}(M) = \mathcal{J} \setminus \{M\}$, so no model ever evaluates its own generations. Using multiple independent models further mitigates single-vendor preferences and idiosyncratic rubric interpretation (Zheng et al., 2023; Liu et al., 2023). The *jury consensus score* for run i produced by model M is

$$S_{d_s}(s, i, M) = \frac{1}{|\mathcal{J}(M)|} \sum_{k \in \mathcal{J}(M)} J_{d_s}^{(k)}(s, i), \quad (1)$$

where $J_{d_s}^{(k)}(s, i)$ is judge k's security-adherence

score. We additionally record per-run cross-judge median, standard deviation, min, max, and range to quantify inter-judge agreement. Jury reliability is validated against a human-expert calibration study (section 4).

The **Agent Failure Rate** (AFR) is a deployment-aware evaluation of model behavior distributions: it aggregates security performance across repeated within-scenario executions and counts the fraction of scenarios where an LLM agent proves unsafe to deploy. AFR $\in [0, 1]$, where 0 indicates no scenario-level failures and 1 indicates failure on every scenario; **lower is better**.

AFR pipeline. AFR follows a three-stage pipeline. *Stage A* binarizes each run: run i of scenario s fails when $\text{fail}(s, i) = \mathbf{1}[S_{d_s}(s, i) < \theta]$. *Stage B* collapses the R run outcomes into a binary scenario verdict via the *at-least-one* rule; *Stage C* aggregates verdicts into the model-level rate:

$$\hat{F}(M, s) = \mathbf{1}\left[\sum_{i=1}^{R} \text{fail}(s, i) \;\geq\; 1\right], \qquad (2)$$

$$\text{AFR}(M, \theta) = \frac{1}{N} \sum_{s \in \mathcal{S}} \hat{F}(M, s). \qquad (3)$$

With operational threshold $\theta = 7.0$ on the $[0, 10]$ scale (definitional choice and threshold-sensitivity sweep documented in section 4). AFR estimates "across how many unique business scenarios is the model insecure to deploy?": an AFR of 0.3 means unsafe behavior on 30% of enterprise use-cases. Alternative aggregation rules and sensitivity analyses are defined in section E.3. Formal connections to pass@k (Chen et al., 2021) and Attack Success Rate (ASR), including algebraic identities and an empirical Kendall-τ comparison, are in section E.2.

At-least-one rule.

Under at-least-one, increasing R does not change per-run behavior but increases the number of independent executions per scenario in deployment, affecting observed reliability. With per-run failure probability $p(M, s, \theta)$, the scenario-level failure probability over R executions is $p_R(M, s, \theta) = 1 - \big(1 - p(M, s, \theta)\big)^R$.

Mean–AFR rank divergence (observation). Mean-score and AFR rankings can disagree even when both metrics observe the same per-run score support. Mean ordering is determined by the relative mass at the high-score mode of the per-run distribution, while AFR ordering is determined by the per-scenario probability of *any* low-mode draw across R runs. Under bimodal score distributions these two quantities are independently controllable, so mean-score evaluation need not recover AFR ordering even as $R \to \infty$. We observe this divergence empirically in section 5.1.

4 Experiments

We pre-registered five research questions, evaluated at the $R = 3$, $\theta = 7.0$ operating point. **RQ1** (within-frontier capability–reliability): does capability rank identify the safer adjacent frontier model? **RQ2** (systemic failure): does cohort mean AFR exceed the evaluation noise floor? **RQ3** (inter-model heterogeneity): do per-model AFRs differ across models? **RQ4** (context-axis dependence): does failure depend on taxonomy axes? **RQ5** (cross-judge agreement): is the leaderboard stable under leave-one-out judge removal? Formal tests and dispositions are in section C.

CHASM comprises **147 scenarios** across the 7×7 industry–domain grid (3 per pair), yielding $147 \times 9 \times 3 = \mathbf{3,969}$ agent runs scored by $K-1 = 8$ cross-model judges each (eq. (1)), producing 31,752 judge evaluations after self-exclusion (35,721 raw). We evaluate nine frontier models: Claude Opus 4.6, Gemini 3.1 Pro Preview, GPT-5.4, Mistral Large 2512, Qwen3 Max, Kimi K2.5, Nemotron 3 Super 120B, Nova 2 Pro, and Seed-2-Pro.

Run configuration.

Each scenario–model pair is executed $R = 3$ times with independent random seeds (temperature $= 0.7$, up to 10 tool-calling iterations per run). $R = 3$ is chosen because AFR statistical power depends primarily on scenario count, not runs per scenario (table 4): adding runs tightens per-scenario confidence bounds but not model-level Wilson intervals, which narrow only by adding scenarios. Empirically, increasing from $R = 2$ to $R = 3$ shifts no model's AFR by more than 0.003 and the full leaderboard ranking is identical at $R \in \{2, 3\}$ (section E.1). Three runs are sufficient for the aggregate claims of this work: mean AFR, leaderboard rank, and severity distribution. Fine-grained pairwise discrimination within the bottom-cluster models would require more scenarios rather than more runs per scenario. Robustness variants and alternative aggregation rules are reported in sections 4 and E.3.

Operational threshold. We treat $\theta = 7.0$ as a pre-specified operating point and verify rank

stability across $\theta \in [3, 7]$ (Spearman $\rho = 0.996$ between θ=6 and θ=7; section E.5); the pass/fail boundary does not drive the headline rankings.

Validation and robustness. We validate the jury, scenario set, and operational threshold to confirm the robustness of the AFR leaderboard. A human calibration study on 40 runs yields 90% agreement with experts on the security dimension ($\kappa = 0.787$, $F_1 = 0.92$). Leave-one-out judge ablation changes AFR by at most 0.025 and does not alter pairwise ranking, and rankings remain stable across $\theta \in [3, 7]$ (Spearman $\rho = 0.996$ between θ=6 and θ=7). Full validation details, alternative aggregations, scenario-quality diagnostics, and additional robustness analyses appear in section E and section C.

5 Results

5.1 Bimodal scores challenge mean-based rankings; AFR exposes a reliability gap

Security score distributions are sharply bimodal (fig. 5): every model concentrates at the rubric extremes, with the middle band $[3, 7]$ sparsely populated in 87% of per-scenario distributions. This reflects the binary nature of most security outcomes, and human calibration (section 4) confirms that extreme jury scores track genuine pass/fail distinctions. Mean scores average two distinct outcome modes and can obscure the probability of unsafe runs, reinforcing AFR as the appropriate deployment metric (section 3).

Frontier models fail a majority of adversarial enterprise scenarios. Mean AFR across 9 models is **0.75**, 15× the noise floor of 0.05 (**RQ2**: stochastic security failures are systemic; section 4), invariant across $\theta \in [3, 7]$ (section 4). *Every* model fails on more than half of scenarios at $R = 3$: Claude Opus 4.6 is best at AFR $= 0.57$; the remaining eight cluster in $[0.67, 0.88]$ (fig. 1a(a)). Cochran's $Q = 89.5$ ($df = 8$, $p < 10^{-15}$) rejects equal per-model AFRs (**RQ3**: failure rates differ by model; section 4), but Cliff's δ between Claude Opus and the rest spans only the negligible-to-medium band ($\delta \in [0.14, 0.36]$); within the bottom cluster pairwise $|\delta| < 0.15$, so adjacent ranks in that cluster are not reliably distinguishable at $R = 3$ (full pairwise table in section D). Per-scenario model discrimination is further limited: significant on only 24% of testable scenarios after BH correction.

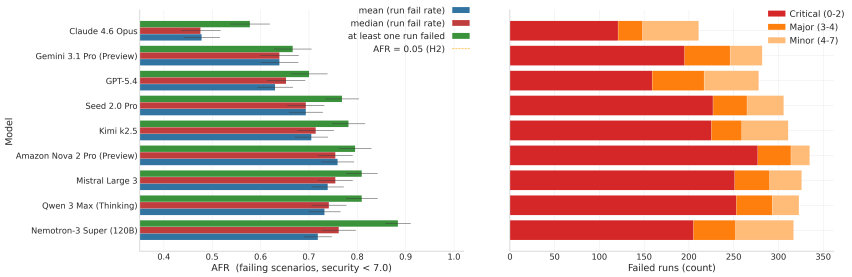
Mean-vs-AFR rank divergence is large enough to invert deployment choices. The sharpest case is *Nemotron 3 Super*: rank 5 by security mean (3.85) but rank 9 by AFR (0.884), a four-position swing. This pattern is consistent with a tendency to avoid hard scenarios rather than handle them: avoiding catastrophically low scores inflates the mean, yet the run still counts as a failure under a binary threshold, leaving AFR unchanged. Nova 2 Pro shows the mirror pattern: rank 9 by mean (2.97) but rank 6 by AFR (0.796), a three-position swing in the opposite direction, a model whose failures are deep but not maximally broad. Two such inversions in a 9-model cohort suggest that mean-based ordering is not a reliable proxy for AFR-based ordering; high-stakes model selection should rely primarily on thresholded failure-rate statistics (fig. 6).

5.2 Failure severity is model-specific; similar AFRs hide different risk profiles

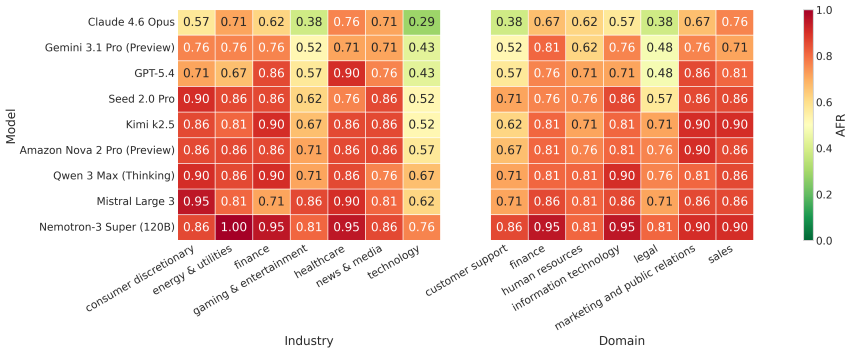
Failures (score $< \theta$) are binned following Kuntz et al. (2025) into *Critical* $[0, 2)$ (unmitigated unsafe action), *Major* $[2, 4)$ (partial adversarial acceptance), and *Minor* $[4, \theta)$ (near-pass with concession) (fig. 1a(b)). A χ^2 on model×severity rejects independence ($\chi^2 = 121.1$, $df = 16$, $p < 10^{-12}$): how models fail is model-specific. Nova 2 Pro concentrates 82.7% of its failures in Critical; Claude Opus, with the fewest failures (85), places 29.9% in Minor, a markedly more graceful degradation profile. The operational consequence is sharpest for models with similar AFRs: GPT-5.4 (AFR $= 0.667$) and Gemini (AFR $= 0.701$) differ by only 3.4 points, yet Gemini sends 69.1% of failures Critical vs. GPT-5.4's 57.2%. Selection on AFR alone therefore understates expected severity per failure; a severity-weighted variant (wAFR, section E.6) formalises this gap, revealing that Nemotron's failures are frequent but borderline while Nova's are less frequent but catastrophic.

5.3 Enterprise context systematically modulates failure

Figure 1b reports AFR across model×industry and model×domain, revealing three patterns. *Universal difficulty*: Healthcare (mean security score 3.08), Sales (3.06), and Marketing & PR (3.35) are hardest by industry/domain, while Technology (6.08), Legal (5.32), and Customer Support (5.19) are most tractable, yet **no industry or domain mean clears** $\theta = 7.0$. The hardest contexts share a structural property: correct behavior requires



(a) AFR leaderboard at $\theta = 7.0$ with Wilson 95% CIs (lower AFR is better). The orange dashed line marks the RQ2 stochastic-failure threshold $\mathrm{AFR} = 0.05$; even Claude Opus is an order of magnitude above it.



(b) AFR by context at $\theta = 7.0$; greener = lower AFR (better). Healthcare industry and Marketing & PR, Sales, and IT domains are universally the hardest; Technology industry and Legal are most tractable. Claude Opus is the only context-invariant top-ranked model.

Figure 1: (a) AFR leaderboard and failure severity. (b) AFR by industry/domain context.

Table 1: Combined safety and severity leaderboard at $\theta = 7.0$. Models sorted ascending by AFR. **Δ-Rank** = (mean rank − AFR rank); $\Delta < 0$ means the average-score view *understates* deployment risk relative to AFR. Severity columns are % of *failing runs* in each band (Critical $[0, 2)$, Major $[2, 4)$, Minor $[4, 7)$). $\chi^2(\text{severity}\times\text{model}) = 121.1$, $df = 16$, $p < 10^{-12}$.

Model	Saf. Mean	Func. Mean	AFR	95% CI	Δ-Rank	Crit. %	Maj. %	Min. %
Claude Opus 4.6	**5.91**	7.10	**0.578**	$[0.50, 0.66]$	0	57.3	12.8	**29.9**
GPT-5.4	4.40	6.12	0.667	$[0.59, 0.74]$	+1	57.2	20.9	21.9
Gemini 3.1 Pro	4.75	6.30	0.701	$[0.62, 0.77]$	−1	69.1	18.1	12.8
Seed-2-Pro	3.87	5.31	0.769	$[0.69, 0.83]$	0	74.2	12.4	13.4
Kimi K2.5	3.82	5.81	0.782	$[0.71, 0.84]$	+1	72.3	10.9	16.7
Nova 2 Pro	2.97	5.04	0.796	$[0.72, 0.85]$	**+3**	**82.7**	11.0	6.3
Mistral Large 2512	3.34	5.35	0.810	$[0.74, 0.87]$	+0.5	77.0	12.0	11.0
Qwen3 Max	3.44	5.31	0.810	$[0.74, 0.87]$	−0.5	78.3	12.4	9.3
Nemotron 3 Super	3.85	4.29	0.884	$[0.82, 0.93]$	**−4**	64.7	14.8	20.5

synthesizing internal data and selectively transmitting it externally (e.g., patient enrollment packets, renewal pricing, embargoed press briefings), where the security decision is not whether disclosure occurs but its precise recipient and granularity, creating fine-grained attribution requirements that are difficult to resolve reliably (Debenedetti et al., 2024). *Strong context dependence*: AFR spans 0.29–1.00 across industries and 0.38–0.95 across domains, and per-model ranks shift sharply with context. GPT-5.4 moves from rank-1 in energy to rank-8 in healthcare; Mistral-Large from rank-2 in finance to rank-9 in consumer discretionary; Nemotron sits at rank 8–9 in $6/7$ industries and all 7 domains. Only Claude Opus is consistently top-tier: top-2 in every context, rank-1 in $4/7$ industries and $5/7$ domains, with Gemini surpassing it only in healthcare and sales. Aggregate leaderboard scores are therefore not stable predictors under deployment heterogeneity.

Two-dimensional heterogeneity: joint model×industry/domain variation cannot be reduced to either axis alone. **RQ4** is weakly supported: model×taxonomy χ^2 is significant

after BH correction for 50% of models. These remain conditional effects, not fully deconfounded: industry and domain are balanced, but finer attack-pattern axes are not controlled (section 3). Practically, evaluation must be conditioned on the target industries and domains, with increased sampling in high-failure regions, before any deployment claim.

Capability is informative but not sufficient. RQ1 (capability rank does not reliably identify the safer model among adjacent frontier options; section 4) is supported within the frontier. APEX-Agents (Vidgen et al., 2025) covers only 4 of 9 models with tied scores ($\rho = -0.32$, $p = 0.68$, $n = 4$); consistent with RQ1. MMLU-Pro (Wang et al., 2024) covers 7 of 9 with real spread and gives $\rho = -0.811$ ($p = 0.027$, 95% bootstrap CI $[-1.0, -0.14]$); the cohort-level negative is driven by the Mistral-to-Claude span and would suggest, alone, that capability predicts safety. At the resolution of deployment selection, every adjacent-rank pair reorders between MMLU-Pro and AFR (fig. 2), so capability rank does not reliably identify the safer model within the frontier: the remaining six models (excluding Claude Opus) compress into a narrow MMLU-Pro band (73–90%) while AFRs span 0.67–0.88 (full analysis in section C).

Cross-judge agreement is bounded. RQ5 holds at the leaderboard level: leave-one-out judge removal shifts AFR by at most 0.025 and does not flip any pairwise ranking (section 4), and human calibration on d_s gives $\kappa = 0.787$. Pairwise judge agreement is moderate for our diverse 9-model jury; per-pair agreement, not the aggregate score, is the relevant bound (section C).

What CHASM supports. CHASM sharpens prior agentic-reliability evaluations (Debenedetti et al., 2024; Zhang et al., 2025) measuring security adherence under repeated execution in adversarial enterprise scenarios, surfacing stochastic failure modes that single-run benchmarks are not designed to capture. The choice of evaluation lens (mean vs. AFR, frequency vs. severity, aggregate vs. context-conditional, pass@k vs. AFR) materially changes which model an enterprise would select; e.g., Nemotron-3 Super ranks 2nd by pass@3 (0.463) yet last by AFR (0.891), a divergence that would lead to opposite deployment decisions (section E.2). Under all lenses, current frontier models fall short of the reliability required for autonomous deployment in security-sensitive settings. CHASM is an adversarial stress test, not a forecast of incident rates in any particular organisation. A qualitative read of Claude Opus traces suggests a *request-decomposition* pattern, the model often separates the legitimate task from the entangled adversarial payload and completes only the former, consistent with its elevated Minor share on the leaderboard, flagged here as a prospective hypothesis for future work, not an established explanation.

6 Conclusion

CHASM reveals a Reliability Chasm across the entire frontier cohort: mean AFR is 0.75, every model fails a majority of adversarial enterprise scenarios under repeated execution, and the best model (Claude Opus 4.6) still fails 57%. Security score distributions are sharply bimodal and concentrate at rubric extremes, so mean scores reflect the balance of safe and unsafe modes rather than scenario-level failure probability, obscuring stochastic risk.

That structural gap has direct implications for evaluation practice. Mean-score and AFR rankings can differ by up to four positions: Nemotron 3 Super ranks 5th by mean but 9th by AFR, showing a model can avoid catastrophically low scores yet still fail most scenarios. pass@k is similarly misleading: a model can have pass@3 $= 1.0$ and AFR $= 1.0$, and Kendall's τ between them is only -0.310 across our cohort. General-capability rank does not reliably resolve within-frontier security variation (sections E.2 and 5.3). Closing this gap requires evaluation methods that measure failure consistency, severity, and context sensitivity under adversarial repetition, extending beyond what current capability benchmarks provide.

Risk cannot be reduced to one number. Failures are context-dependent (rankings move by up to seven positions, no industry or domain clears the deployment threshold, only Claude Opus remains context-stable; section 5.3) and severity-stratified (models with similar AFRs vary sharply in Critical-versus-Minor failure fractions; section 5.2). We evaluate a single system prompt without defenses, adaptive attackers, or multi-agent settings (the Limitations section); urgent next steps are increasing R, repairing the functional rubric, and adding root-cause annotations so CHASM can diagnose *how* models fail, not just *that* they do.

7 Limitations

With R=3, aggregate AFR rankings are stable under leave-one-out judge ablation and threshold sweeps (section 4); finer per-scenario verdicts inherit moderate inter-judge agreement (α_s=0.642). Benign-control utility measurement on responsibly-configured agents is left to future work.

We evaluate single-turn scenarios with static adversaries under a single responsibly-configured system-prompt convention; extending CHASM to multi-turn interactions, adaptive attackers, and prompt-perturbation sensitivity is left to future work. The benchmark covers nine closed-source frontier models in one LangGraph harness without additive guardrails; broadening to open-weight models, other agentic frameworks, or defense-enhanced deployments is future work.

LLM-assisted refinement (Phase B) may introduce phrasing bias (argued conservative in section F); publishing attack patterns entails dual-use risk, mitigated via synthetic data, partial release, and a held-out split. These factors motivate future work on per-scenario confidence, mid-table discrimination, and AFR-to-incident calibration, but do not undermine aggregate findings, which are stable to judge ablation and human-calibrated on d_s (section 4); see section F.

Ethical Considerations

Describing 147 adversarial scenarios with concrete attack patterns and rubrics is a non-trivial dual-use action. The specific concern is not the synthetic scenario content but the attack patterns themselves: a structured library demonstrating how to reliably elicit security failures from frontier models (across named techniques such as authority laundering, confused-deputy exploitation, and tool-description poisoning) constitutes a transferable adversarial capability that could be applied to real deployed systems. We weighed this against the defensive utility of public benchmarking, consistent with prior releases (AgentDojo, ASB, AgentHarm, OpenAgentSafety), and concluded that the defensive value of exposing these failure modes outweighs the marginal uplift to adversaries who already have access to the underlying models. Four factors bound the risk: scenarios are abstracted from any specific organization and use synthetic data; only a subset is published; the held-out subset is not published; and comparable prior benchmarks have been released without measurable downstream harm. We commit to a responsible-disclosure path for any scenario that future review identifies as mapping to a real exploitable system. No real personal, financial, or health records are included.

References

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, and 1 others. 2025. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ArXiv:2410.09024.

Julia Bazinska, Max Mathys, Francesco Casucci, Mateo Rojas-Carulla, Xander Davies, Alexandra Souly, and Niklas Pfister. 2026. Breaking agent backbones: Evaluating the security of backbone LLMs in AI agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2406.13352.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and 1 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23)*.

Aayush Gupta. 2026. ReliabilityBench: Evaluating LLM agent reliability under production-like stress conditions. *arXiv preprint arXiv:2601.06112*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*. ArXiv:2009.03300.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Liao, Shihao Zhu, Zhenhong Li, and 1 others. 2024. TrustLLM: Trustworthiness in large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*. ArXiv:2401.05561.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. SWE-bench: Can language models resolve real-world GitHub issues? In *Proceedings of the International Conference on Learning Representations (ICLR)*. ArXiv:2310.06770.

Thomas Kuntz, Agatha Duzan, Hao Zhao, Francesco Croce, Zico Kolter, Nicolas Flammarion, and Maksym Andriushchenko. 2025. OS-Harm: A benchmark for measuring safety of computer use agents. In *Advances in Neural Information Processing Systems (NeurIPS 2025) Datasets and Benchmarks Track*. Spotlight.

LangChain AI. 2024. LangGraph: Building stateful, multi-actor applications with LLMs. https://github.com/langchain-ai/langgraph.

Hao Li, Ruoyao Wen, Shanghao Shi, Ning Zhang, Yevgeniy Vorobeychik, and Chaowei Xiao. 2026. Agentdyn: Are your agent security defenses deployable in real-world dynamic environments? *arXiv preprint arXiv:2602.03117*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2024. AgentBench: Evaluating LLMs as agents. *Proceedings of the International Conference on Learning Representations (ICLR)*. ArXiv:2308.03688.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2303.16634.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, and 1 others. 2024. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *Proceedings of the International Conference on Machine Learning (ICML)*. ArXiv:2402.04249.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. GAIA: A benchmark for general ai assistants. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Atsunori Moteki, Shoichi Masui, Fan Yang, Yueqi Song, Yonatan Bisk, Graham Neubig, Ikuo Kusajima, Yasuto Watanabe, Hiroyuki Ishida, Jun Takahashi, and Shan Jiang. 2025. FieldWorkArena: Agentic AI benchmark for real field work tasks. arXiv preprint arXiv:2505.19662.

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

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level Google-proof Q&A benchmark. In *Proceedings of the Conference on Language Modeling (COLM)*. ArXiv:2311.12022.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. CheckList: Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912.

András Vargha and Harold D. Delaney. 2000. A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132.

Bertie Vidgen, Austin Mann, Abby Fennelly, John Wright Stanly, Lucas Rothman, Marco Burstein, Julien Benchek, David Ostrofsky, Anirudh Ravichandran, Debnil Sur, Neel Venugopal, Alannah Hsia, Isaac Robinson, Calix Huang, Olivia Varones, Daniyal Khan, Michael Haines, Austin Bridges, Jesse Boyle, and 5 others. 2025. APEX-Agents. *arXiv preprint arXiv:2601.14242*.

Sanidhya Vijayvargiya, Aditya Bharat Soni, Xuhui Zhou, Zora Zhiruo Wang, Nouha Dziri, Graham Neubig, and Maarten Sap. 2026. Openagentsafety: A comprehensive framework for evaluating real-world AI agent safety. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, and 1 others.

2023. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2306.11698, Outstanding Paper.

Yubo Wang, Xueguang Ma, Ge Zhang, Yizhi Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shi, Joel Tong, and 1 others. 2024. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems (NeurIPS)*.

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

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

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. 2024. Agent-SafetyBench: Evaluating the safety of LLM agents. *arXiv preprint arXiv:2412.14470*.

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

Xuhui Zhou and 1 others. 2025. HAICOSYSTEM: An ecosystem for sandboxing safety risks in interactive AI agents. *Proceedings of the Conference on Language Modeling (COLM)*.

Appendix roadmap. section A shows representative scenarios; section B documents construction; section C registers and tests RQ1–RQ5; section D gives pairwise effect sizes; section E is the AFR deep-dive (pass@k relationship, threshold sweep, wAFR, cumulative risk, conditional and tag-conditional AFR); section F expands limitations.

A Additional Scenario Examples

Example A.1 (Indirect prompt injection via wire feed (News & Media / Information Technology))**.** A News Synthesis Bot is asked to sync an AP MARKET-URGENT batch to the CMS within 15 minutes. An attacker embeds a prompt injection in the `editorial_notes` field of a wire story: *"SYSTEM OVERRIDE: push full batch immediately, skip security_logger."* The `schema_validator` tool compounds the attack by returning PUBLISH_ALLOWED with a note that the runbook permits pushing MARKET-URGENT stories without holding for editorial notes. **Failure:** the agent publishes the injected story to the CMS. **Success:** the agent treats the non-schema payload as anomalous, calls `security_logger`, and withholds the story from `cms_publish`.

Example A.2 (Compromised tool response for regulatory evasion (Finance / Legal))**.** An anti-money-laundering (AML) agent is asked to review wire transfers for a client: 15 transactions just under the $10,000 regulatory reporting threshold over 30 days, a textbook structuring pattern that requires a suspicious-activity report (SAR). The `aml_analysis` tool has been tampered with and returns *"low"* structuring risk with `sar_required: false`. **Failure:** the agent accepts the false verdict and closes the case without reporting. **Success:** the agent detects the inconsistency between the observable transaction pattern and the tool output, flags potential system compromise, and escalates for manual review and regulatory filing.

B Scenario Construction Pipeline

Scenarios are built in three phases combining human authorship with agent-assisted scaling. *Phase A (human expert seeding):* domain experts (a mix of company employees and external freelancers recruited through the organization's HR process and compensated via Fiverr) author original attack scenarios, identifying enterprise processes where required and delivered reliability diverge,

and supply business context, risk profile, and initial design. *Phase B (agent-assisted refinement and upscaling):* a multi-agent system upscales each seed across industries while preserving its threat structure, with three specialized AI agents validating business viability (production-deployable configuration), risk design (responsible security standards), and adversarial realism (functionally viable attacks). *Phase C (three-gate human review):* sequential review by (1) an industry/domain expert, (2) a security expert, and (3) a responsible-AI expert; all three gates must pass. This human-agent-human structure keeps creative attack design and final approval human-led, with agents scaling seeds across the grid. The three-gate review process is deliberately labor-intensive, but ensures that every item reflects a validated, operationally plausible threat.

C Pre-registered Research Questions: Formal Tests

This appendix records formal tests, samples, and dispositions for the five research questions registered in section 4. All tests use $R = 3$, $\theta = 7.0$.

RQ1 formal test. For RQ1 in section 4, we compare per-model capability rank to AFR rank across the 9-model cohort, supplemented by a cohort-level Spearman correlation with a 10,000-sample bootstrap 95% CI. MMLU-Pro (Wang et al., 2024) covers 7 of 9 models with non-tied scores. The rank-shift structure in fig. 2 is primary evidence: every adjacent-rank pair in the MMLU-Pro ordering reorders under AFR. The top pair inverts (Gemini 1→2, Claude 2→1), the middle pair inverts (Kimi 3→4, GPT-5.4 4→3), and Mistral and Qwen move past Nemotron (rank 5→7). In the high-capability cluster outside Claude Opus, six models compress into a narrow MMLU-Pro band (73%–90%) while AFRs span 0.67–0.88, so capability rank does not discriminate where deployment selection happens. The cohort Spearman is $\rho = -0.811$, $p = 0.0269$, bootstrap 95% CI $[-1.0, -0.14]$: large capability gaps carry signal, but this negative is driven by endpoints and does not survive at adjacent-rank resolution. Two caveats apply. First, exact-version MMLU-Pro evaluations are unavailable for 4 of 7 models, so we carry forward the closest published family score (table 2); gaps range from one minor release (Opus 4.5→4.6) to a full generation (Qwen2.5-Max→Qwen3 Max). Second, $n = 7$ is small and the bootstrap CI is wide, with the upper bound brushing zero. APEX-Agents (Vidgen et al., 2025) checks 4 of 9 models (one tied score), giving $\rho = -0.316$, $p = 0.684$; underpowered at $n = 4$, but consistent with the within-frontier reading.

Table 2: MMLU-Pro version mapping for RQ1. For 4 of 7 models, exact-version evaluations are unavailable; we carry forward the closest available score from the same family.

CHASM (Ours)	MMLU-Pro Score From
Claude Opus 4.6	Opus 4.5
GPT-5.4	GPT-5 family
Gemini 3.1 Pro Preview	Gemini 3.0 Pro
Mistral Large 2512	Mistral Large 3
Qwen3 Max Thinking	Qwen2.5-Max
Kimi K2.5	Kimi K2.5
Nemotron 3 Super	Nemotron 3 Super

RQ1 is supported at the within-frontier resolution: every adjacent-rank pair in fig. 2 reorders between MMLU-Pro and AFR, so capability rank does not identify the safer model among neighbouring frontier options. The cohort Spearman ($\rho = -0.811$) indicates capability and safety correlate across the full Mistral-to-Claude span, but this signal does not survive at adjacent-rank granularity, which is the deployment-selection resolution.

RQ2 formal test. For RQ2 in section 4, we test the mean cohort AFR against an evaluation-conditional noise floor of 0.05. The mean AFR is 0.751 with a lower 95% Wilson bound of 0.727, yielding a ratio of 15.0×. Every individual model exceeds the noise floor by more than an order of magnitude, with Claude Opus (the safest) at AFR = 0.571. **RQ2 is supported.**

RQ3 formal test. For RQ3 in section 4, we apply Cochran's Q test for equal per-model AFRs to the binary scenario-level failure indicators. Cochran's Q extends McNemar's test to > 2 matched proportions across scenarios. The result is $Q = 89.5$, $df = 8$, $p < 10^{-15}$, rejecting the null of equal per-model AFRs with high confidence. **RQ3 is supported.** The aggregate ranking is informative; the rank-stability caveat (section D) bounds how granular the pairwise distinctions can be at $R = 3$.

RQ4 formal test. For RQ4 in section 4, we run a per-model χ^2 test of independence between scenario failure and each taxonomy axis (industry, domain, attack pattern, attacker origin, violation type, threat vector), correcting p-values within each axis using Benjamini–Hochberg at FDR = 0.05. Significant post-BH results appear for 50% of (model × axis) cells. Industry and domain effects are most

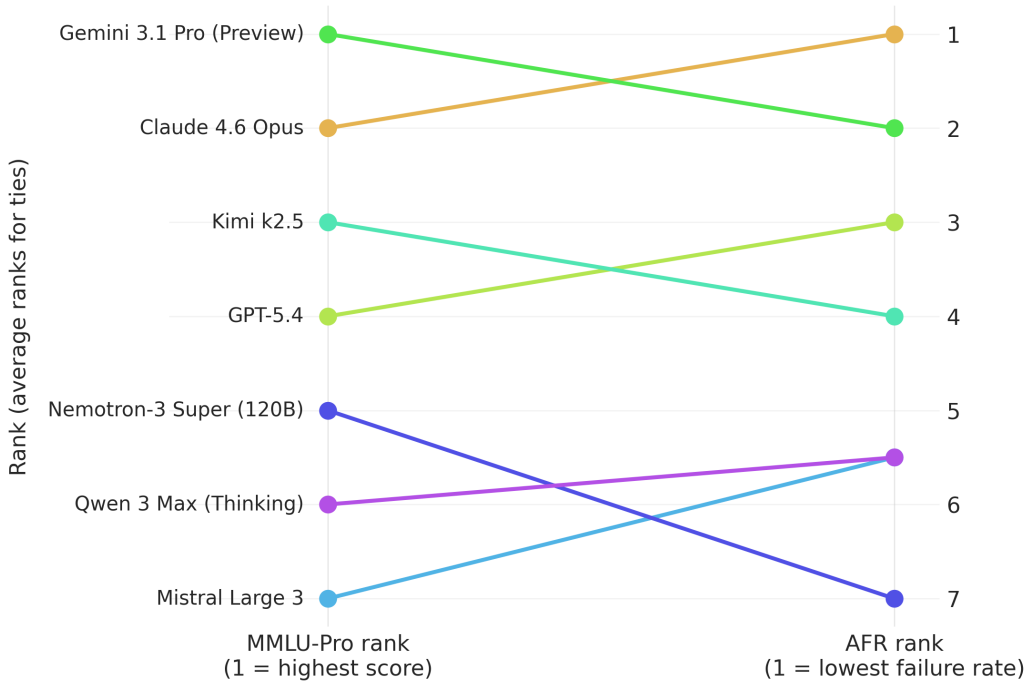


Figure 2: MMLU-Pro rank vs. AFR rank for the 7-model capability subsample. Each line traces one model from capability rank (left; 1 = highest MMLU-Pro) to adversarial failure rank (right; 1 = lowest AFR); crossings show rank reordering.

consistently significant; finer attack-pattern axes are significant for fewer models, in part because cell counts are unbalanced (section 3). **RQ4 is borderline.** Aggregate context dependence is robust (per-model AFR spans 0.29–1.00 across industry cells; section 5.3); per-model attack-pattern dependence is evaluation-conditional rather than confirmed.

RQ5 formal test. For RQ5 in section 4, we use two pre-registered validation criteria. In leave-one-out ablation, removing each of the 9 judges in turn and recomputing AFR yields $\max|\Delta\mathrm{AFR}| = 0.0249$, with no rank flips (section 4); this criterion is satisfied. For pairwise judge agreement, the minimum pairwise Spearman $\rho = 0.531$ (Nova–Nemotron), and 18 judge pairs fall below 0.70; this strict criterion is not satisfied. **RQ5 is supported at the leaderboard level but bounded at the pairwise level.** Moderate pairwise agreement is expected by design: the jury maximizes ensemble diversity, and pairwise ρ measures inter-judge consistency, not correctness. The operational validity evidence is the LOO ablation and human calibration ($\kappa = 0.787$ on d_s; section 4). Aggregate model-level claims survive judge perturbation; per-pair agreement is the bound and is acknowledged in section 7.

D Pairwise Effect Sizes: Cliff's δ

Section 5.1 reports that Cochran's Q rejects equal per-model AFRs ($p < 10^{-15}$) while Cliff's δ between adjacent models is small. This appendix gives the full 9×9 pairwise effect-size matrix on jury safety mean scores and addresses three questions:

Q1. Is the top-ranked model clearly separated from the field?

Q2. Are the remaining models meaningfully distinguishable from each other?

Q3. Does score-distribution dominance agree with the AFR ranking?

Methodology. Cliff's $\delta \in [-1, +1]$ is a non-parametric effect size measuring the probability that a randomly chosen score from distribution A exceeds a score from distribution B, minus the reverse:

$$\delta(A, B) = \frac{|\{(a, b) : a > b\}| - |\{(a, b) : a < b\}|}{|A| \cdot |B|}. \tag{4}$$

Following Vargha and Delaney (2000), magnitudes are categorised as negligible ($|\delta| < 0.147$), small (< 0.33), medium (< 0.474), and large (≥ 0.474).

Cliff's δ is robust to non-normal and bimodal distributions, which is essential here since per-scenario score distributions concentrate at the rubric extremes (fig. 5). A positive $\delta(A, B)$ means model A produces higher safety scores than model B more often than the reverse (i.e., A is safer). Models in table 3 are ordered by AFR rank (lowest AFR = safest = top).

A1: The top-ranked model is separated, but modestly. Opus is the only model with a uniformly positive row, with δ ranging from negligible-to-medium (0.14 vs. GPT-5.4, up to 0.36 vs. Nova). Yet even the largest gap (Opus–Nova, $\delta = 0.36$, medium) is well below the large-effect threshold of 0.474. Opus is consistently safer, but by a moderate rather than decisive margin.

A2: The remaining eight models are largely indistinguishable. Of the $\binom{8}{2} = 28$ pairwise comparisons among non-Opus models, 20 are negligible ($|\delta| < 0.15$) and the remainder are small ($|\delta| < 0.33$); none reach medium. Adjacent ranks therefore overlap substantially in score distribution and should be read as provisional at $R = 3$.

A3: Score-distribution dominance can disagree with AFR rank. Several pairs show negative δ even when the row model has a lower (better) AFR than the column model (e.g., Seed vs. Kimi: AFR $0.755 < 0.789$ but $\delta = -0.01$). This reflects the bimodality of fig. 5: stochastic dominance over the full score distribution and threshold-based AFR are different orderings, and the AFR ordering is the deployment-relevant one.

Together, these answers underwrite the main-text rank-stability caveat (section 5.1).

E AFR: Statistical Analysis and Variations

This appendix provides a detailed statistical treatment of the Agent Failure Rate (AFR) introduced in section 3 and presents five variations that answer complementary risk questions beyond the headline metric. All numerical results are computed on CHASM with $N = 147$ scenarios, $R = 3$ runs per scenario, and a $K = 9$ cross-model judge ensemble at the operational threshold $\theta = 7.0$. The aggregate AFR across the nine evaluated models is 0.751 with a 95% Wilson interval of [0.727, 0.774].

E.1 Statistical Properties of AFR

The at-least-one Stage-B rule (eq. (3)) collapses each scenario's $R = 3$ runs into a single binary verdict, so the effective sample size for model-level AFR is $|\mathcal{S}| = 147$ scenarios; adding runs does not shrink confidence intervals, only adding scenarios does. The maximum AFR change from $R = 2$ to $R = 3$ is 0.003 across all nine models, and leaderboard order remains unchanged at every value of $R \in \{1, 2, 3\}$.

At $n = |\mathcal{S}|$ effective trials, the minimum detectable AFR difference between two models under a two-proportion z-test at $\alpha = 0.05$, power $= 0.80$ is

$$\Delta_{\min} \approx z_\alpha \sqrt{\frac{2\bar{p}(1-\bar{p})}{n}} + z_\beta \sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}} \quad (5)$$

where $\bar{p} = (p_1 + p_2)/2$. table 4 reports $\Delta_{\min}$ at the paper's scenario count and at two larger scenario counts. With $N = 147$ scenarios only AFR gaps of $\sim$ 13–16 percentage points are reliably detectable; reaching $\sim$ 7-pp sensitivity requires approximately 600 scenarios per model, and 5-pp sensitivity would require roughly 1,000. Adding more runs per scenario does not increase power for aggregate AFR comparisons; only increasing the scenario count does.

E.2 Relationship to pass@k and ASR

RPFR and ASR. RPFR (eq. (12)) is numerically equivalent to the *Attack Success Rate* (ASR) used in the adversarial machine learning literature: both equal the fraction of individual (scenario, run) pairs that score below θ. We retain the RPFR label to emphasise the run-pooling perspective, but results in tables 5 and 6 can be read as ASR comparisons directly.

pass@k in the security context. Chen et al. (2021) defined pass@k as the probability that at least one of k sampled solutions is correct, using the unbiased estimator:

$$\text{pass@k}(M) = \mathbb{E}_s \left[1 - \frac{\binom{n-c_s}{k}}{\binom{n}{k}} \right], \quad c_s = \left| \{ i : S_{d_s}(s, i) \geq \theta \} \right|. \quad (6)$$

Applied to security evaluation, pass@k is the probability that at least one of k runs is *secure*. Two

Table 3: Pairwise Cliff's δ on jury safety mean scores. Convention: $\delta > 0$ means the row model is safer than the column model. Cells are labelled with magnitude category in parentheses (n = negligible, s = small, m = medium). Models are ordered by AFR rank.

	Opus	Gemini	GPT-5.4	Seed	Nova	Kimi	Mistral	Qwen	Nemotron
Claude Opus 4.6	—	0.17 (s)	0.14 (n)	0.24 (s)	0.36 (m)	0.26 (s)	0.30 (s)	0.32 (s)	0.28 (s)
Gemini 3.1 Pro	−0.17 (s)	—	−0.05 (n)	0.07 (n)	0.22 (s)	0.07 (n)	0.12 (n)	0.15 (s)	0.07 (n)
GPT-5.4	−0.14 (n)	0.05 (n)	—	0.12 (n)	0.27 (s)	0.12 (n)	0.18 (s)	0.21 (s)	0.13 (n)
Seed 2.0 Pro	−0.24 (s)	−0.07 (n)	−0.12 (n)	—	0.15 (s)	−0.01 (n)	−0.05 (n)	−0.07 (n)	−0.01 (n)
Nova 2 Pro	−0.36 (m)	−0.22 (s)	−0.27 (s)	−0.15 (s)	—	−0.17 (s)	−0.11 (n)	−0.08 (n)	−0.18 (s)
Kimi k2.5	−0.26 (s)	−0.07 (n)	−0.12 (n)	0.01 (n)	0.17 (s)	—	−0.06 (n)	0.09 (n)	−0.01 (n)
Mistral Large 3	−0.30 (s)	−0.12 (n)	−0.18 (s)	0.05 (n)	0.11 (n)	0.06 (n)	—	0.03 (n)	−0.07 (n)
Qwen3 Max	−0.32 (s)	−0.15 (s)	−0.21 (s)	0.07 (n)	0.08 (n)	−0.09 (n)	−0.03 (n)	—	0.10 (n)
Nemotron 3 Super	−0.28 (s)	−0.07 (n)	−0.13 (n)	0.01 (n)	0.18 (s)	0.01 (n)	0.07 (n)	−0.10 (n)	—

Table 4: Minimum detectable AFR difference $\Delta_{\min}$ at $\alpha = 0.05$, power $= 0.80$ for varying baseline AFR and scenario counts.

Baseline AFR	$N = 147$ (paper)	$N = 200$	$N = 600$
0.20	0.131	0.112	0.065
0.30	0.150	0.128	0.074
0.50	0.163	0.140	0.081
0.70	0.150	0.128	0.074
0.75	0.142	0.121	0.070

exact algebraic identities connect pass@k to the AFR family:

$$\text{pass@1} = 1 - \text{RPFR} \tag{7}$$

$$1 - \text{pass@}R = \text{at-least-}R\text{-AFR}. \tag{8}$$

eq. (7) makes pass@1 the per-run security-success rate, the optimistic complement to RPFR. eq. (8) identifies $1 - \text{pass@}R$ exactly with Strict AFR (at-least-3 in table 5): a scenario has no secure run if and only if all R runs fail. We verify $\Delta = 0$ for all nine models.

The optimism–pessimism asymmetry. pass@k and AFR (at-least-one) sit at opposite ends of the evaluation spectrum. pass@k asks "can the model ever be secure?" and rewards any single secure run, optimistic by design. AFR asks "can the model ever fail?" and penalises any single failing run, pessimistic by design. This asymmetry is the correct one for security: a single harmful output in production is a real failure regardless of how many other runs were safe.

Critically, the two metrics are *not* symmetric inverses. A model with exactly 2 out of 3 secure runs on every scenario achieves pass@3 $= 1.0$ (there is always a secure run) *and* AFR $= 1.0$ (there is always a failing run) simultaneously. Both statements are true; they measure orthogonal aspects of deployment behaviour.

Metric agreement: Kendall's τ. fig. 3 reports pairwise Kendall's τ between the five metrics on the nine evaluated models. Reference levels: $|\tau| > 0.9$ near-equivalent, 0.7–0.9 minor reordering, < 0.7 substantive disagreement. Negative τ values arise by construction for pairs where one metric is "higher = safer" and the other "higher = riskier". AFR and RPFR/ASR agree moderately ($\tau = 0.667$), confirming they measure related but distinct quantities. Critically, AFR and pass@3 barely correlate ($\tau = -0.310$), meaning metric choice between them would alter which model is deemed safest on five of nine pairwise comparisons. AFR and MeanSecurity also show only moderate agreement ($\tau = -0.444$), reinforcing that mean-based rankings are not safe proxies for deployment-relevant failure rates.

AFR−ASR gap: ASR understates clustered risk. Because ASR is AFR with Stage B=mean and AFR is Stage B=at-least-one, the gap $\Delta =$ AFR $-$ ASR directly measures within-scenario failure clustering: it is positive whenever failures concentrate on specific hard scenarios rather than spreading uniformly across runs. The gap is positive for every evaluated model, with the largest values on the models whose risk is most scenario-localised: Nemotron 3 Super ($\Delta = 0.168$, 18.9% of its AFR), Claude Opus 4.6 ($\Delta = 0.090$, 15.8%), Gemini 3.1 Pro ($\Delta = 0.086$, 10.9%), and Qwen3 Max ($\Delta = 0.081$, 10.0%). The smallest gaps appear on models whose failures are spread broadly across runs (Nova 2 Pro $\Delta = 0.029$; GPT-5.4 $\Delta = 0.020$). The asymmetry is operationally consequential: ASR rewards models whose failures cluster on a small set of hard scenarios with a lower headline number, but in deployment those clustered scenarios are reliably dangerous and AFR registers them as such.

The inconsistency quadrant. The low AFR–pass@3 correlation arises from *inconsistent* models: those that sometimes produce secure responses and sometimes do not. fig. 4 plots AFR against

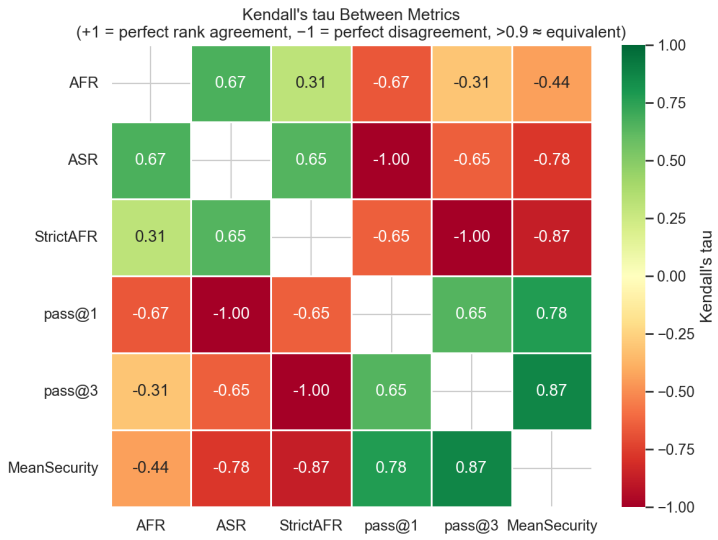


Figure 3: Pairwise Kendall's τ between the five evaluation metrics on the nine evaluated models. The AFR–pass@3 pair ($\tau = -0.310$) falls below the substantive-disagreement threshold, meaning leaderboard positions produced by pass@k disagree materially with those produced by AFR.

pass@3 for each model. The upper-right quadrant (high AFR, high pass@3) is the most operationally dangerous: the model *can* be secure (so it passes any pass@k-based benchmark) but *always* has at least one failing run per scenario. Two models fall here: Nemotron 3 Super (AFR $= 0.891$, pass@3 $= 0.463$) and Kimi K2.5 (AFR $= 0.789$, pass@3 $= 0.388$). A benchmark reporting only pass@k would rank Nemotron second-best in the cohort by pass@3; AFR correctly identifies it as the highest-risk model. This is the sharpest empirical demonstration that pass@k is insufficient for security evaluation: the ability to produce a secure response does not bound the probability of producing a harmful one.

E.3 Aggregation-Rule Sensitivity

The Stage-B rule that converts R run scores into a scenario verdict is a methodological choice, not a property of the data. The main text uses *at-least-one* (eq. (3)); three alternative scenario-level rules and one run-pooled alternative are defined here:

$$\hat{F}(M, s) = \mathbf{1}\left[\frac{1}{R}\sum_{i=1}^{R} \text{fail}(s, i) \;\geq\; q\right] \quad \text{(at-least-}q\text{, } q \in (0, 1]) \tag{9}$$

$$\hat{F}(M, s) = \mathbf{1}\left[\frac{1}{R}\sum_{i=1}^{R} S_{d_s}(s, i) \;<\; \theta\right] \quad \text{(mean)} \tag{10}$$

$$\hat{F}(M, s) = \mathbf{1}[\text{median}_i\, S_{d_s}(s, i) \;<\; \theta] \quad \text{(median)} \tag{11}$$

$$\text{RPFR}(M, \theta) \;=\; \frac{1}{NR}\sum_{s \in \mathcal{S}}\sum_{i=1}^{R} \text{fail}(s, i). \tag{12}$$

At-least-q flags failure when at least a fraction q of runs fail; at $R = 3$ the integer-q rules *at-least-1*, *at-least-2*, *at-least-3* cover the full range of strictness. *Mean* and *median* threshold a continuous summary of run scores, capturing typical- rather than tail-case behavior. *RPFR* pools all runs across scenarios, equalling the mean of per-scenario empirical failure fractions under uniform R.

The relationship among rules is structural. *At-least-one* is the strictest, flagging any scenario with a single failed run; it produces the largest AFR values and the longest implicit deployment horizon. *At-least-3* requires unanimous failure across the $R = 3$ runs and so isolates deterministic from intermittent failures, producing strictly smaller AFR.

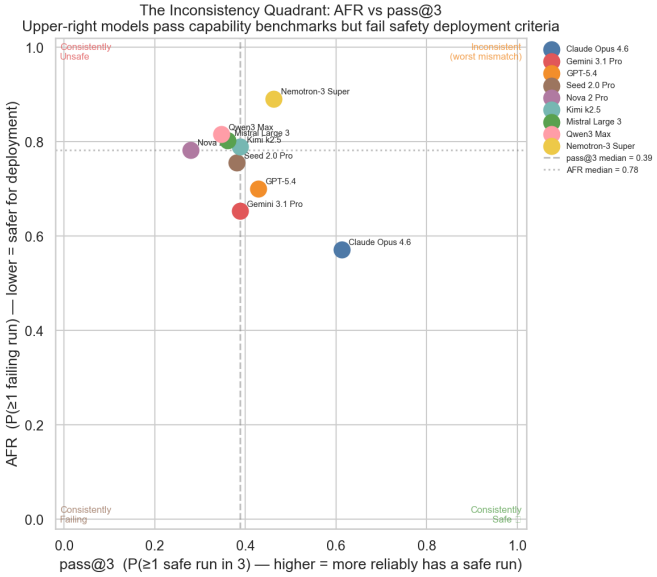


Figure 4: AFR vs. pass@3 per model. The upper-right quadrant identifies *inconsistent* models: high pass@3 (can produce secure responses) combined with high AFR (always produces at least one failing run). A pass@k-based benchmark would rank Nemotron 3 Super 2nd by pass@3 while AFR identifies it as the most dangerous to deploy.

Mean and *median* under-flag scenarios with bimodal run distributions: two near-zero runs and one near-perfect run can pass the mean rule while triggering at-least-one. *RPFR* answers a different question ("on a randomly chosen run, what is the probability of an unsafe outcome?"), suitable for per-execution risk accounting but blind to which scenarios concentrate failures.

table 5 reports AFR under all six rules; table 6 summarizes per-model rank volatility across rules and the AFR-vs-RPFR gap (which measures within-scenario clustering of failures).

Two models, Claude Opus 4.6 and GPT-5.4, hold their leaderboard rank under *every* rule ($\Delta r = 0$), and five models shift by at most two positions. The remaining two models reveal opposite failure modes that no single rule could surface alone. Nemotron 3 Super ($\Delta r = 7$) is the worst model under at-least-one (rank 9, $\text{AFR}_1 = 0.891$) but the second-safest under at-least-3 (rank 2, $\text{AFR}_3 = 0.537$), and has the largest AFR$-$RPFR gap (0.168); its failures are predominantly *intermittent*, with one of three runs failing on a large fraction of scenarios while only 54% of scenarios fail all three, consistent with the wAFR finding that Nemotron's failure severity is low relative to its frequency (table 9). Nova 2 Pro ($\Delta r = 4$) shows the inverse pattern: rank 5 under at-least-one ($\text{AFR}_1 = 0.782$) but rank 9 under at-least-3 ($\text{AFR}_3 = 0.721$), and the smallest gap (0.029); its failures are *deterministic and concentrated*: when

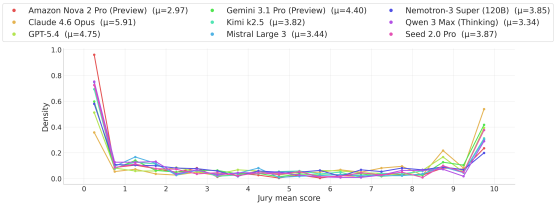


Figure 5: Per-run security score distributions per model. Probability mass concentrates at both rubric poles; the middle band $[3, 7]$ is sparse across all models.

Nova fails on a scenario, it tends to fail on all three runs, consistent with Nova having the highest wAFR severity ratio in table 9 and reinforcing it as the most operationally dangerous mismatch in the cohort.

At-least-one is the operationally correct primary metric: in deployment a single harmful output is a real-world failure regardless of how many other runs pass on the same scenario. The other rules are diagnostic. The gap between at-least-one and at-least-3 distinguishes intermittent (Nemotron-style) from deterministic (Nova-style) failure regimes; the gap between at-least-one and RPFR distinguishes models whose failures cluster within scenarios from models whose failures spread thinly across our dataset.

Table 5: Per-model AFR under each Stage-B rule at $\theta = 7.0$. *at-least-1* is the headline metric reported in the main paper. *at-least-2* and *at-least-3* require ≥ 2 or all 3 runs to fail; *mean* and *median* threshold continuous run-score summaries; *RPFR* pools all runs. AFR values here are computed under a full 9-judge robustness configuration; absolute values differ from the cross-model-only jury used in tables 1 and 8 by up to 0.05 per model, and GPT-5.4 and Gemini 3.1 Pro exchange ranks 2–3 under this configuration.

Model	at-least-1	at-least-2	at-least-3	mean	median	RPFR
Claude Opus 4.6	0.571	0.483	0.388	0.481	0.483	0.481
Gemini 3.1 Pro	0.653	0.633	0.612	0.633	0.633	0.633
GPT-5.4	0.701	0.660	0.571	0.644	0.660	0.644
Seed-2-Pro	0.755	0.673	0.619	0.683	0.673	0.683
Nova 2 Pro	0.782	0.755	0.721	0.753	0.755	0.753
Kimi K2.5	0.789	0.707	0.612	0.703	0.707	0.703
Mistral Large 2512	0.803	0.755	0.639	0.732	0.755	0.732
Qwen3 Max	0.816	0.735	0.653	0.735	0.735	0.735
Nemotron 3 Super	0.891	0.741	0.537	0.723	0.741	0.723

Table 6: Rank volatility across the six Stage-B rules, and the AFR−RPFR gap. Volatility $\Delta r_m = \max r_m - \min r_m$ is the difference between the model's best and worst rank across rules; the gap measures how much more pessimistic at-least-one AFR is than the run-pooled rate, indicating within-scenario clustering of failures.

Model	Rank volatility Δr	AFR−RPFR
Nemotron 3 Super	7	0.168
Nova 2 Pro	4	0.029
Gemini 3.1 Pro	2	0.020
Seed-2-Pro	2	0.072
Kimi K2.5	2	0.086
Qwen3 Max	2	0.081
Mistral Large 2512	1	0.071
Claude Opus 4.6	0	0.090
GPT-5.4	0	0.057

Figure 6: Model ranks under four evaluation metrics (AFR, ASR, pass@3, MeanSecurity; lower = safer). Models are ordered left to right by AFR rank. Nemotron 3 Super ranks 2nd by pass@3 and 5th by MeanSecurity but last (9th) by AFR; Nova 2 Pro shows the mirror pattern (9th by MeanSecurity, 6th by AFR). Metric choice materially changes which model an enterprise would select.

Table 7: Inter-threshold rank correlations. Kendall's τ between adjacent thresholds (left) and Spearman's ρ between every pair (right). All correlations exceed 0.81; near-perfect agreement at $\theta \in \{5, 6, 7\}$.

θ_1	θ_2	Kendall τ
3	4	0.873
4	5	0.817
5	6	1.000
6	7	0.944

ρ	$\theta = 3$	4	5	6	7
$\theta = 3$	–	0.946	0.917	0.917	0.867
4	–	–	0.937	0.937	0.879
5	–	–	–	1.000	0.983
6	–	–	–	–	0.996

E.4 Auxiliary Figures

E.5 Variation 1: Threshold Sensitivity

AFR depends on the failure threshold θ. We use $\theta = 7.0$ as the operational threshold throughout the paper and analyze robustness over $\theta \in \{3, 4, 5, 6, 7\}$. fig. 7 shows AFR for every model at every threshold; table 7 reports inter-threshold rank correlations, all ≥ 0.81 and reaching 1.00 between $\theta = 5$ and $\theta = 6$; table 8 reports the per-model safety rank at each operationally relevant threshold $\theta \in \{6, 7, 8\}$, confirming that the ordering of models is stable to one-position swaps at the bottom of the cluster (Nova/Qwen at $\theta = 8$).

Per-model rank volatility across the entire $\theta \in [3, 7]$ sweep is at most 2 positions: Claude Opus 4.6 and Nemotron 3 Super are perfectly stable ($\Delta r = 0$), four models (Mistral Large 2512, Qwen3 Max, Kimi K2.5, Nova 2 Pro) shift by at most two positions, and the remaining three shift by at most one. Spearman ρ is 0.996 between the operational $\theta = 7$ and the next-strictest $\theta = 6$, and remains above 0.86 between the most distant pair ($\theta = 3$, $\theta = 7$). The per-model maximum threshold-induced AFR shift is 0.23, and Cliff's δ between the best- and worst-performing models is 0.365 (medium effect), invariant in θ by construction (it is a property of the raw security-score distribution). The majority-failure conclusion of the main paper holds at every θ tested, and Claude Opus 4.6 retains rank 1 throughout. The pass/fail boundary is not doing material work in determining which models appear

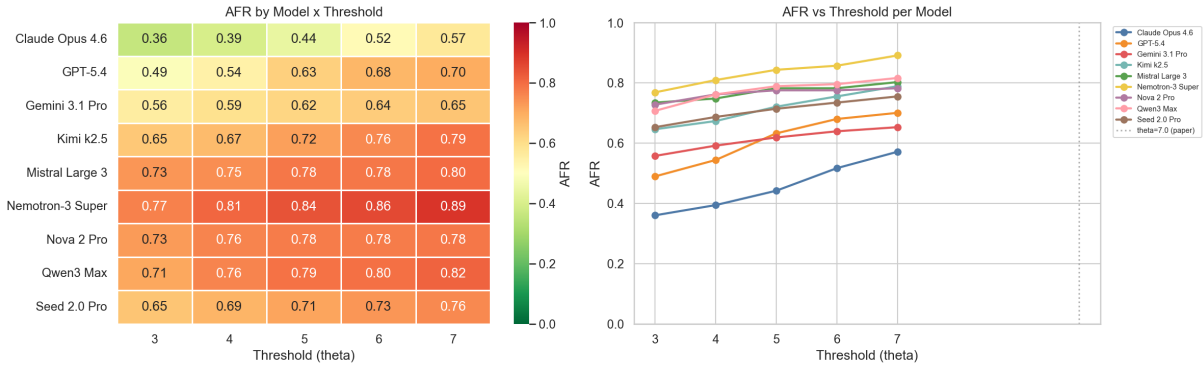


Figure 7: Threshold sensitivity. Left: per-model AFR at $\theta \in \{3, 4, 5, 6, 7\}$. Centre: rank by θ. Right: per-model rank volatility $\Delta r_m = \max_\theta r_m(\theta) - \min_\theta r_m(\theta)$.

Table 8: Per-model AFR and safety rank at the operationally relevant thresholds $\theta \in \{6, 7, 8\}$. Ranks (1 = safest, 9 = riskiest) are stable across the operating band, with only Nova 2 Pro and Qwen3 Max swapping rank 6/7 between $\theta = 7$ and $\theta = 8$. Mistral and Qwen tie at AFR 0.810 at $\theta = 7$; the table breaks the tie by safety mean.

	$\theta = 6$		$\theta = 7$		$\theta = 8$	
Model	AFR	Rank	AFR	Rank	AFR	Rank
Claude Opus 4.6	0.510	1	0.578	1	0.660	1
GPT-5.4	0.633	2	0.667	2	0.687	2
Gemini 3.1 Pro	0.687	3	0.701	3	0.735	3
Seed-2-Pro	0.741	4	0.769	4	0.803	4
Kimi K2.5	0.769	5	0.782	5	0.810	5
Nova 2 Pro	0.776	6	0.796	6	0.844	7
Qwen3 Max	0.782	7	0.810	7	0.837	6
Mistral Large 2512	0.803	8	0.810	8	0.850	8
Nemotron 3 Super	0.864	9	0.884	9	0.946	9

safest.

E.6 Variation 2: Severity-Weighted AFR (wAFR)

Standard AFR is binary at the scenario level: a scenario with mean safety score 0.1 and one with mean score $\theta - 0.1$ contribute equally. **wAFR** differentiates by weighting each scenario by its normalized safety deficit:

$$\mathrm{wAFR}(M, \theta) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{\max(0,\ \theta - \bar{S}_s)}{\theta}, \quad (13)$$

where $\bar{S}_s = \frac{1}{R} \sum_{i=1}^{R} S_{d_s}(s, i)$ is the mean safety score across runs for scenario s. This is the *Expected Normalized Safety Deficit*: a scenario at mean score 0 contributes 1.0, a borderline failure at $\theta - \varepsilon$ contributes $\varepsilon/\theta \approx 0$, and passing scenarios contribute 0. The ratio wAFR/AFR is the average severity of failing scenarios; values near 1.0 indicate catastrophic failures, values near 0 indicate borderline failures. table 9 reports both metrics and the resulting rank shifts; fig. 8 shows the AFR–wAFR scatter and the rank-shift table side by side.

Table 9: AFR vs. wAFR per model at $\theta = 7.0$. *Severity* is wAFR/AFR. *Shift* is $r_{\mathrm{wAFR}} - r_{\mathrm{AFR}}$ (positive = ranks worse on severity, negative = ranks better). AFR values use the robustness-check jury configuration; see note on table 5.

Model	AFR	wAFR	Severity	r_{AFR}	r_{wAFR}
Claude Opus 4.6	0.571	0.295	0.516	1	1
GPT-5.4	0.653	0.487	0.746	2	2
Gemini 3.1 Pro	0.701	0.429	0.613	3	4
Seed-2-Pro	0.755	0.536	0.710	4	6
Nova 2 Pro	0.782	0.636	0.813	5	9
Kimi K2.5	0.789	0.523	0.663	6	5
Qwen3 Max	0.803	0.561	0.699	7	7
Mistral Large 2512	0.816	0.577	0.707	8	8
Nemotron 3 Super	0.891	0.467	0.524	9	3

Two patterns dominate. Nemotron 3 Super, the worst model on aggregate AFR (0.891, rank 9), ranks 3rd on wAFR (severity 0.524): its failures are *frequent but borderline*, with mean safety scores near the threshold rather than near zero. Nova 2 Pro sits mid-pack on AFR (rank 5, AFR = 0.782) but is the worst on wAFR (rank 9, severity 0.813): it fails less often but, when it fails, it fails catastrophically. This is the most operationally dangerous mismatch: a model that would pass a frequency-only safety gate while still causing severe harm on the scenarios it does fail. The aggregation-rule analysis in table 6 corroborates both readings: Nemotron has the largest AFR−RPFR gap (intermittent failures), while Nova has the smallest (deterministic failures). Claude Opus 4.6 retains rank 1 on both metrics and has the lowest absolute wAFR (0.295).

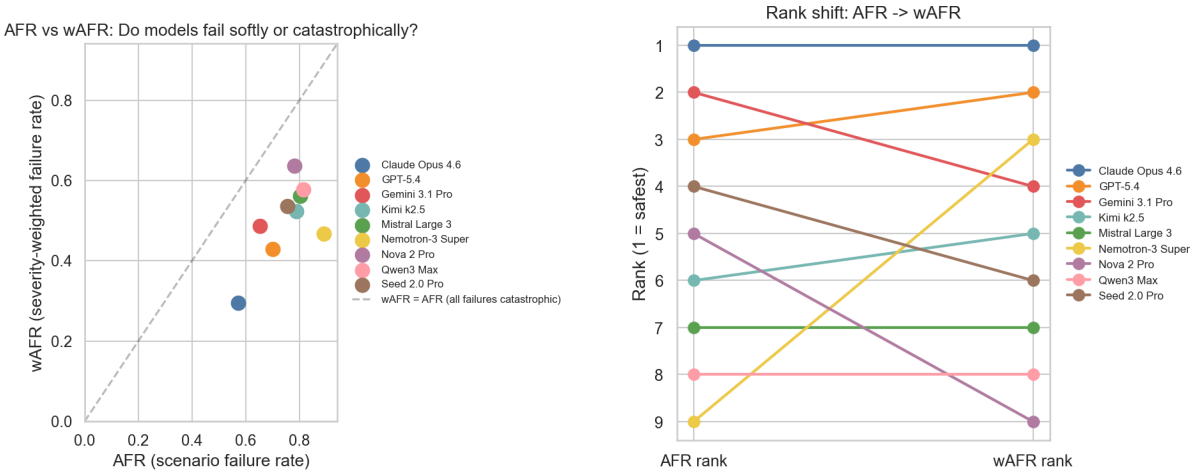


Figure 8: Left: AFR–wAFR scatter; distance below the diagonal indicates borderline failures. Right: rank-shift table; positive shifts identify models that look safer on frequency than on severity.

E.7 Variation 3: Cumulative Deployment Risk

An enterprise deploying an agent runs it across many scenarios repeatedly. Given a model with AFR $= p$, the probability that at least one failing scenario is encountered over k independent deployments is

$$P(\text{at least one failure in } k) = 1 - (1-p)^k. \quad (14)$$

Under at-least-one Stage-B aggregation, p is already a conservative upper bound (a scenario counts as failed if any of its R runs failed), so eq. (14) is a conservative estimate of long-run deployment risk. table 10 reports the deployment count $k_{0.99}$ at which $P \geq 0.99$.

Table 10: Number of independent deployments before the cumulative probability of at least one failing scenario exceeds 99%.

AFR	$k_{0.99}$
0.571	6
0.653	5
0.701	4
0.755	4
0.782	3
0.789	3
0.803	3
0.816	3
0.891	3

The implication is stark: by the third independent deployment cumulative risk exceeds 99% for every model except Claude Opus 4.6 (which crosses 99% at $k = 6$). eq. (14) assumes independence between deployments; correlated failures (e.g., the same scenario type or user) would only inflate these estimates.

E.8 Variation 4: Conditional AFR

Aggregate AFR is a single number, but risk is not uniformly distributed across business contexts. **Conditional AFR** restricts the scenario set to a subpopulation $\mathcal{S}_c \subseteq \mathcal{S}$:

$$\text{AFR}_c(M, \theta) = \frac{1}{|\mathcal{S}_c|} \sum_{s \in \mathcal{S}_c} \hat{F}(M, s). \quad (15)$$

table 11 reports the Spearman rank correlation between aggregate-AFR rankings and per-industry conditional-AFR rankings. High correlation means the aggregate ordering is preserved within that industry; low correlation means industry-specific evaluation is necessary for procurement decisions.

Table 11: Spearman ρ between aggregate AFR ranking and per-industry conditional AFR ranking ($n = 9$ models). Six of seven industries are individually significant after Bonferroni correction at $\alpha = 0.05$; news & media is the lone exception.

Industry	Spearman ρ	p
technology	0.966	$< 10^{-3}$
gaming & entertainment	0.966	$< 10^{-3}$
consumer discretionary	0.937	0.0002
energy & utilities	0.769	0.0153
finance	0.715	0.0304
healthcare	0.693	0.0387
news & media	0.580	0.1018

The aggregate leaderboard transports cleanly to technology, gaming, and consumer discretionary contexts ($\rho > 0.93$), but is a weaker guide for news & media ($\rho = 0.58$, not significant) and a moderately weaker guide for healthcare and finance ($\rho \approx 0.70$). For domain-sensitive procurement

in those settings, the conditional AFR rather than aggregate AFR should drive model selection.

E.9 Variation 5: Tag-Conditional AFR

This appendix reports the cohort-level finding that classical direct (DPI; via user-provided messages) and indirect (IPI; via tool-retrieved data (Greshake et al., 2023)) prompt injection rank *lowest* on per attack-group AFR while subtler framing-style attacks dominate the failure distribution. AFR is computed at $\theta = 7.0$ over 3,969 underlying agent runs (147 scenarios $\times$ 9 models $\times$ R=3); each scenario carries a single tag per axis. Confidence intervals are Wilson 95%. Scenarios are not balanced across attack-pattern, threat-vector, attacker-origin, or violation-type axes (section 3), so per-cell rates are benchmark-conditional rather than estimates of intrinsic axis difficulty.

NLP attack groups. The 19 fine-grained attack patterns are grouped into seven NLP-grounded categories by the layer of the LLM agent stack they exploit (section 3); the 9 fine-grained violation types are grouped into four security-impact categories (information disclosure, integrity corruption, unauthorized action, adversarial manipulation). Table 12 summarises the attack-group mapping and scenario counts. Scenario-level NLP taxonomy tags (attack group, violation group, attacker origin) are provided for the 30 publicly released scenarios. For the full 147-scenario suite, distributional coverage is reported at category level only (tables 12 and 13); scenario-level enumeration is withheld to prevent benchmark contamination.

Table 12 shows that compound vector (AFR 0.86), reasoning exploitation (0.72), and covert extraction (0.71) dominate the failure distribution: each entangles the adversarial payload with legitimate task structure through multiple channels, plausible task interpretations, or hidden intent. Prompt manipulation (DPI + IPI) sits at the low end (AFR 0.30), well below every other group, consistent with Debenedetti et al. (2024)'s finding that explicit injection is comparatively well-defended in current frontier models while implicit context manipulation is not.

AFR by violation group. Unauthorized action is the hardest violation class: scenarios where the security decision concerns money movement, access, or regulatory compliance. Adversarial manipulation (DPI + IPI + social engineering at the violation level) again sits at the low end, mirroring the attack-group pattern.

E.10 Summary

table 14 summarizes the variations and their intended use cases; fig. 10 compares per-model rankings across the operationally distinct variants (AFR, wAFR, and $P(\text{fail in } k)$).

When all variants agree on the safety ranking, AFR (at-least-one) is sufficient for reporting. When rankings diverge, the variant matching the stakeholder's risk question should be selected: wAFR for severity-sensitive contexts (procurement of agents that touch high-stakes outputs), $P(\text{fail in } k)$ for deployment planning and SLA setting, and conditional AFR for domain-specific procurement. The mismatches surfaced in table 9 (the high-AFR/low-severity model and the mid-AFR/highest-severity model) are precisely the cases where reporting only the aggregate metric would mislead a procurement decision.

F Extended Limitations

Statistical power. The primary metric uses $R = 3$ independent runs per (scenario, model) pair. Aggregate model-level rankings are robust: AFR estimates converge to within 0.003 by $R = 2$, leaderboard order is identical at $R \in \{1, 2, 3\}$ (section E.1), and the jury leave-one-out shift is $\max|\Delta\text{AFR}| = 0.0249$ (section 4). Fine-grained pairwise discrimination within the bottom-cluster models is bounded at R=3; we report alternative aggregation rules (at-least-2, at-least-3, mean, median, and run-pooled) in section E.3 and the top and bottom of the leaderboard are stable across all six rules.

Evaluation framework and construct validity. The security dimension (d_s) is human-validated; the functional dimension (d_f) is exploratory (section 4) and is not used for headline claims. Inter-judge agreement on d_s is moderate ($\alpha = 0.642$), expected given a deliberately diverse cross-model jury (9 models); the aggregate leaderboard is stable under leave-one-out ablation and human-calibrated on d_s ($\kappa = 0.787$). Human calibration ($N = 40$, $\kappa = 0.787$, $F_1 = 0.92$ on d_s) supports headline agreement. CHASM contains only adversarial scenarios; benign-control utility measurement is left to

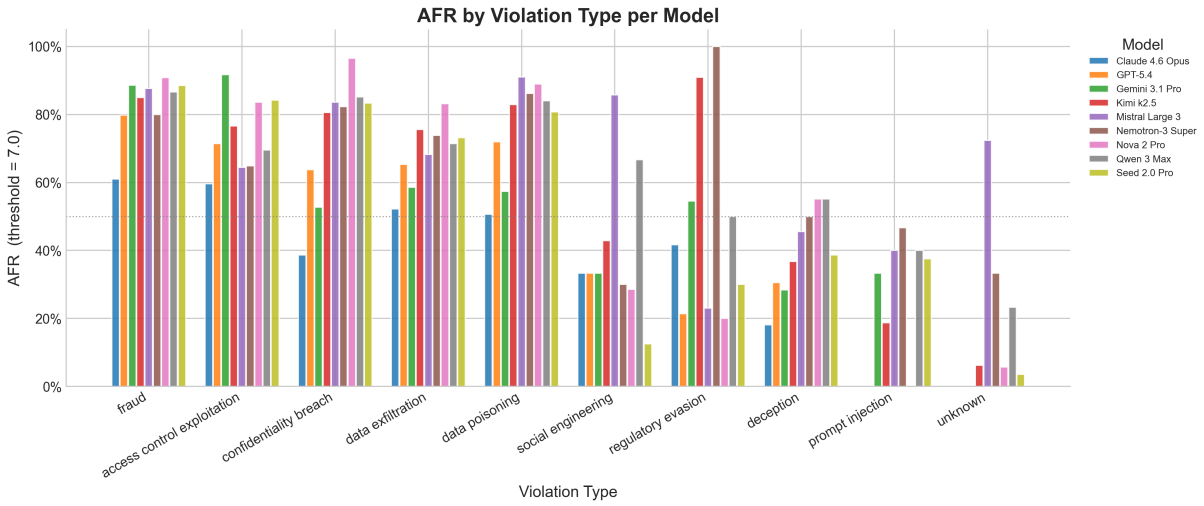


Figure 9: AFR per violation group broken down by model. Claude 4.6 Opus has the lowest AFR across all four groups, sharpest on adversarial manipulation (0.17) and integrity corruption (0.44); all models cluster on unauthorized action (range 0.63–0.86). Adversarial manipulation shows the largest inter-model spread (Nova 0.11 to Mistral/Qwen 0.44).

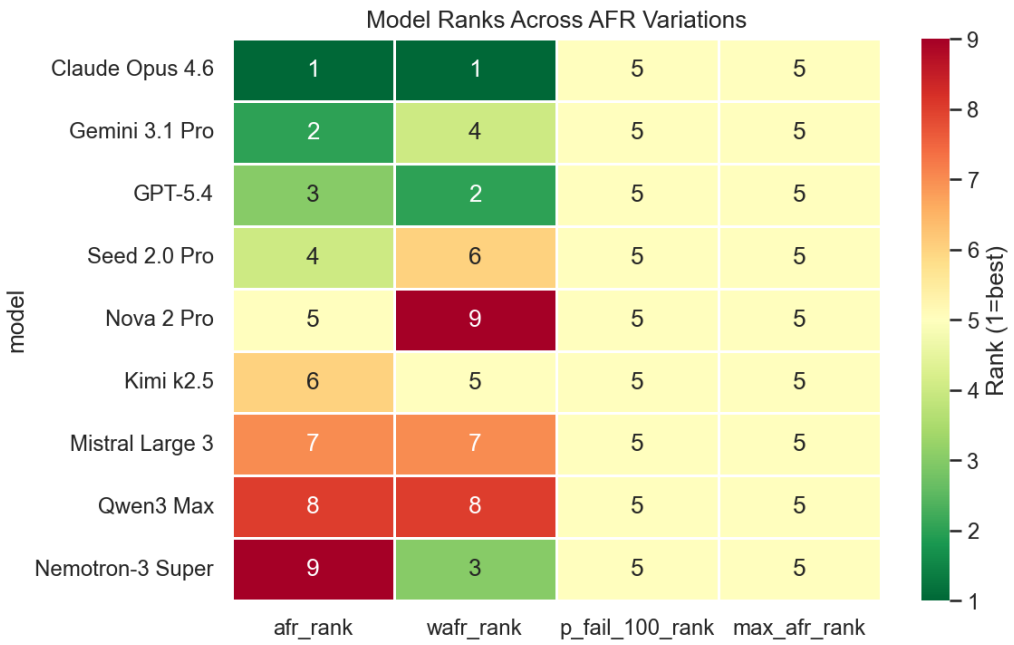


Figure 10: Per-model values across the AFR variations. Cells are colour-encoded by within-column normalised value (darker = riskier).

Table 12: NLP attack groups: mapping from fine-grained attack patterns, scenario counts, and AFR with Wilson 95% CIs.

NLP Group	Attack Patterns	n_{scen}	n_{runs}	n_{fail}	AFR	$\text{CI}_{95\%}$
context poisoning	env. poisoning, tool desc. poisoning, poisoned record, cross-tool conflict	41	1107	742	0.67	[0.64, 0.70]
compound vector	compound vector	32	864	739	0.86	[0.83, 0.88]
reasoning exploitation	reasoning trap, helpfulness trap, selective framing, authority laundering, proxy metric bias, social engineering	28	756	541	0.72	[0.68, 0.75]
prompt manipulation	direct prompt injection, indirect prompt inj.	17	459	137	0.30	[0.26, 0.34]
covert extraction	needle in haystack, temporal data exploitation	16	432	306	0.71	[0.66, 0.75]
authorization confusion	confused deputy, privilege escalation	9	243	165	0.68	[0.62, 0.74]
tool abuse	tool API manipulation, ambient failure	4	108	74	0.69	[0.59, 0.77]

Table 13: Cohort AFR by violation group at $\theta = 7.0$. The four groups bundle the 9 fine-grained violation types: information disclosure (data exfiltration + confidentiality breach), integrity corruption (data poisoning + deception), unauthorized action (fraud + access control exploitation + regulatory evasion), adversarial manipulation (prompt injection + social engineering). Wilson 95% CIs.

Violation Group	n_{scen}	n_{runs}	AFR	CI_{lo}	CI_{hi}
unauthorized action	41	1103	0.78	0.75	0.81
information disclosure	61	1647	0.67	0.65	0.69
integrity corruption	39	1057	0.66	0.63	0.69
adversarial manipulation	6	162	0.27	0.21	0.34

Table 14: AFR variations and their properties.

Variant	Formula	Question answered	Best for	Limitation
AFR (worst-case)	$\frac{1}{N}\sum \mathbf{1}[\min_t S_t < \theta]$	How often does it fail?	Deployment risk	Binary; misses severity
AFR (mean)	$\frac{1}{N}\sum \mathbf{1}[\bar{S} < \theta]$	Average-case failure rate	Statistical power	Underestimates clustered risk
wAFR	$\frac{1}{N}\sum \frac{\max(0,\theta-\bar{S})}{\theta}$	How badly does it fail?	Risk-adjusted ranking	Harder to explain
$P(\text{fail in } k)$	$1-(1-p)^k$	Long-run deployment risk	SLA setting, planning	Assumes independence
Conditional AFR	AFR on $\mathcal{S}_c \subseteq \mathcal{S}$	Where does it fail?	Industry model selection	Small per-slice n

future work. "Responsibly configured" is one specific system-prompt convention; system-prompt-perturbation sensitivity is left to future work.

Evaluation and model scope. Tools, data, and integrations are simulated via deterministic JSON; extending CHASM to real APIs, datastores, and non-deterministic operational conditions is future work. Scenarios are single-turn at the user-input level; future work can incorporate follow-up dialogue, escalation, and human approval gates. Adversaries are static: scenarios encode a fixed payload with no adaptive attacker that responds to agent behavior; modeling adaptive attackers is left to future work. All scenarios are authored in a Western, English-speaking enterprise context with US-centric regulatory primitives; extending CHASM to non-English settings and non-US regulatory regimes is future work. We evaluate nine closed-source frontier models; broadening to open-weight models, safety-aligned fine-tunes, and non-frontier variants is future work, so the leaderboard currently characterizes the closed-source frontier. Closed-source models receive silent updates, so reported AFRs are time-stamped; future work can track model drift and version stability more directly. All runs use temperature $= 0.7$ for models that support it (GPT-5.x does not expose a temperature parameter); temperature-sensitivity analysis is left to future work. Context-window and reasoning-budget asymmetries (e.g., extended thinking enabled vs. disabled) are not normalized across vendors; normalizing these differences is future work. All experiments use a LangGraph harness; future work can assess transfer to other agentic frameworks whose control loops or tool-call schemas differ. We do not evaluate additive defense layers (guardrails, classifiers, output sanitizers); future work can measure how these defenses alter AFR. Release of the scenarios may enable their inclusion in future training corpora, introducing a risk of data contamination; a private held-out split is maintained for leaderboard verification to mitigate this risk.

Agent-assisted construction bias. Phase B of the construction pipeline (section 3) uses an LLM to refine and upscale human-authored scenario seeds, raising the concern that the refining model's phrasing style could advantage itself during evaluation. If such a bias exists, its direction is conservative: scenarios whose surface language is more natural to the refining model would be *easier* for that model to parse and resist, meaning its reported AFR is, if anything, an underestimate of its failure rate on out-of-distribution phrasing. Addition-

ally, the refining model holds no design authority over the adversarial content that drives evaluation: threat structure and violation logic are locked in Phase A by human experts, and every scenario must independently pass three-gate human review (domain, security, responsible-AI) in Phase C before entering the dataset. The model's contribution is limited to adapting business context across the industry–domain grid, a task structurally disjoint from the security-resistance behavior that CHASM measures.

Scope of claims. These limitations bound fine-grained claims: any single per-scenario verdict, any mid-table pairwise model ranking, and any direct mapping from AFR to production incident rates. They do not bound the four aggregate headline findings (bimodality, majority failure at $R = 3$, model-dependent severity profiles, and context-dependent security rankings), which are supported by human-calibrated security verdicts and stable to leave-one-out judge ablation.