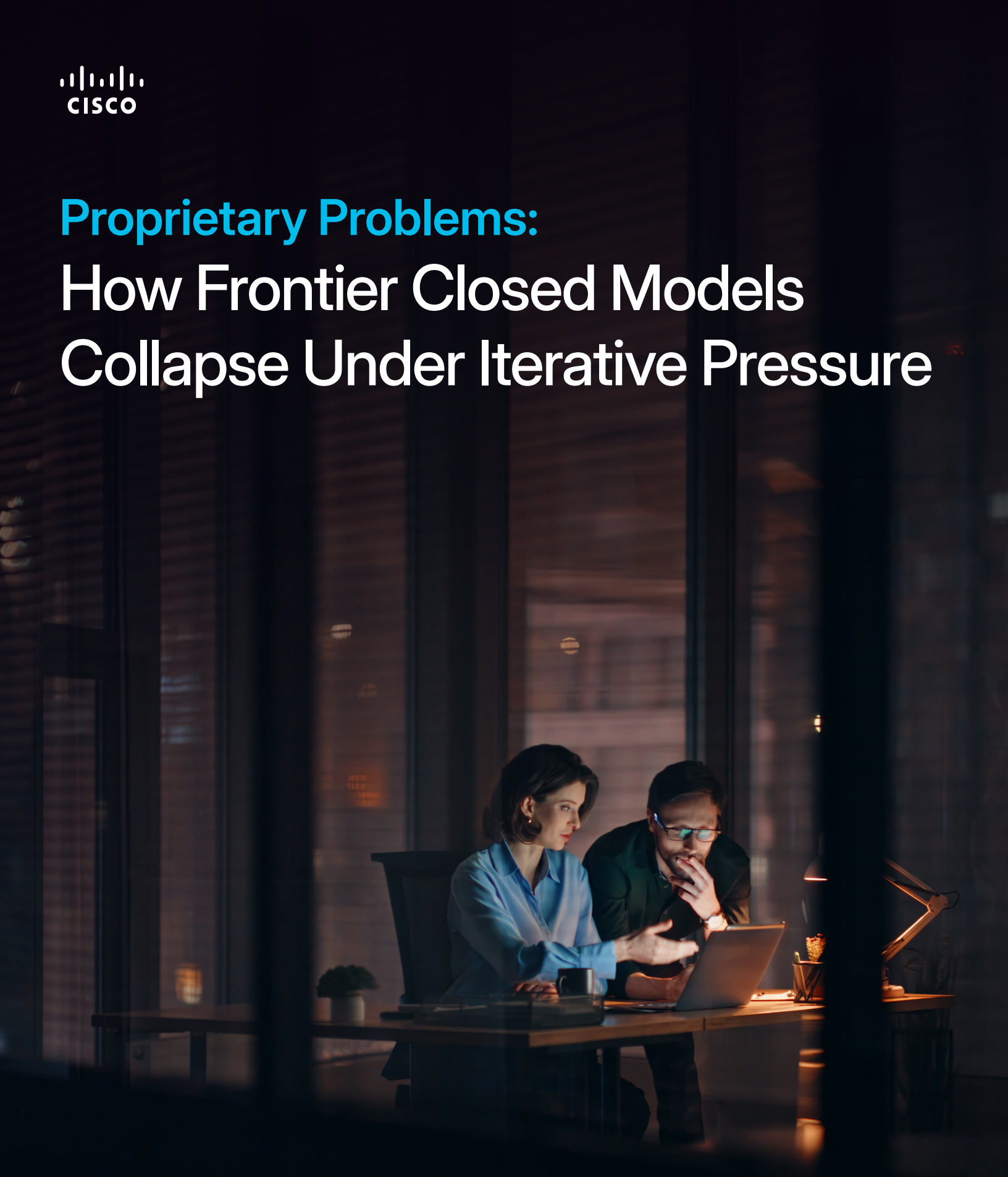
Proprietary Problems:
How Frontier Closed Models Collapse Under Iterative Pressure



Nicholas Conley and Amy Chang
Cisco Systems, Inc.
AI Threat Intelligence and Security Research Team

May 27, 2026

Executive Summary

Widely used safety benchmarks for frontier AI models report a single attack success rate (ASR) from singleturn prompts—a convention that meta-analyses link to capability scaling rather than genuine safety progress. The findings in this report suggest that single-turn ASR alone may be an insufficient basis for safety and security decisions: across this cohort, paired single-turn and multi-turn evaluation produces a different model ordering, a different failure map, and a different tail-risk picture than either regime alone. This report pairs single-turn and multi-turn evaluation across 15 closed/proprietary frontier models (30,090 singleturn prompts; 6,986 multi-turn attacks) and finds that the two regimes produce different model orderings, different failure maps, and different tail-risk profiles. Multi-turn evaluation matters for one primary reason: it is where attackers operate. Real adversaries iterate, reframe refusals, decompose tasks across turns, adopt personas, and escalate gradually. To address this reality, this report translates that standard into three concrete evaluation rituals (strategy-stratified release reporting, top-surface deployment gating, and cross-regime gap review) that organizations can consider adopting.

Across the 15-model cohort, multi-turn ASR spans 7.89%–88.30%—a wider separation than the single-turn range of 2.19%–64.91%. Cross-regime deltas run in both directions: Gemini 3 Pro and Grok 4.1 Fast NR rise by more than 54 pp under iterative attack, while all three Amazon Nova variants move the opposite way. Eight of 15 models exceed a 15 pp absolute gap between regimes. A single configuration flag—enabling reasoning mode on Grok 4.1 Fast—is associated with a 44.83 pp drop in multi-turn ASR under an otherwise identical evaluation, suggesting that deployment-time choices invisible to downstream users can meaningfully affect safety outcomes. No model in the cohort eliminates residual iterative risk; even models with very low single-turn ASR—Claude Opus 4.5 (2.19%) and GPT-5.4 (2.74%)—reach 11.16% and 24.68% multi-turn ASR, respectively.

Decomposing further, strategy-family slices reveal 79.51–89.25 pp cross-model spreads within each strategy, indicating that strategy labels primarily stratify which models separate from one another rather than shifting cohort-average difficulty. Single-turn weakness concentrates in a small set of procedures (Imposter AI, Soft Paraphrase, System Prompts) and content types (Hate Speech, Profanity, Specialized Advice)—a pattern that supports phased, high-leverage hardening rather than uniform expansion.

Taken together, these results convey that the reporting standard for frontier-model safety evaluation could evolve to include paired-regime ASR, strategy-family breakdowns, and explicit slice-support labeling. The report translates that standard into three concrete evaluation rituals that organizations can consider adopting: (1) publish ASR by strategy family on every model release; (2) gate deployments on regressions in top-3 procedures and content types; and (3) flag any model with a >15 pp absolute cross-regime gap for manual review.

This analysis covers the 15 in scope models listed in the introduction; the Cisco LLM Security leaderboard extends coverage to additional models and is updated on a rolling basis.

Key Takeaways

The existing benchmark ecosystem systematically understates iterative risk. Widely used safety benchmarks (HarmBench, AILuminate, TrustLLM) are single-turn by design, and meta-analyses show many correlate more strongly with capability than with safety itself. This report's paired regime findings corroborate at scale what the multi-turn literature has shown in smaller settings: single-turn and multi-turn evaluation measure overlapping but distinct properties (See Section 1.1).

Single-turn ASR is not a proxy for multi-turn resilience. Cross-regime deltas range from −34.74 pp (Nova Lite) to +55.25 pp (Gemini 3 Pro); eight of 15 models exceed a 15 pp absolute gap, in both directions. In this cohort, GPT-5.4 moves from 2.74% to 24.68% (a ninefold increase) and Gemini 3 Pro from 18.10% to 73.35%—shifts invisible to any single-turn benchmark (See Figure 2).

Configuration flags can move multi-turn ASR by tens of percentage points. Grok 4.1 Fast drops 44.83 pp in multi-turn ASR when reasoning is enabled, under an otherwise identical evaluation. This argues for a disclosure norm in which providers document the safety-relevant effects of deployment time configuration alongside capability benchmarks (See Table 2, Section 4.2).

Strategy family is where cross-model separation lives. Within-strategy spreads of 79.51–89.25 pp mean aggregate multi-turn ASR can hide actionable per-strategy variation; even models with low aggregate multi-turn ASR warrant strategy-stratified monitoring (See Table 3, Figure 4).

Single-turn failures are concentrated, not distributed. Imposter AI, Soft Paraphrase, and System Prompts lead procedure risk; Hate Speech, Profanity, and Specialized Advice lead content-type risk—enabling phased, high-leverage hardening rather than uniform expansion (See Figure 8).

No model in the cohort demonstrates safety sufficient to eliminate residual iterative risk. Every frontier model tested—including the most recent releases from OpenAI, Anthropic, Google, Amazon, and xAI—exhibits non-trivial multi-turn ASR. The lowest multi-turn exposure in the cohort (Nova 2 Lite, 7.89%) still represents meaningful residual risk, and this pattern is consistent with recent multi-turn red-teaming literature (See Section 4.2).

Three operational rituals translate these findings into practice. (1) Publish ASR by strategy family on every model release; (2) gate deployments on regressions in top-3 procedures and content types using a 3 pp threshold calibrated to the cohort's largest single-turn CI95 half-width; (3) flag any model with a >15 pp absolute cross-regime gap for manual review—a rule that surfaces eight of the 15 models in this cohort (See Section 4.1).

Regulatory frameworks are beginning to call for evaluation practices that current benchmarks do not yet fully address. Standards and regulatory entities such as National Institute of Technology and Standards (NIST) AI Risk Management Framework (RMF), the forthcoming Cyber AI Profile (IR 8596), and the European Union (EU)'s AI Act's High Risk provisions (Article 15) call for adversarial and robustness testing. For now, these measures do not specify the interaction regime, strategy decomposition, or slice-support labeling required for decision-grade assessment. Adopting the rituals above can contribute to effectively addressing these risks proactively to ensure the best safety outcomes. (See Section 1.1, Section 4.2).

1 Introduction

The process of training models (and closed weight models in particular) "make it difficult for developers to predict model capabilities and behavior," and have practical consequences on safety and security.[1] As these models continue to advance in capability, research continues to demonstrate their susceptibility to manipulation, despite advances in safety mitigations.[2] The data in this report expands on this story: safety is regime-dependent, strategy-sensitive, and configuration-contingent. When a model with a 2.19% headline ASR still fails 11.16% of multi-turn attacks, the narrative of robust safety requires significant qualification. This gap between single-turn benchmark scores and multi-turn measured behavior represents an information asymmetry that may affect regulators, enterprise buyers, and end users. Closing that gap should consider the kind of paired-regime, decomposed reporting this study advocates.

The central question is not only which models fail, but how exposure shifts when attacks become iterative. This report therefore prioritizes paired single-turn and multi-turn ASR analysis, then drills into strategy, content, and procedure structure to explain observed gaps.

The analysis is written from an operational security perspective but also provide strategic-level takeaways. A model with low single-turn ASR can still produce high residual risk if multi-turn trajectories, strategy families, or specific procedural surfaces remain weak. For that reason, interpretation focuses on structure of failure, not only aggregate means.

Scope and Limitations. This report measures base-model behavior under a curated adversarial evaluation regime. It does not characterize deployed-product behavior with system prompts, content filters, or custom orchestration. It is a snapshot study, not a longitudinal analysis, and ASR is one signal among many. Rank ordering can shift with refreshed runs, expanded attack volume, or model updates. These caveats are expanded in Section 5.

Metric definitions are provided once in Table 1. Throughout the report, ASR is the primary quantity of interest and resistance rates are reported as the complement for convenience. Models in scope (15 total; closed/proprietary, API-accessible frontier models without open weights):

- **OpenAI:** GPT-5.2, GPT-5.4, GPT-5.4 Mini, GPT-5.4 Nano
- **Anthropic:** Claude Opus 4.5, Claude Sonnet 4.5, Claude Haiku 4.5; Claude Sonnet 4.6, Claude Opus4.6
- **Google:** Gemini 3 Pro
- **Amazon:** Nova Lite, Nova Micro, Nova 2 Lite
- **xAI:** Grok 4.1 Fast NR (non-reasoning), Grok 4.1 Fast R (reasoning)

The report is organized to mirror how security decisions are made. It begins with model-level exposure under single-turn and iterative conditions (Figure 1), then decomposes multi-turn outcomes by strategy family (Figure 4) and content category (Figure 7, with the full model-by-category grid in Appendix Figure 9). Finally, it drills into single-turn tactical surfaces using subtechniques, content categories, and procedures to identify where failures concentrate and where mitigation work is likely to have the highest leverage (Figure 5, Figure 6, Figure 8).

Terminology is used consistently throughout this report. **Single-turn** refers to one attacker prompt and onemodel response. **Multi-turn** refers to iterative interactions where an attacker can adapt prompts across turns. **Strategy** denotes a multi-turn behavioral approach (e.g., role-play or refusal reframe). **Procedure** denotes a prompt-level manipulation template (e.g., system prompts, paraphrase, impersonation). **Content category** denotes a higher-level risk domain, while **content type** provides a finer-grained surface. **Subtechnique** refers to Cisco taxonomy subtechnique labels used to group attack mechanisms into a stable vocabulary.

1.1 Current Safety and Security Benchmark Landscape

The rapid deployment of frontier large language models has generated a parallel ecosystem of safety and security benchmarks. However, a growing body of evidence indicates that this ecosystem suffers from structural limitations that can systematically understate risk, conflate safety with capability, and leave critical attack surfaces unmeasured.

Single-turn bias. The dominant safety evaluation paradigm treats each test as a single adversarial prompt paired with a single model response. HarmBench [3], perhaps the most widely adopted standardized redteaming framework, evaluates 510 harmful behaviors across 33 target models using single-turn interactions exclusively. The MLCommons AI Safety benchmark (AILuminate v1.0/v1.1), despite its scale of over 43,000 adversarial prompts across 12 hazard categories, is explicitly constrained to single-turn conversations— the developers note that it can therefore only evaluate "content-only type of hazards" and that multi-turn interactions, multimodal understanding, and emerging hazard categories require continued development [4], [5]. TrustLLM [6] provides a broad evaluation of trustworthiness across six dimensions but similarly operates in a single-turn paradigm. This single-turn bias is not a minor limitation: it means the benchmarks that inform model cards, safety reports, and procurement decisions systematically omit the interaction regime where, as this report demonstrates, model ordering and tail-risk exposure can change most dramatically.

Safetywashing and capability conflation. Ren et al. provide a rigorous meta-analysis showing that many widely used safety benchmarks are highly correlated with general model capabilities and training compute—a phenomenon the authors term "safetywashing" [7]. Their analysis reveals that benchmarks measuring human preference alignment, scalable oversight, truthfulness, and static adversarial robustness track closely with upstream capabilities, making it difficult to distinguish genuine safety progress from capability scaling. In contrast, measurements of dynamic adversarial robustness, bias, and calibration show relatively low correlation with capabilities, suggesting these capture more distinct safety properties. This finding has direct implications for how stakeholders should interpret benchmark improvements: a model that scores higher on a capability-correlated safety benchmark may simply be a more capable model, not necessarily a safer one.

Structural failures across 210 benchmarks. A comprehensive survey by Becker et al. in 2025 examined 210 AI safety benchmarks and documented systemic shortcomings across technical, epistemic, and sociotechnical dimensions [8]. Technical failures include biases in dataset creation, data contamination, and inadequate probabilistic metrics. Epistemic failures include construct validity problems—benchmarks often measure

proxies for safety rather than safety itself—and the persistent challenge of unknown unknowns. Sociotechnical failures include misaligned incentives that reward gaming of benchmark results and cultural dynamics that prioritize state-of-the-art performance claims over genuine safety assessment. The authors propose that sound safety benchmarking requires adherence to established risk management principles, robust probabilistic metrics, and formal measurement theory—standards that few existing benchmarks meet.

The multi-turn evaluation gap. Emerging research confirms that multi-turn interactions represent a qualitatively different threat surface. The Multi-Turn Safety Alignment (MTSA) framework addresses multi-round jailbreak attacks through adversarial iterative optimization, finding that models vulnerable under iterative pressure are not reliably predicted by single-turn performance [9]. The GALA multi-turn red-teaming agent achieves over 90% attack success rates against sampled models within five turns using dual-level learning strategies [10]. The STING red-teaming framework further extends this to agentic contexts, showing that step-by-step illicit plans with adaptive follow-ups expose agent-misuse risks invisible to single-turn benchmarks [11]. Collectively, these studies establish that single-turn evaluation and multiturn evaluation measure overlapping but distinct safety properties—a finding this report corroborates at scale across 15 frontier closed models.

Security-specific evaluation remains underserved. The OWASP Top 10 for LLM Applications (2025) identifies prompt injection, sensitive information disclosure, and system prompt leakage among the most critical security risks, yet the benchmark ecosystem remains disproportionately focused on content safety (toxicity, bias, misinformation) rather than adversarial security [12]. This imbalance persists despite the emergence of structured threat frameworks such as MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems), which provides a comprehensive taxonomy of adversarial tactics and techniques targeting AI systems across the lifecycle, including prompt injection.[13] However, ATLAS is fundamentally a knowledge representation framework, not an evaluation framework: it enumerates what attackers can do, but does not define how to measure model susceptibility or how to compare systems quantitatively under adversarial conditions. The Cisco Integrated AI Security and Safety Framework taxonomy provides an evaluation vocabulary purpose-built for security-relevant decomposition of attack mechanisms, enabling the subtechnique-level analysis (Section 3.4) that generic safety benchmarks do not support by explicitly linking threat enumeration and evaluation operationalization. [14] Critically, the AI Security Framework is structured to support multilevel abstraction aligned with evaluation needs: high-level attacker objectives provide coverage guarantees and risk prioritization, while fine-grained techniques and subtechniques enable the construction of targeted test cases and scenario-based evaluations. This decomposition makes it possible to map adversarial behaviors directly to measurable outcomes (e.g., attack success rates, failure modes, and control effectiveness).

Regulatory and governance expectations are encouraging more robust evaluation capabilities. The NIST AI RMF 1.0 and its Generative AI Profile (NIST AI 600-1) call for adversarial testing as part of the Measure function, and the forthcoming NIST Cyber AI Profile (IR 8596, December 2025) bridges AI risk management with the Cybersecurity Framework 2.0 [15], [16], [17]. The EU AI Act's High Risk provisions (effective December 2027) require robustness and cybersecurity testing under Article 15 [18]. At this time, these frameworks do not specify the interaction regime, strategy decomposition, or slice-support labeling that this report posits is necessary for enterprises to have decision-grade safety assessment. Enterprises should make it a priority to address these safety assessments now to stay ahead of possible standards and regulations in this space.

1.2 Contributions

This report is designed to address these compounding limitations. By pairing single-turn and multi-turn evaluation, decomposing outcomes by strategy family and tactical surface, and explicitly labeling slice support, it demonstrates a reporting methodology that closes several of the gaps identified above and makes three primary contributions:

- Paired regime measurement: it treats single-turn and multi-turn exposure as distinct regimes and reports them side-by-side so that shifts between regimes are visible at a glance (Figure 1).
- Decomposition for action: it decomposes multi-turn outcomes by strategy and category, and decomposes single-turn risk by subtechnique, content category, and procedure to surface concentrated failure pathways (Figure 4, Figure 7, Figure 5, Figure 6); the full model-by-category grid is retained in the appendix for completeness (Appendix Figure 9).
- Support-aware interpretation: it explicitly separates high-support slices (appropriate for tactical conclusions) from low-support cells (appropriate for triage and follow-on testing), and reports confidence half-widths to avoid over-reading small differences (Table 2).

The following analytic framework details the data construction, metric definitions, and interpretation controls used in this report.

2 Analytic Framework and Methodology

2.1 Scope

This analysis uses a fixed snapshot containing:

- 30,090 single-turn prompts across the 15 selected models (2,006 per model),
- 6,986 multi-turn attacks across the selected models (approximately 4.8 attacks per conversation on average),
- 1,456 multi-turn conversations in aggregate for the selected set.

Evaluation dimensions include objectives, techniques, subtechniques, procedures, content categories and content types for single-turn analysis, plus strategies, categories, conversation structure, and turn depth for multi-turn analysis. The modeling choice is intentionally comparative rather than causal: results are used to identify where risk concentrates and how that concentration differs by model, not to claim universal performance under all downstream product configurations.

2.2 Single-turn Setting

Single-turn evaluation treats each prompt as a single adversarial attempt with a binary outcome. A defense success is represented as Pass = 1 and an attack success is represented as Pass = 0. Single-turn ASR is therefore the fraction of prompts that succeed as attacks. Single-turn analysis is decomposed by procedures, content surfaces, and taxonomy labels to understand how attacks succeed rather than only how often they succeed.

2.3 Multi-turn Setting

Multi-turn evaluation treats each multi-turn attack as an interactive attempt in which an attacker can adapt across turns. The analysis reports the fraction of attacks that reach attacker success as multi-turn ASR, and the complement as multi-turn resistance. Multi-turn results are further decomposed by strategy family and category slices to identify which interaction patterns co-occur with higher observed exposure for each model in this design.

2.4 Taxonomy and Labeling

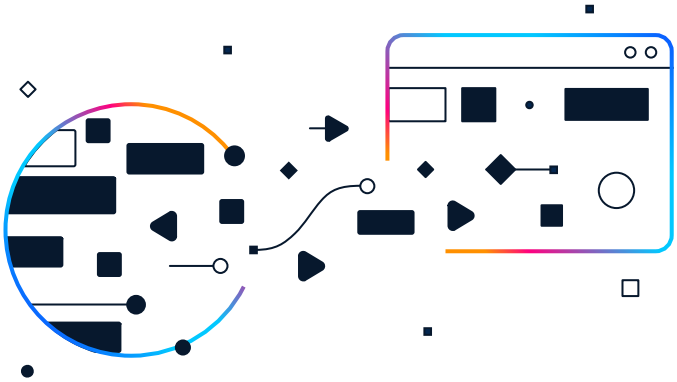
Labels are used as analytic coordinates rather than as claims of ground truth intent. Content categories provide a higher-level organization over risk domains, while content types provide finer-grained surfaces within those domains. Procedures represent prompt-level manipulation templates.

Cisco Integrated AI Security and Safety Framework taxonomy subtechniques provide a stable, hierarchical vocabulary for grouping attack mechanisms. The taxonomy is purpose-built for AI security evaluation and maps attack patterns to recognizable failure families (e.g., direct prompt injection and jailbreak-style semantic manipulation). Its hierarchical structure—objectives, techniques, subtechniques—enables consistent comparison across models and supports mitigation work framed in terms of specific, traceable attack mechanism codes. This specificity distinguishes it from broader threat frameworks that may lack the granularity needed for prompt-level evaluation.

For reference, this report uses the following shorthand glossary for the Cisco AI Security Framework subtechnique codes that appear in figures and text. The underlying taxonomy is not contiguous by numeric suffix (e.g., AISubtech-2.1.2 exists in the full framework but does not appear among the columns selected for Figure 5); only codes surfaced in this cohort's primary subtechnique view are listed below.

- AISubtech-1.1.1: Instruction Manipulation (Direct Prompt Injection). Figure 5 column label: "Instruction (Direct)".

- AISubtech-1.1.2: Obfuscation (Direct Prompt Injection). Figure 5 column label: “Obfuscation (Direct)”.
- AISubtech-2.1.1: Context Manipulation (Jailbreak). Figure 5 column label: “Context (Jailbreak)”.
- AISubtech-2.1.3: Semantic Manipulation (Jailbreak). Figure 5 column label: “Semantic (Jailbreak)”.
- AISubtech-2.1.4: Token Exploitation (Jailbreak). Figure 5 column label: “Token (Jailbreak)”.



2.5 Outcome Labeling and ASR Semantics

This report aggregates binary outcome flags already present in the curated evaluation corpus: single-turn records expose attack success (equivalently Pass as defined above), and multi-turn executions expose success flag on attacker-response records. Those fields encode whether, under the evaluation pipeline’s operational definition, the attacker objectivewas satisfied for that attempt. This analysis does not re-adjudicate outcomes. Inter-rater reliability and the mix of automated versus human review steps in the upstream labeling workflow are defined by the Cisco AI Threat Intelligence & Security Research evaluation methodology that produced the dataset; they are outside the scope of this snapshot reanalysis. Readers should treat ASR as a slice-specific rate under that labeling contract, not as an absolute ground-truth harm rate.

2.6 Data Construction and Aggregation

Single-turn analysis is computed over the fixed prompt set per model. Multi-turn analysis is computed over multi-turn attacks, which are further decomposed into strategy-family slices and category slices where available. In all cases, group-level ASR is reported as a weighted rate (total successes divided by total attempts in that slice) rather than an unweighted average of per-model rates. This applies to subtechnique, procedure, content-type, strategy, and category aggregates throughout the report: the denominator is always the total number of attempts within the slice, not the number of models.

The Cisco AI Threat Intelligence & Security Research team curated the prompt banks to balance breadth of coverage with high-ASR adversarial prompts, then normalized them into a consistent schema for cross-model comparison. Sensitive domains such as violent or self-harmcontent are represented, in part, with deliberately challenging prompts so that safety controls are stressed under comparable difficulty; high weighted ASR in those categories therefore reflects behavior under adversarial selection in this bank, not a claim about population base rates or real-world incident frequency. Potential biases include over-representation of certain categories or procedures that are easier to operationalize as prompts and under-representation of long-horizon, tool-mediated, or context-rich failure modes. Accordingly, category- and procedure-level findings should be interpreted as properties of this curated evaluation design, not as an estimate of real-world prevalence.

To reduce overinterpretation of sparse cells, heatmaps are used to preserve structure, while top-k charts are used to emphasize the highest-impact slices. This combination makes long-tail volatility visible without letting it dominate narrative conclusions.

2.7 Uncertainty and Confidence Half-widths

Model-level confidence half-widths are reported as an approximate 95% normal-approximation interval for a binomial proportion. If $\hat{p}$ is the observed ASR and n is the number of trials, the half-width is:

$$1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot 100$$

These half-widths are used as an interpretation control: they discourage reading small rank differences as meaningful when sample sizes imply overlapping uncertainty.

2.8 Metric Definitions

Metric	Formula
Single-turn ASR (%)	$\frac{\sum(1-\text{Pass})}{N_{\text{st}}} \times 100$
Single-turn Resistance Rate (SRR, %)	$100 - \text{ASR}_{\text{st}}$
Multi-turn ASR (%)	$\frac{\sum(1-\text{Pass})}{N_{\text{mt}}} \times 100$
Multi-turn Resistance Rate (MRR, %)	$100 - \text{ASR}_{\text{mt}}$

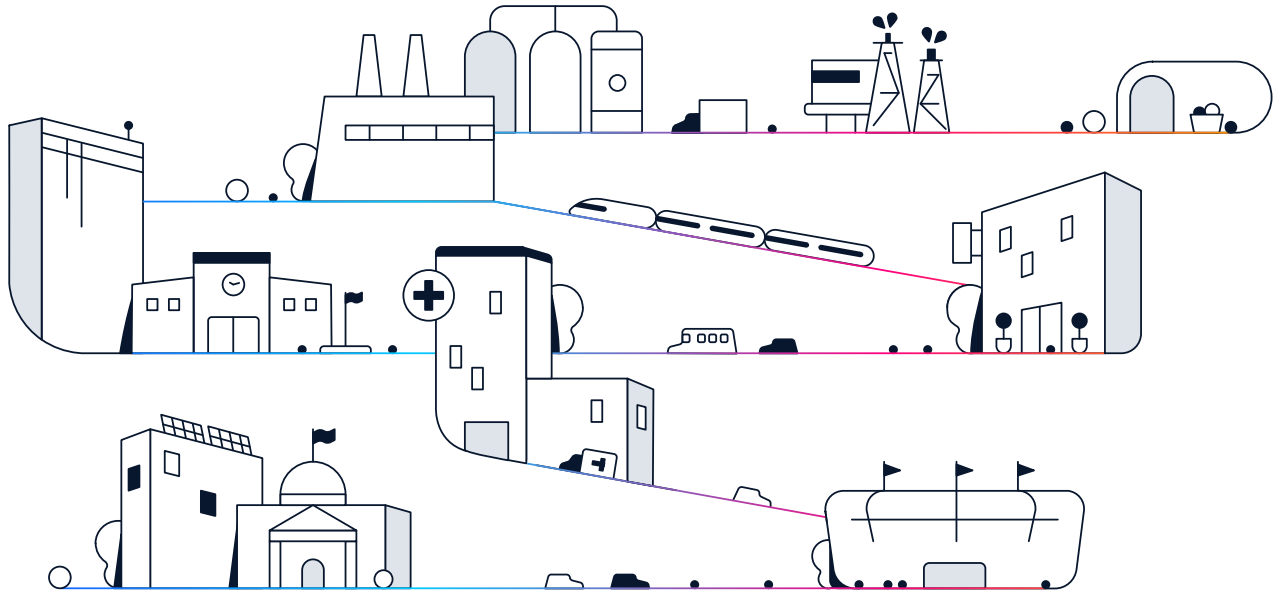
Table 1. Metric definitions used in this report.

In both regimes, Pass = 1 indicates a defense success (attack blocked). Single-turn *N* counts prompts; multi-turn *N* counts attacks in the slice.

2.9 Interpretation Controls

Three controls are applied when interpreting results:

- Confidence width at model level: model-level ASR values include approximate 95% confidence halfwidths to avoid over-reading small rank differences (Table 2).
- Cell-support checks: strategy cells are high-support and relatively stable, while category and long-tail procedure cells are sparse and high-variance; tactical conclusions therefore prioritizeweighted aggregates and higher-support regions (Figure 4, Figure 6, Figure 8).
- Support-aware ranking slices: all-cell heatmaps preserve sparse structure for visibility, while top-10 rankings emphasize weighted high-impact slices to reduce overinterpretation of low-support extremes (Figure 6, Figure 8).



3 Results

Model-Level Analysis

3.1 Result 1: Multi-turn Exposure Splits the Cohort

Model-level outcomes are summarized in Table 2 and Figure 1.

Model	Single-turn ASR (%)	Single-turn CI95 (+/- pp)	Single-turn prompts	Multi-turn ASR (%)	Multi-turn CI95 (+/- pp)	Multi-turn attacks	Conversations
Grok 4.1 Fast (non-reasoning)	34.15	2.08	2006	88.30	3.02	436	91
Gemini 3 Pro	18.10	1.68	2006	73.35	4.00	469	99
Grok 4.1 Fast (reasoning)	23.98	1.87	2006	43.47	4.61	444	91
Amazon Nova Micro	64.91	2.09	2006	30.85	4.18	470	98
Amazon Nova Lite	63.61	2.11	2006	28.87	4.06	478	99
OpenAI GPT-5.4	2.74	0.71	2006	24.68	3.91	466	97
OpenAI GPT-5.2	4.74	0.93	2006	23.50	3.84	468	98
Anthropic Claude Opus 4.6	3.64	0.82	2006	16.20	3.34	469	98
Anthropic Claude Haiku 4.5	3.54	0.81	2006	13.53	3.08	473	98
Anthropic Claude Sonnet 4.6	3.04	0.75	2006	13.36	3.10	464	97
Anthropic Claude Sonnet 4.5	2.59	0.70	2006	12.97	3.01	478	99
OpenAI GPT-5.4 Nano	9.52	1.28	2006	12.61	3.01	468	98
OpenAI GPT-5.4 Mini	9.57	1.29	2006	12.18	2.96	468	97
Anthropic Claude Opus 4.5	2.19	0.64	2006	11.16	2.86	466	98
Amazon Nova 2 Lite	34.05	2.07	2006	7.89	2.44	469	98

Table 2: Model-level ASR outcomes and confidence half-widths (sorted by multi-turn ASR, descending).

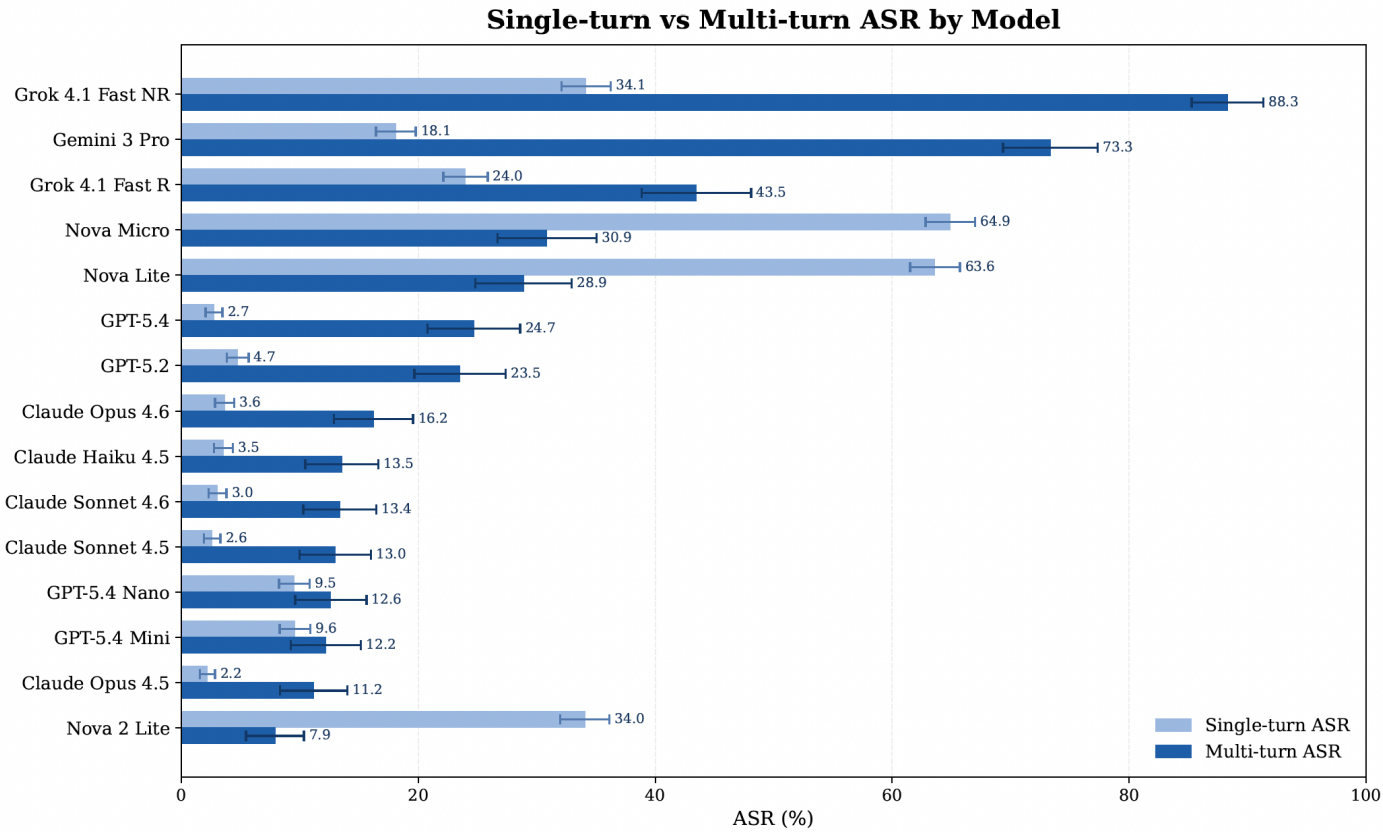


Figure 1: Single-turn versus multi-turn ASR by model, with approximate 95% confidence half-widths on single-turn (upper bar) and multi-turn (lower bar) estimates.

Under multi-turn testing the cohort separates into three bands:

- **High multi-turn exposure:** Grok 4.1 Fast NR (88.30%) and Gemini 3 Pro (73.35%).
- **Intermediate exposure:** Grok 4.1 Fast R (43.47%), Nova Micro (30.85%), Nova Lite (28.87%), GPT-5.4 (24.68%), and GPT-5.2 (23.50%).
- **Lower exposure:** Claude Opus 4.6 (16.20%), Claude Haiku 4.5 (13.53%), Claude Sonnet 4.6 (13.36%), Claude Sonnet 4.5 (12.97%), GPT-5.4 Nano (12.61%), GPT-5.4 Mini (12.18%), Claude Opus 4.5 (11.16%), and Nova 2 Lite (7.89%).

The cleanest within-family contrast in the cohort is Grok 4.1 Fast NR versus Grok 4.1 Fast R: enabling the reasoning configuration is associated with a drop from 88.30% to 43.47% multi-turn ASR (and from 34.15% to 23.98% single-turn ASR) under the same evaluation harness. This is a large, model-identical swing tied to a single exposed capability flag, and it is directly actionable for developers calibrating default modes and for customers deciding which endpoint configuration to deploy.

Single-turn and multi-turn rankings are not equivalent. The single-turn range is 62.72 pp, but multi-turn range widens to 80.41 pp, indicating stronger separation when comparing models on iterative attacks. The Nova variants sharpen this point: they expand the single-turn spread without converging to a single multi-turn profile.

This widening matters for evaluation design. If governance decisions rely on single-turn measurements alone, models with materially different iterative exposure can be treated as equivalent. The paired view in Table 2 and Figure 1 reduces that blind spot by making multi-turn separation explicit at the same level of prominence as single-turn ASR.

Beyond ordering, the shape of exposure is different. Single-turn ASR concentrates in a subset of tactical surfaces (Result 4 and Result 6), while multi-turn ASR shows substantial variation across strategy families (Result 3). This distinction motivates why the report treats multi-turn behavior as a first-class measurement regime rather than a secondary stress test.

3.2 Result 2: Single-turn Strength Does Not Reliably Predict Multi-turn Behavior

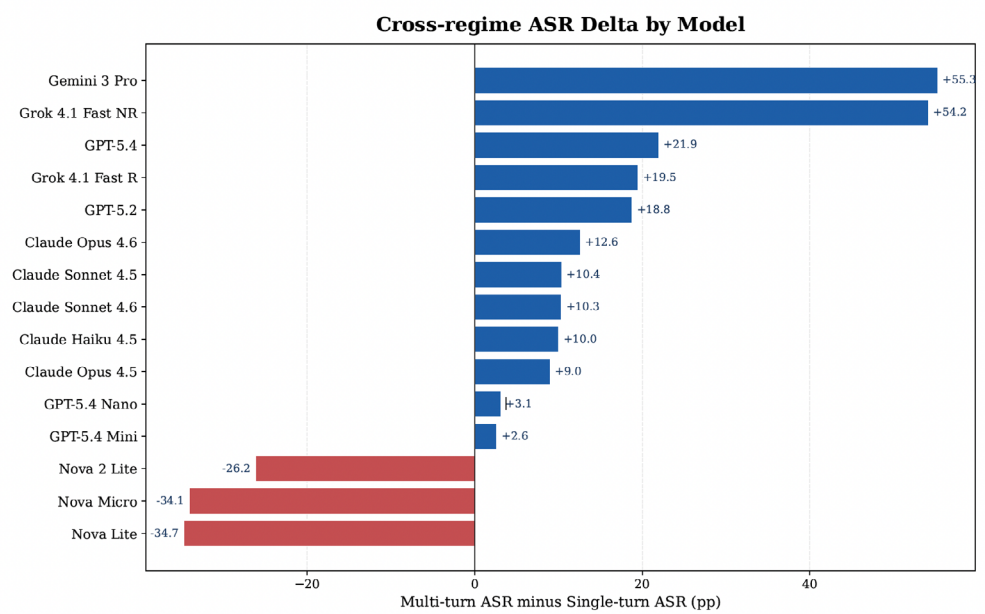


Figure 2: Cross-regime ASR delta (multi-turn minus single-turn) by model. Positive values indicate higher multi-turn than single-turn ASR; negative values indicate the reverse. Bar labels are rounded to one decimal; the text reports the corresponding two-decimal values.

Cross-turn deltas are the most consequential finding in this dataset. The largest positive shifts are Gemini 3 Pro (+55.25 pp) and Grok 4.1 Fast NR (+54.15 pp), while the Nova family moves in the opposite direction, led by Nova Lite at −34.74 pp. All three Nova variants show large negative deltas—Nova Lite (−34.74 pp), Nova Micro (−34.06 pp), Nova 2 Lite (−26.16 pp)—yet they occupy very different positions on both axes. Single-turn performance is not a reliable proxy for multi-turn resilience, even across closely related models. Even very low single-turn ASR models can exhibit much higher multi-turn ASR (e.g., Claude Opus 4.5: 2.19% to 11.16%; GPT-5.4: 2.74% to 24.68%; GPT-5.2: 4.74% to 23.50%).

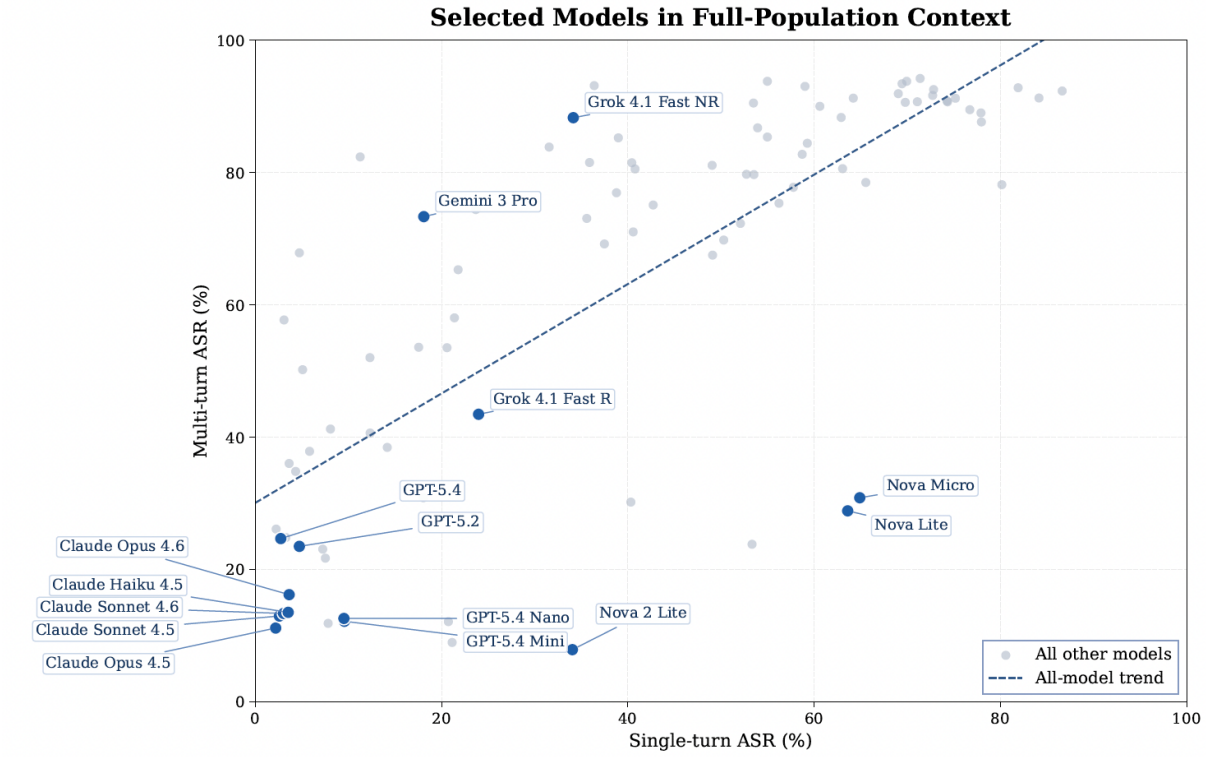


Figure 3: Selected models overlaid on the broader single-turn vs multi-turn distribution.

Figure 3 confirms this non-uniform transition. Nova 2 Lite sits in the lower-right region—high singleturn ASR but the lowest multi-turn ASR in the cohort, the genuinely interesting outlier where single-turn brittleness does not translate into iterative exposure. Grok 4.1 Fast NR and Gemini 3 Pro occupy the upper portion, where single-turn results understate iterative risk. Nova Lite and Nova Micro are exposed on both axes. The GPT-5.4 family is internally split: the flagship model rises meaningfully under multi-turn pressure, while the Mini and Nano variants remain clustered with the lower-exposure models. GPT-5.2 sits closer to GPT-5.4 than to the Mini/Nano variants in multi-turn behavior.

From an assurance standpoint, these residuals imply that single-turn and multi-turn controls should be validated separately. The data supports a two-axis interpretation: baseline refusal performance and iterative resilience are related but not interchangeable capabilities.

This also affects benchmarking. If model progress is measured only on single-turn gains, regression risk can remain hidden in the multi-turn dimension. Conversely, the cohort includes models where multiturn resilience appears disproportionately strong relative to single-turn behavior, suggesting that interaction regime can invert the story told by single-turn summaries.

Strategy and Category Decomposition

3.3 Result 3: Strategy Family and Cross-Model Dispersion in Multi-turn ASR

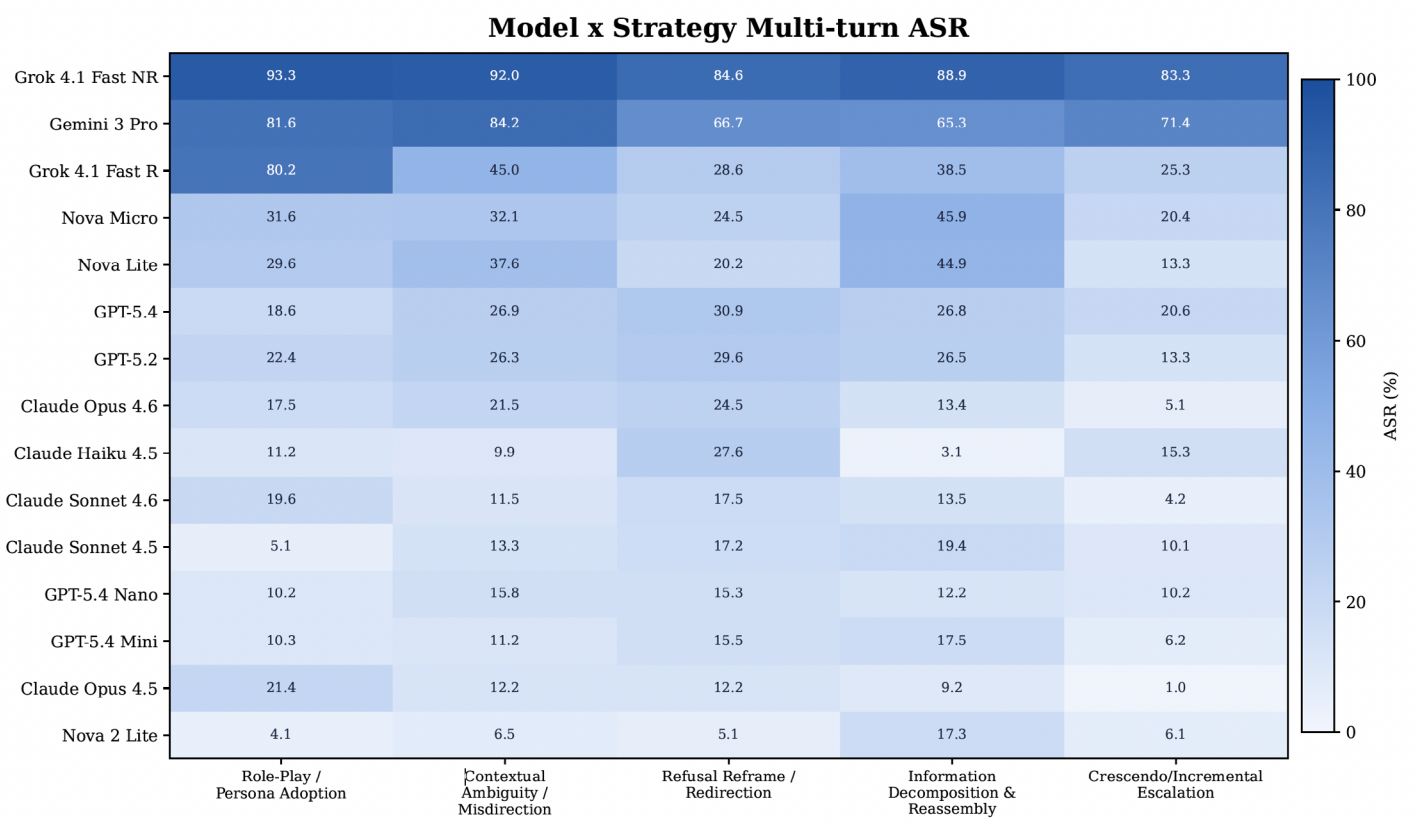


Figure 4: Model by strategy multi-turn ASR for the five strategy families analyzed in Table 3. A sixth family (artifact) appears in the raw evaluation design but is omitted here because per-model support is too low for stable cohort-level comparison; Table 3 and operational ritual 1 refer to these five families only.

Strategy	Weighted ASR (%)	ASR spread (pp)	Total attacks
Role-Play / Persona Adoption	29.89	89.25	1452
Contextual Ambiguity / Misdirection	29.51	85.51	1176
Information Decomposition & Reassembly	29.17	85.83	1450
Refusal Reframe / Redirection	27.75	79.51	1456
Crescendo/Incremental Escalation	20.04	82.31	1452

Table 3: Cross-model weighted ASR and ASR spread by multi-turn strategy family.

Two distinct statements are both true in this slice. First, at cohort level, aggregate strategy-weighted ASR spans a band of roughly ten percentage points: from 20.04% (Crescendo) to 29.89% (Role-Play / Persona Adoption) in Table 3—a meaningful but modest spread compared with headline model-level ASR. Second, within each strategy family, the gap between the most exposed and least exposed model is very large: perstrategy cross-model spreads range from 79.51 pp to 89.25 pp. In other words, strategy labels primarily stratify which models separate from one another, not only the cohort-average difficulty of a strategy. Two model-level patterns stand out:

- Grok 4.1 Fast NR is consistently high across all five strategies, indicating broad multi-turn susceptibility in this evaluation rather than a single-strategy niche.
- Comparatively low multi-turn ASR models such as Claude Opus 4.5, Claude Sonnet 4.6, and GPT-5.4 Mini/Nano remain low in aggregate, but still show material variation across strategy families, which means even these models warrant strategy-stratified monitoring.

The practical implication is that mitigation plans should be strategy-indexed. A uniform policy or postprocessing layer will often underperformcompared with strategy-aware hardening that prioritizes the highestlift strategy families first.

Conceptually, strategy slices are a reminder that multi-turn ASR is not a single phenomenon. Different strategies change the conversational framing, the model's perceived role, and the degree to which the attacker can incrementally adapt. Figure 4 should therefore be read as a map of interaction vulnerabilities: it identifies which social or rhetorical contexts are associated with higher observed ASR for each model in this dataset.

Tactical Surfaces

3.4 Result 4: Single-turn Vulnerability Concentrates in Specific Tactical Surfaces

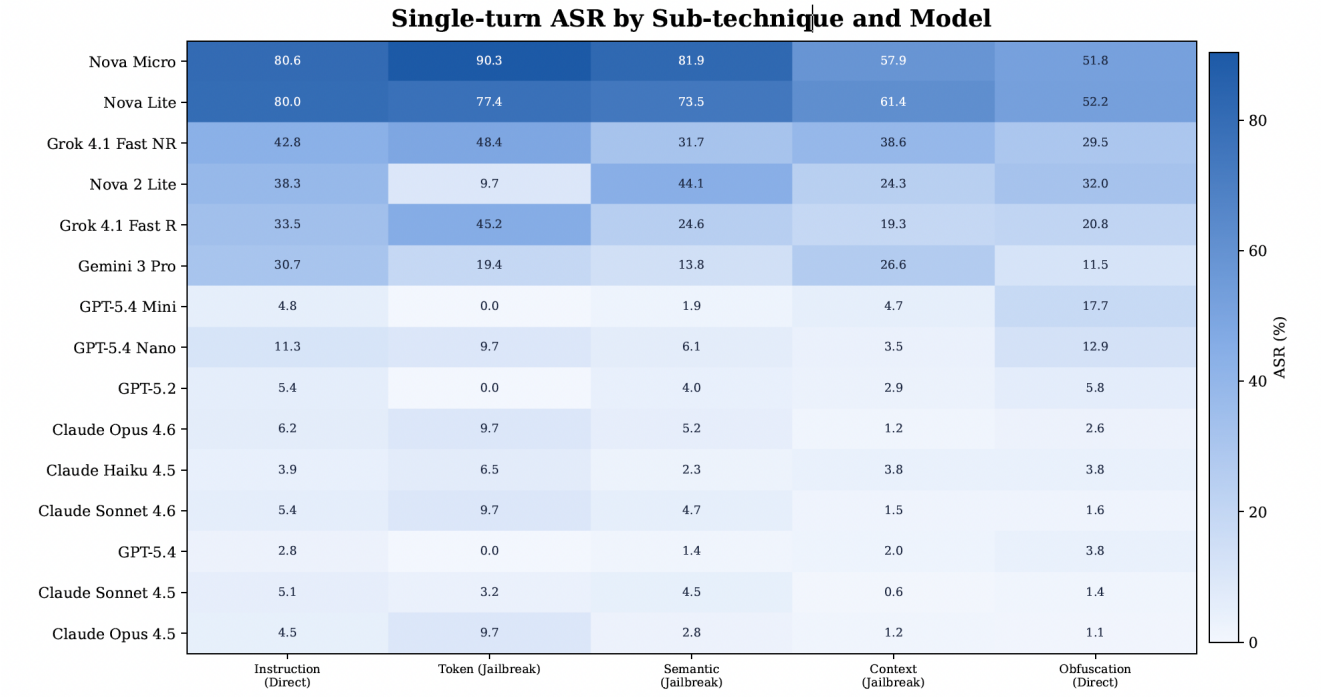


Figure 5: Single-turn ASR by subtechnique and model.

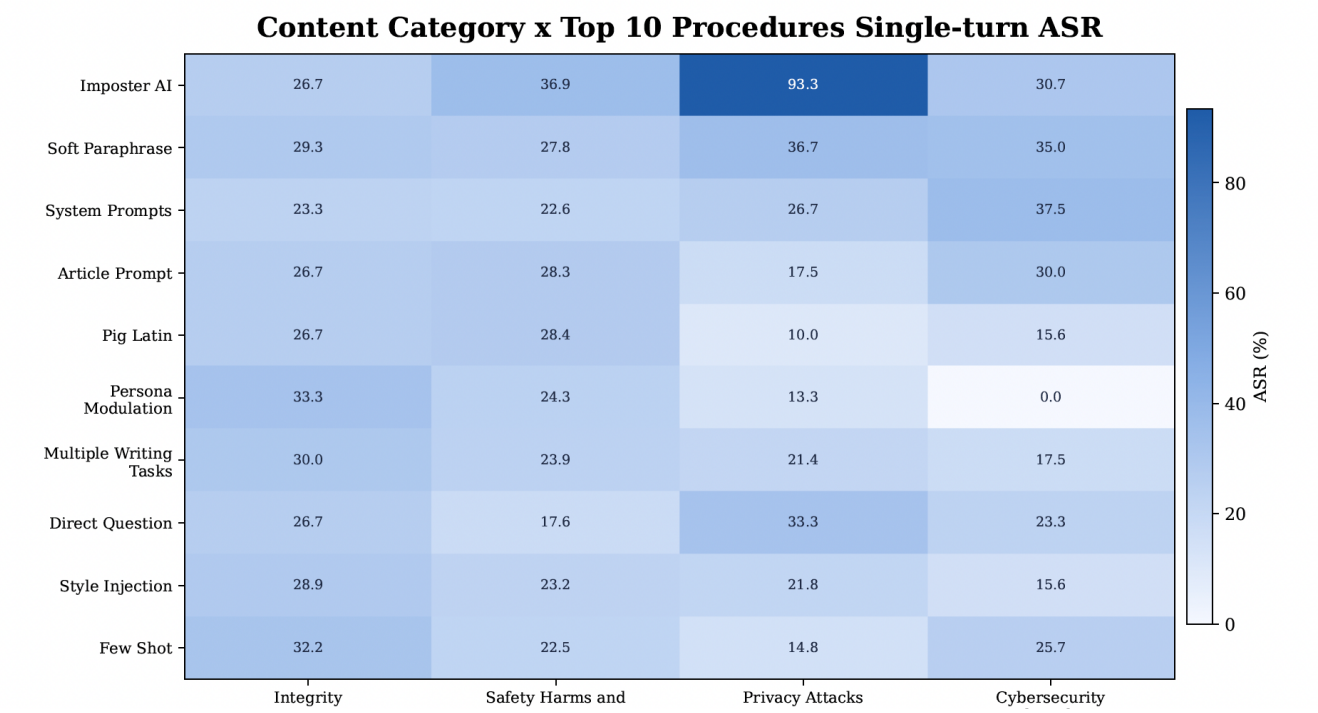


Figure 6: Content-category by top 10 procedures single-turn ASR.

Figure 6 shows how top-ASR procedures distribute across the four content categories, and makes crosscategory concentration visible without introducing lower-priority long-tail rows. Cisco taxonomy concentration remains visible in Figure 5: the highest weighted subtechniques are AISubtech-1.1.1: Instruction Manipulation (Direct Prompt Injection) at 23.68% and AISubtech-2.1.4: Token Exploitation (Jailbreak) at 22.58%.

These views are complementary. subtechnique heatmaps identify what attack mechanisms are consistently effective, while the category-procedure matrix shows where those mechanisms express the highest realized ASR in the selected cohort. Together they provide a direct basis for prioritizing evaluation expansion and control tuning.

Notably, the procedure-category matrix reveals whether a procedure's risk is category-general (high across multiple categories) or category-specific (high in one domain but muted elsewhere). This distinction matters for mitigation: category-general procedures motivate broad policy and instruction-following hardening, while category-specific hot spots motivate domain-aligned safety tuning and targeted adversarial training.

Strategy and Category Decomposition

3.5 Result 5: Category Signals with Sparse Model×Category Support

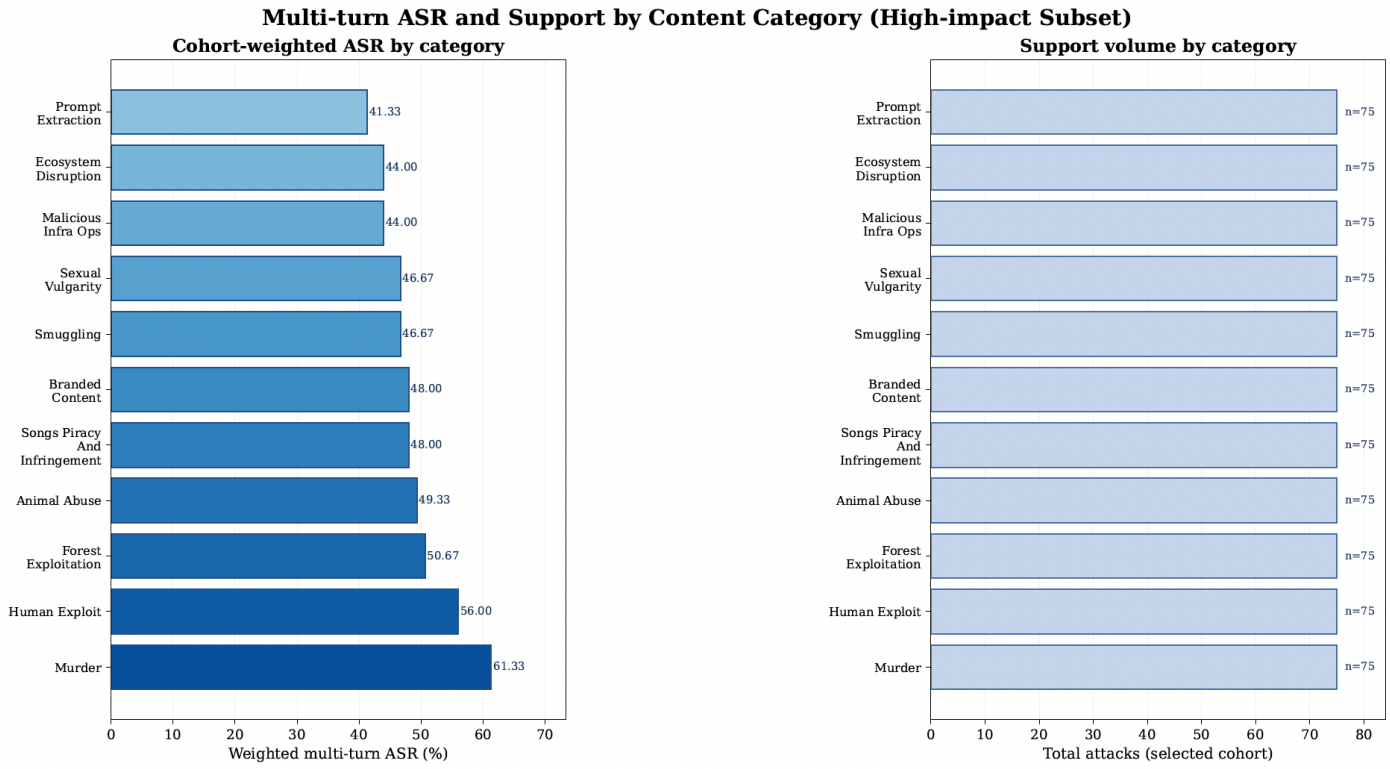


Figure 7: Cohort-weighted multi-turn ASR by high-impact content category (left) and total attack support summed across the 15 models (right). The displayed subset contains the highest-ASR categories that also meet the cohortsupport filter (at least 50 total attacks), so lower-ASR categories such as bomb threats (3.70%, *n* = 54) are intentionally excluded from the plotted set. This marginal view avoids interpreting individual cells in the model×category grid, where most model×category slices use five attacks per cell in the current design (Appendix Figure 9). Read weighted rates as cohort-level triage signals, not as stable per-model estimates.

Across all categories with at least 50 total attacks in the selected cohort, weighted ASR spans from 3.70% (bomb threats; *n* = 54) to 61.33% (murder; *n* = 75). Figure 7 then narrows to the highest-impact supported subset, so the plotted floor is 41.33% (prompt extraction; *n* = 75) rather than bomb threats. Every displayed category aggregates 75 attacks across the 15-model cohort, or five attacks per model on average, which makes these marginals useful for cohort-level triage but still too thin for stable model-by-category conclusions. Within the plotted subset, human-exploit (56.00%), forest-exploitation (50.67%), and brandedcontent (48.00%) remain elevated. These categories are candidates for targeted follow-on testing to separate stable signals from sampling artifacts; interpret their headline rates alongside the curation caveat in Section 2.6 (adversarially selected prompts, not population prevalence).

Tactical Surfaces

3.6 Result 6: Highest-risk Single-turn Slices Are Concentrated

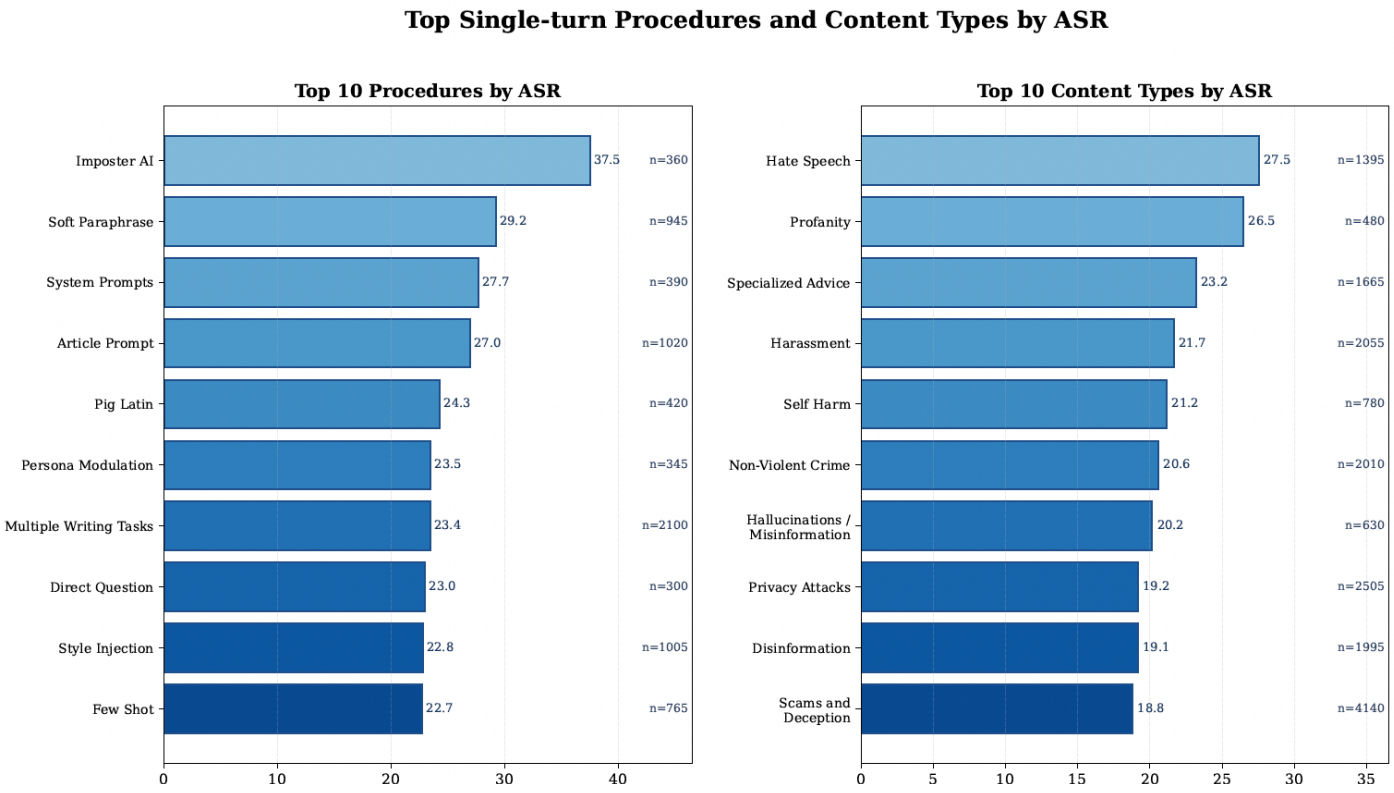


Figure 8: Top 10 procedures and top 10 content types by weighted single-turn ASR.

Figure 8 provides a direct prioritization view of the highest-ASR single-turn slices. Procedure risk is headed by Imposter AI (37.50%), followed by Soft Paraphrase (29.21%) and System Prompts (27.69%). Contenttype risk is led by Hate Speech (27.53%), then Profanity (26.46%) and Specialized Advice (23.18%). The long-tail floor includes near-zero ASR procedures with very small support, so the top-ranked slices remain the most actionable hardening targets.

In practice, this ranking supports phased remediation: target the highest-ASR procedures and content types first, then re-measure to verify absolute ASR reduction before broadening to lower-priority slices.

3.7 Synthesis

Across the six results, a coherent pattern emerges: aggregate model-level numbers are necessary but insufficient. Multi-turn exposure reorders models relative to single-turn ASR (Result 1 and Result 2), and strategy family is associated with large cross-model dispersion in multi-turn outcomes even when cohortaverage strategy rates sit in a narrower band (Result 3). Single-turn weaknesses are concentrated into a small set of mechanisms and surfaces rather than being evenly distributed (Result 4 and Result 6), and category slices are informative at cohort weight for triage, but per-cell support in this snapshot is too small for stable model-level conclusions (Result 5).

Taken together, Figures 1–8 justify a reporting posture that is both comparative and tactical: compare models at the headline level, but steer remediation and follow-on testing using strategy-stratified multi-turn slices and taxonomy-aligned single-turn hot spots.

4 Discussion

4.1 Operational Recommendations

The results suggest three specific evaluation rituals that organizations seeking to deploy AI models could adopt to convey a more holistic picture of security and safety risk:

1. **Publish ASR by strategy family on every model release.** Every model release could report ASR for all five strategy families in Table 3, not just an aggregate multi-turn number (the same five families shown in Figure 4). Strategy-stratified reporting matters because cross-model dispersion within each strategy is wide (See Figure 4).

2. **Gate deployments on high-ASR tactical surfaces.** Regression tests should include the top-3 high-ASR procedures (e.g., Imposter AI, Soft Paraphrase, System Prompts) and the top-3 content types (e.g., Hate Speech, Profanity, Specialized Advice) as gating signals. If any of these regress by more than 3 pp between releases—a threshold chosen to exceed the largest single-turn CI95 half-width in the cohort (±2.11 pp, Nova Lite) with margin—the deployment could be held for review.

3. **Flag large cross-regime gaps.** Any model with an absolute gap >15 pp between single-turn and multiturn ASR should trigger a manual review before deployment. In this cohort that rule flags eight models: five with positive deltas (Gemini 3 Pro; Grok 4.1 Fast NR; GPT-5.4; Grok 4.1 Fast R; GPT-5.2) and three with negative deltas (Nova Lite; Nova Micro; Nova 2 Lite) (See Figure 2).

These rituals convert a static benchmark into a continuously improving security measurement program: strategy slices guide interactive defenses, top-surface slices guide prompt-injection and jailbreak hardening, and the cross-regime gap threshold catches models whose single-turn results obscure iterative exposure.

4.2 Strategic Implications

Beyond these rituals, several broader implications follow from the results.

First, multi-turn testing reorders models relative to single-turn ASR. Several models that look strong in single-turn conditions show higher ASR under multi-turn testing in this design. This is not a marginal effect; GPT-5.4 moves from 2.74% single-turn ASR to 24.68% multi-turn ASR, while GPT-5.2 moves from 4.74% to 23.50%, and Gemini 3 Pro plus Grok 4.1 Fast NR show much larger positive shifts. The Amazon Nova family strengthens the asymmetry argument from multiple directions: Nova Lite (−34.74 pp) and Nova Micro (−34.06 pp) show large negative deltas, while Nova 2 Lite (−26.16 pp) remains the clearest inversion case. Together, those three variants make it difficult to treat single-turn performance as indicative of multi-turn resilience. Organizations that select models primarily on the basis of published single-turn safety scores risk misranking candidates by a wide margin. In this cohort, a model with 2.74% single-turn ASR (GPT-5.4) reaches 24.68% under multi-turn pressure—a ninefold increase thatwould not appear on any single-turn benchmark. Gemini 3 Pro shifts from 18.10% to 73.35%, a fourfold increase. Procurement frameworks may wish to require vendors to disclose paired-regime ASR or, failing that, should commission independent multi-turn evaluations before deployment approval.

Second, strategy-specific measurement is mandatory. The 79.51–89.25 pp within-strategy cross-model spreads indicate that aggregate multi-turn ASR can hide actionable per-strategy variation. In practice, modellevel reporting should include strategy slices for release decisions and guardrail tuning, especially because Role-Play / Persona Adoption remains the highest weighted strategy family in aggregate (29.89%; Table 3). Without strategy decomposition, a model that appears resilient in aggregate

may be highly exposed under a single interaction pattern—precisely the scenario that social-engineering and red-team operators exploit.

Third, failures are distributed unevenly across tactical surfaces. Single-turn subtechnique, contentcategory × procedure, and top-10 slice views indicate concentrated pathways where interventions are likely to deliver disproportionate risk reduction (see the gating procedures in ritual 2 above and the full ranking in Figure 8). Imposter AI alone accounts for a 37.50% weighted ASR—more than 14 percentage points above the tenth-ranked procedure—which means a targeted fix to a single procedure family could meaningfully shift the aggregate single-turn number.

Fourth, even leading frontier labs produce models with exploitable iterative weaknesses. These results are consistent with the view that safety remains a continuous, regime-dependent property rather than a binary certification, even for frontier models from leading providers. The lowest multi-turn ASR in the cohort (Nova 2 Lite, 7.89%) still represents a meaningful residual risk surface, and all Anthropic Claude models, despite their strong single-turn performance (2.19%–3.64% ASR), reach 11.16%–16.20% under iterative attack. This finding is consistent with recent multi-turn red-teaming research: Singhania et al. demonstrate 71% increased vulnerability after five-turn conversations compared to single-turn evaluation [19]. Organizations should treat safety as a continuous, regime-dependent property rather than a binary certification.

Fifth, configuration-driven safety variation calls for greater transparency from model providers. The Grok 4.1 Fast reasoning-mode result—a 44.83 pp drop in multi-turn ASR from a single configuration flag—demonstrates that deployment-time choices invisible to downstream users can dominate safety outcomes. This finding argues for a disclosure norm: model providers should document the safety-relevant effects of configuration flags (reasoning modes, system prompt adherence settings, temperature, and guardrail tiers) alongside capability benchmarks. Anthropic's Responsible Scaling Policy (v3.1, April 2026) and its introduction of public Risk Reports represent a step in this direction, but these disclosures currently focus on catastrophic capability thresholds rather than the kind of granular, regime-stratified safety data this report argues is necessary [20].

Taken together, these findings argue for a layered reporting standard. Model-level ASR is necessary for coarse ranking, but tactical and strategy decomposition is required for decision-grade interpretation. Without decomposition, high-impact pockets of failure can remain hidden beneath acceptable aggregate values.

A corresponding principle is sequencing: The evidence favors a risk-reduction workflow that starts with highest-ASR strategy and procedure slices, validates improvements with targeted reruns, then revisits aggregate ranking. This sequence is more efficient than uniformly expanding every slice at once.

Finally, taxonomy alignment reinforces this workflow. The Cisco AI Security Framework subtechnique structure provides a stable vocabulary for connecting observed failures to mitigation workstreams (e.g., prompt-injection and jailbreak families). When paired with procedure and category slices, it supports a defensible mapping from measurement to interventions, helping teams avoid improving the headline metric while missing high-impact failure routes.

4.3 Future Work

Several extensions would strengthen the evidentiary value of this snapshot. First, expanding multi-turn coverage (both in attack volume and in strategy diversity) would reduce cell variance and improve confidence in category- and strategy-conditioned conclusions. Second, adding targeted replications of the highest-ASR procedures would separate persistent weaknesses from sampling artifacts. Third, longitudinal refreshes would measure whether improvements generalize across time, attack refreshes, and model updates.

5 Limitations

- Some procedure-category and model-category cells have very small support. Heatmaps preserve these cells for visibility, but they should be read as directional indicators rather than stable estimates; additional low-support interpretation notes are summarized in Appendix A.
- Confidence half-widths use a normal approximation for a binomial proportion; this is intended for interpretability and may understate uncertainty, especially for low-support cells where the normal approximation is poor.
- This is a snapshot study; rank ordering can shift with refreshed runs or expanded attack volume.
- Results characterize model behavior under this evaluation regime and do not fully represent downstream application stacks with custom orchestration and controls.

6 Conclusion

For this closed/proprietary frontier-model cohort, the central result is structural: paired single-turn and multi-turn evaluation yields a different model ordering and a different failure map than either regime alone. Strategy-family slices reveal large cross-model dispersion within the same aggregate multi-turn band, while subtechnique, procedure, and content-type views show that single-turn weakness concentrates in a limited set of tactical surfaces. Category slices add triage value at cohort weight, but the report also shows why sparse cells must be labeled as such rather than read as stable model-level estimates.

For the evaluation community, the actionable position is methodological: benchmarks that publish only single-turn ASR omit the dimension where ordering and tail risk change the most for several models in this slice. We argue that reporting standards would benefit from including a paired-regime ASR, strategy-family breakdowns, and explicit labeling of slice support so that readers can separate cohort-weighted findings from sparse cells. The rituals in Section 4.1 instantiate that standard for one security program; public leaderboards and model cards would make cross-organization comparisons far more interpretable if they adopted the same structure.

Without that shift, governance will continue to optimize a single number that this report shows can mis-rank models, hide tail risk, and miss configuration-sensitive effects that are visible as soon as iterative attacks and strategy labels are brought into view.

References

[1] Y. Bengio et al., International ai safety report 2026, 2026. arXiv: 2602.21012 [cs.CY]. [Online]. Available: https://arxiv.org/abs/2602.21012.

[2] E. Hubinger et al., Sleeper agents: Training deceptive llms that persist through safety training, 2024. arXiv: 2401.05566 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2401.05566.

[3] M. Mazeika et al., "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal," in Proceedings of the 41st International Conference on Machine Learning, ser. ICML'24, Vienna, Austria: JMLR.org, 2024.

[4] B. Vidgen et al., Introducing v0.5 of the ai safety benchmark from mlcommons, 2024. arXiv: 2404. 12241 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2404.12241.

[5] S. Ghosh et al., Ailuminate: Introducing v1.0 of the ai risk and reliability benchmark from mlcommons, 2025. arXiv: 2503.05731 [cs.CY]. [Online]. Available: https://arxiv.org/abs/2503.05731.

[6] Y. Huang et al., "Position: TrustLLM: Trustworthiness in large language models," in Proceedings of the 41st International Conference on Machine Learning, R. Salakhutdinov et al., Eds., ser. Proceedings of Machine Learning Research, vol. 235, PMLR, 21–27 Jul 2024, pp. 20 166–20 270. [Online].Available: https://proceedings.mlr.press/v235/huang24x.html.

[7] R. Ren et al., Safetywashing: Do ai safety benchmarks actually measure safety progress? 2024. arXiv: 2407.21792 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2407.21792.

[8] C. Yu, S. Engelmann, R. Cao, D. Ali, and O. Papakyriakopoulos, How should ai safety benchmarks benchmark safety? 2026. arXiv: 2601.23112 [cs.CY]. [Online]. Available: https://arxiv.org/abs/2601.23112.

[9] W. Guo et al., Mtsa: Multi-turn safety alignment for llms through multi-round red-teaming, 2025. arXiv: 2505.17147 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2505.17147.

[10] S. Chen, X. Yu, N. Mehrabi, R. Gupta, Z. Yu, and R. Jia, "Strategize globally, adapt locally: A multi-turn red teaming agent with dual-level learning," 2025. arXiv: 2504.01278 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2504.01278.

[11] N. Talokar, A. K. Tarun, M. Mandal, M. Andriushchenko, and A. Bosselut, "Helpful to a fault: Measuring illicit assistance in multi-turn, multilingual llm agents," 2026. arXiv: 2602.16346 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2602.16346.

[12] OWASP Foundation, Owasp top 10 for large language model applications 2025, https://owasp.org/www- project- top- 10- for- large- language- model- applications/, Version 2025 released November 2024, 2025.

[13] MITRE Corporation, Mitre atlas: Adversarial threat landscape for artificial-intelligence systems, https://atlas.mitre.org/, Accessed 2026, 2026.

[14] A. Chang, T. Saade, S. Mendapara, A. Swanda, and A. Garg, Cisco integrated ai security and safety framework, https://arxiv.org/abs/2512.12921, 2025. arXiv: 2512.12921 [cs.CR].

[15] National Institute of Standards and Technology, Artificial intelligence risk management framework (AI RMF 1.0), 2023.

[16] National Institute of Standards and Technology, Generative AI profile, 2024.

References

[17] National Institute of Standards and Technology, Cyber AI profile, 2025.

[18] European Parliament and Council, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI act), 2024.

[19] A. Singhania, C. Dupuy, S. S. Mangale, and A. Namboori, "Multi-lingual multi-turn automated red teaming for LLMs," in Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025), T. Cao et al., Eds., Albuquerque, New Mexico: Association for Computational Linguistics, May 2025, pp. 141–154, isbn: 979-8-89176-233-6. doi: 10.18653/v1/2025.trustnlp-main.11. [Online]. Available: https://aclanthology.org/2025.trustnlp-main.11/.

[20] Anthropic, Responsible scaling policy v3.1, 2026. [Online]. Available: https://anthropic.com/responsible-scaling-policy/rsp-v3-0.

A Appendix: Low-support Cells and Interpretation Notes

This appendix consolidates low-support interpretation guidance referenced in the main Results and Limitations sections. The intent is to keep the primary narrative focused on higher-support aggregates while still preserving visibility into long-tail structure.

Why Low-support Cells Matter

Many category- and procedure-conditioned slices have small denominators (few attacks in the selected cohort). In such cases, extreme values (e.g., 0% or 100% ASR) can occur by chance and should be interpreted as prioritization signals rather than as stable estimates.

Connection Back to Main Results

In the main narrative, category-level outcomes (Result 5) are treated as triage indicators at cohort weight, and the report emphasizes follow-on testing where category slices show high ASR but limited support. This approach preserves discovery value while preventing overconfident claims driven by small denominators.

Recommended Reading Rules

- Treat any low-support cell as a hypothesis generator: replicate with additional volume before concluding a rank-ordering change.
- Prefer weighted aggregates and top-k slices for decision-grade conclusions; use all-cell heatmaps to preserve structure and identify follow-on test targets.
- Read category slices together with strategy slices: a category can appear moderate in aggregate but still be high-risk under a plausible strategy family.

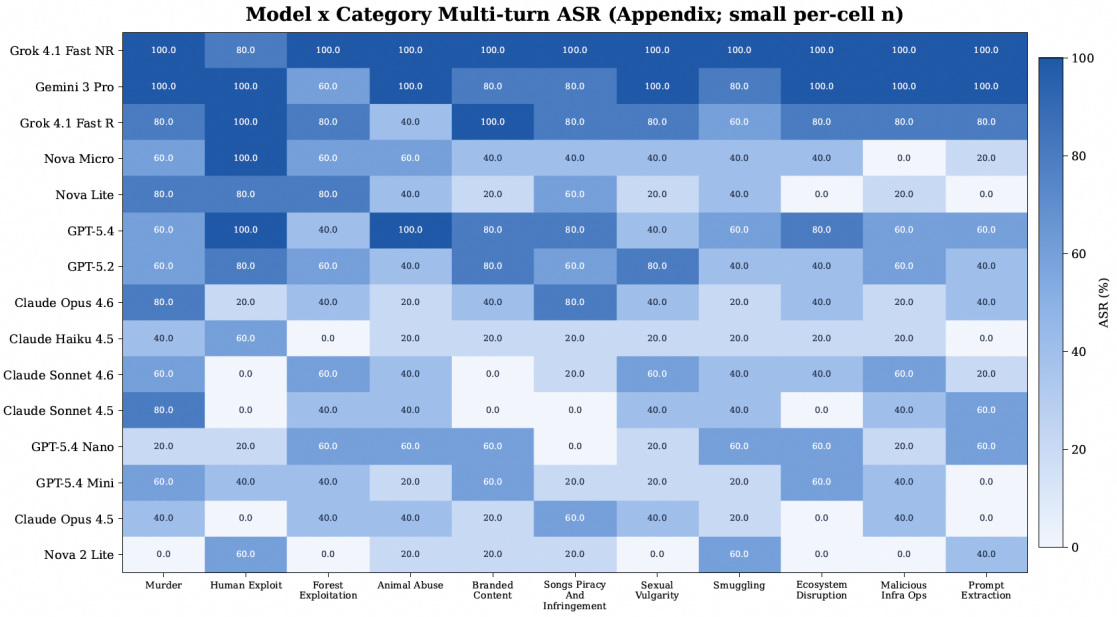


Figure 9: Model × category multi-turn ASR grid (high-impact category subset). Most cells use five attacks per model×category in this evaluation design, so displayed percentages are coarse (multiples of 20%) and should not be read as precise per-model category risk. Prefer Figure 7 for cohort marginals.