NIST AI RMF Playbook Contribution

Suggested Actions for Adversarial Testing of AI Systems with Tool Access

Submitted by: Brandon Arbour, Principal AI Security Researcher, CLS Security Labs LLC

Date: April 2026 | Contact: brandon@clsecuritylabs.com

Public Resource: https://clsecuritylabs.com/blog.html

---

1. Summary

Closed Loop Security Labs (CLS Labs) respectfully submits the following suggested actions for consideration in the NIST AI RMF Playbook. These suggestions address three specific gaps identified through empirical adversarial testing of 273+ large language models across 15+ inference providers, comprising 1.5M+ adversarial probes with multi-vendor cross-judge validation.

Our publicly available research (https://clsecuritylabs.com/blog.html) documents findings that standard safety benchmarks underreport operational risk by a significant margin when AI systems are deployed with tool access, agentic pipelines, or retrieval-augmented generation (RAG) configurations. These findings are relevant to organizations applying the MEASURE and MANAGE functions of the AI RMF to production AI deployments.

CLS Security Labs is a privately held, for-profit AI security assessment firm based in Colorado. This submission constitutes a free, publicly available resource from a for-profit entity, consistent with Playbook inclusion criteria.

2. Suggested Action: MEASURE 2.6

AI RMF Subcategory: MEASURE 2.6 — AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment conditions. The measurements are documented.

Gap Identified

Current Playbook guidance for MEASURE 2.6 does not distinguish between text-only evaluation and evaluation of AI systems with tool access. Our empirical testing across 101 models demonstrates that this distinction is material: well-defended production models exhibit 2x–9x breach rate increases when tool access is enabled. Specific examples from our warehouse: Claude Opus 4 exhibits a 16.2% text-only breach rate versus 38.9% with tool access (2.4x); Llama 3.3-70B exhibits 7.5% text-only versus 42.6% with tools (5.7x); Llama 4 Maverick exhibits 10.4% text-only versus 46.6% with tools (4.5x). The pattern is most pronounced in the

models organizations actually deploy in production: 47 of 101 models tested showed higher breach rates with tool access than without.

Models that score well on standard safety benchmarks (JailbreakBench, AdvBench, HarmBench) demonstrated dramatically higher breach rates when adversarial probes were delivered through tool-calling interfaces, function parameters, and multi-step agentic workflows. Text-only benchmarks do not capture this risk.

Suggested Action

- For AI systems deployed with tool access, function-calling capabilities, or agentic workflows, adversarial testing should include probes delivered through the tool interface, not solely through the text/chat interface. Testing should evaluate whether the AI system can be induced to execute dangerous tool calls, disclose its tool configuration to an attacker, or propagate compromised outputs to downstream agents.
- Organizations should document the difference between text-only breach rates and tool-connected breach rates for the same model, as this delta represents the risk introduced by the deployment configuration rather than the model itself.

Supporting Evidence

CLS Labs conducted assessments using 134 attack modules across 273+ models, including dedicated tool-injection modules that deliver adversarial probes through function-calling interfaces, MCP protocol connections, and multi-step agentic workflows. Across 101 models with complete forge data, 47 exhibited higher breach rates when tool access was enabled. GPT-4.1-nano produced 22 actual tool calls with dangerous parameters in a single assessment. Models that defended well against text-based attacks (e.g., Claude Opus 4: 16.2% text-only breach rate rising to 38.9% with tool access) still disclosed their full tool configuration to attackers at rates up to 71%. Full methodology and results are published at https://clsecuritylabs.com/blog.html and https://clsecuritylabs.com/research.html.

3. Suggested Action: MEASURE 2.7

AI RMF Subcategory: MEASURE 2.7 — AI system security and resilience, as identified in the MAP function, are evaluated and documented.

Gap Identified

Current guidance does not address the variability inherent in LLM-based evaluation of adversarial testing results. Our research demonstrates that breach rate assessments vary from 41.9% to 67.1% depending on which vendor's model serves as the judge, with inter-judge agreement averaging 59.3% across attack categories. Single-vendor scoring introduces systematic bias that organizations may not recognize.

Suggested Action

- When using AI-based systems to evaluate adversarial testing results (e.g., LLM judges scoring whether a model response constitutes a breach), organizations should employ multiple independent scoring systems from different vendors to reduce evaluation bias. Results should be reported as a validated range rather than a single point estimate.

- Organizations should document inter-judge agreement rates across evaluation categories and flag categories with low consensus (below 50% agreement) for additional human review.

Supporting Evidence

CLS Labs employs a three-vendor cross-judge pipeline (Google Gemini, Anthropic Claude, Meta Llama) for all adversarial testing results. Categories with high inter-judge agreement (e.g., Resource Exhaustion at 86%, Secret Exposure at 86%) are treated as high-confidence findings. Categories with low agreement (e.g., Code Generation at 29%, Output Manipulation at 29%) are flagged for manual review. This methodology is documented at https://clsecuritylabs.com/research.html.

4. Suggested Action: MANAGE 2.1

AI RMF Subcategory: MANAGE 2.1 — Resources required to manage AI risks are taken into account, along with viable non-AI alternative systems, deployment approaches, or other AI system reconfigurations.

Gap Identified

Organizations deploying AI systems through third-party inference providers (e.g., serverless API endpoints) may experience materially different security postures depending on the deployment configuration, even when using the identical model weights. Our testing documented cases where the same model exhibited a 0% breach rate on a provider's serverless endpoint versus a 27.8% breach rate on a dedicated deployment, due to undocumented provider-level safety filtering applied only to serverless infrastructure.

Additionally, AI-based severity scoring (used to prioritize remediation) overestimates evasion sophistication in 80–90% of cases when compared against empirical proxy ground truth data. Organizations relying solely on model-based severity assessments may misprioritize their remediation efforts.

Suggested Action

- Organizations should conduct adversarial testing on the specific deployment infrastructure they intend to use in production, not on alternative configurations (e.g., serverless endpoints when production will run on dedicated or self-hosted infrastructure). Security assessments performed on a different deployment configuration should be clearly documented as non-representative of production risk.
- Organizations should validate defense effectiveness empirically by measuring actual block rates against known adversarial inputs, rather than relying solely on model-based severity assessments. Where feasible, pre-deployment and post-deployment breach rates should be compared to quantify remediation effectiveness.

Supporting Evidence

CLS Labs documented provider-level filtering discrepancies across multiple inference providers during cross-provider assessments of identical model weights. Additionally, our defense proxy achieved a reduction from initial breach rates to 100% block rate across 448,000+ classified

records, measured empirically through re-testing, not through model-based estimation. This methodology is documented at https://clsecuritylabs.com/blog.html.

CLS Labs documented a complete defense improvement cycle: initial proxy block rate of 55% was diagnosed through per-category analysis revealing category fragmentation (19 index categories vs 368 warehouse categories) and 4 poisoned vectors causing false positives. Surgical remediation (correcting the category mapping, removing poisoned vectors, and embedding 3,041 targeted vectors from underrepresented attack domains) improved block rate to 99.9% while reducing false positive rate from 9.0% to 4.8%. This empirical improvement cycle demonstrates the MANAGE function applied to AI defense systems themselves, not just the AI systems being defended.

5. Suggested Action: MAP 1.5

AI RMF Subcategory: MAP 1.5 — Organizational risk tolerances are determined and documented.

Gap Identified

The current Playbook does not address risks specific to multi-agent AI systems where multiple AI components communicate and delegate tasks to one another. Our testing across 69 models demonstrates that cross-agent contamination, where a compromised upstream agent's output is trusted by downstream agents, succeeds at a 23.5% average breach rate, with individual models ranging from 0% to 87.5%. In multi-agent verification patterns (where Agent B reviews Agent A's output for quality assurance), both agents share the same model vulnerabilities, meaning both can agree on a compromised conclusion.

Suggested Action

- For AI systems employing multi-agent architectures, risk mapping should include cross-agent trust boundaries as a distinct risk category. Organizations should assess whether compromised output from one agent can propagate through the pipeline and influence downstream agents' decisions or actions.
- Multi-agent verification patterns (where one agent checks another's work) should not be treated as security controls unless the agents use different underlying models or include independent, non-AI validation steps.

Supporting Evidence

CLS Labs executed 2,875 cross-agent probes across 72 models, achieving a 21.9% average breach rate on cross-agent contamination, multi-agent coordination, and identity spoofing attacks. Breach rates ranged from 0% to 87.5% across models, with cost-optimized models used in agent pipelines (e.g., GPT-4.1-nano) exhibiting the highest rates. Across all 14 agent-specific attack modules, 11,599 probes were executed against 96 models with a 29.3% aggregate breach rate. Full agent security assessment methodology and results are published at https://clsecuritylabs.com/research.html.

6. About the Submitting Organization

Closed Loop Security Labs LLC is a Colorado-registered AI security assessment firm specializing in adversarial red team evaluation and defense of large language models, AI agents, and agentic AI systems. CLS Labs has tested 273+ models across 15+ inference providers, executing 1.5M+ adversarial probes across 368 attack categories in 14 security domains, utilizing 134 custom-built attack modules. The firm maintains 66,334 defense classifier vectors and has conducted 67,000+ severity assessments using three-vendor cross-judge validation. All findings are mapped to OWASP Top 10 for LLMs, MITRE ATLAS, and the NIST AI RMF.

The firm's principal researcher, Brandon Arbour, is a former network security engineer at Verizon with experience in security audit automation and AI integration strategy at the executive level.

All referenced research is publicly available at https://clsecuritylabs.com. This submission constitutes a free resource from a for-profit entity, consistent with NIST AI RMF Playbook inclusion criteria.