NIST AI RMF Use Case Profile

Application of AI RMF Functions to Adversarial Assessment of GPT-4.1 and Agentic AI Systems

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. Organization

Closed Loop Security Labs LLC (CLS Labs), a Colorado-based AI security assessment firm. CLS Labs conducts adversarial red team evaluations of large language models and AI agent systems, with findings mapped to NIST AI RMF, OWASP Top 10 for LLMs, and MITRE ATLAS.

2. AI System Under Assessment

Multiple AI systems assessed, including OpenAI GPT-4.1 (across all variant sizes), GPT-4o, Claude Sonnet 4, Claude Opus 4, Gemini 2.5 Pro/Flash, Llama 3.3-70B, Llama 4 Scout/Maverick, and 260+ additional models. Assessments covered both text-only deployments and tool-connected agent configurations across 15+ inference providers.

3. AI RMF Functions Applied

GOVERN

CLS Labs established organizational policies governing adversarial testing methodology, including: classification of findings using a standardized severity framework (Attack Impact Score across 5 dimensions: Compromise Depth, Action Depth, Privilege Level, Persistence, and Evasion); data handling policies for attack prompts and breach evidence; and a competitive language policy protecting proprietary methodology from public disclosure while maintaining transparency in published results.

MAP

Risk mapping was conducted across 368 attack categories in 14 security domains, organized into a structured taxonomy covering: direct prompt injection, indirect prompt injection (XPIA), information disclosure, jailbreaking, denial of service, supply chain and RAG poisoning, output integrity manipulation, and agentic reasoning attacks. Each category was mapped to OWASP

Top 10 for LLMs, MITRE ATLAS techniques, and relevant NIST AI RMF subcategories. For agent assessments, 134 attack modules covering agent exploitation, MCP protocol attacks, containment escape, OT/ICS, robotics/VLA, autonomous vehicles, supply chain, and infrastructure attacks were mapped as distinct risk categories.

MEASURE

Adversarial testing was executed using a multi-tool assessment pipeline (Forge Engine, Garak, PyRIT, and custom exploit chains) generating 1.5M+ adversarial probes. Results were scored through a three-vendor cross-judge pipeline (Google Gemini, Anthropic Claude, Meta Llama) to eliminate single-vendor evaluation bias. Key measurement finding: standard safety benchmarks reported an 84.3% defense rate for GPT-4.1, while operational adversarial testing across 3,595 probes revealed a 55% breach rate, a 3.5x underreporting of actual risk. For agent assessments, tool-connected testing revealed 2x–9x breach rate increases in well-defended production models compared to text-only evaluation (e.g., Claude Opus 4: 16.2% text-only vs 38.9% with tools; Llama 3.3-70B: 7.5% vs 42.6%; 47 of 101 models showed higher breach rates with tool access).

MANAGE

Verified findings were remediated through a three-tier defense stack: Tier 1 signal filters for known attack patterns (<5ms latency), Tier 2 semantic defense classifier with 66,334 vectors across 368 categories (~15ms latency), and Tier 3 LLM judge for novel threats (~200ms latency). Defense effectiveness was validated empirically by re-running the original attack suite against the defended system, achieving 99.9% block rate across 6,103 test probes spanning 326 categories, subsequently validated at 100% block rate across 448,000+ warehouse records, with a 4.8% false positive rate on benign traffic. Certified Risk Profiles (CRPs) were issued for each verified vulnerability with remediation guidance mapped to specific NIST AI RMF subcategories.

4. Key Outcomes

- Demonstrated that standard safety benchmarks underreport operational breach rates by 3.5x for production deployments
- Documented that tool access increases breach rates by 2x–9x in well-defended production models (47 of 101 models tested)
- Established a reproducible three-vendor cross-judge methodology that eliminates single-vendor evaluation bias
- Built defense classifier achieving 99.9% block rate with 4.8% false positive rate across 368 attack categories, validated at 100% block rate across 448,000+ warehouse records
- Documented a complete empirical defense improvement cycle (diagnosing category fragmentation, removing poisoned classifier vectors, and embedding targeted vectors from underrepresented attack domains), improving block rate from 55% to 99.9% while reducing false positive rate from 9.0% to 4.8%, demonstrating the MANAGE function applied to AI defense systems themselves
- Scaled adversarial testing to 1.5M+ probes across 273+ models, the largest independently produced adversarial assessment dataset publicly documented

- Published all methodology and findings as a free, publicly available resource at clsecuritylabs.com

5. Lessons Learned

- The MEASURE function would benefit from explicit guidance distinguishing between text-only and tool-connected adversarial evaluation, as the risk profiles differ substantially
- Multi-vendor scoring is essential for reliable adversarial evaluation; single-judge breach rates varied from 41.9% to 67.1% for the same dataset
- Provider-level safety filtering on serverless endpoints can create a false sense of security that does not transfer to dedicated or self-hosted deployments
- Cross-agent contamination in multi-agent systems represents an emerging risk category not currently addressed in the AI RMF Playbook
- Defense classifier effectiveness is highly sensitive to category coverage and vector quality. Expanding from 19 categories to 368 categories while surgically removing 4 poisoned vectors improved block rate from 55% to 99.9% and reduced false positive rate from 9% to 4.8%, demonstrating that targeted data quality improvements outperform volume increases in classifier defense

6. Public Resources

- GPT-4.1 Assessment Blog: https://clsecuritylabs.com/blog.html
- Cross-Model Research Results: https://clsecuritylabs.com/research.html
- CLAP Protocol Methodology: https://clsecuritylabs.com/clap.html
- Organization Website: https://clsecuritylabs.com