Red Teaming Generative AI

The New Attack Surface

Manish Pandey
Cybersecurity Professional
linkedin.com/in/manishkp
Maharshi Markandeshwar University | February 2026

Why Red Team GenAI?

> **1.8 Million attacks. 22 frontier models. Every single model broke.**
>
> — UK AI Safety Institute / Gray Swan, 2025

97%

of organizations reported GenAI security issues

Viking Cloud, 2026

72%

surge in AI-assisted cyberattacks

SecurityWeek, 2025

19%

describe their GenAI security posture as confident

Lakera Report, 2025

890%

surge in GenAI traffic across enterprises

Palo Alto Networks

From Networks to Neural Networks

Your security skills transfer — the attack surface is new, the mindset is not

Traditional Security	GenAI Equivalent
SQL Injection	Prompt Injection
XSS (Cross-Site Scripting)	Indirect Prompt Injection
Privilege Escalation	Jailbreaking / Excessive Agency
Supply Chain Attack	Model Poisoning / Rogue MCP
Buffer Overflow	Context Window Overflow
Social Engineering	Persuasion-Based Attacks (PAP)
Reverse Engineering	System Prompt Extraction
Fuzzing	Adversarial Prompt Scanning

The GenAI Attack Surface

Layer 1: Model Layer

Training data, weights, fine-tuning pipeline

Layer 2: Prompt Layer

System prompt, user input, instruction hierarchy

Layer 3: Context Layer

RAG retrieval, vector databases, documents

Layer 4: Integration Layer

Tool calling, APIs, MCP servers, plugins

Layer 5: Agent Layer

Autonomous workflows, planning, memory

More capability = more attack surface. You're testing a system, not just a model.

The Attack Playbook

Attack Category	Target	OWASP
Direct Prompt Injection	Prompt Layer	LLM01
Indirect Prompt Injection	Context Layer	LLM01
Jailbreaking	Safety Alignment	LLM01
System Prompt Extraction	Business Logic	LLM07
Data Exfiltration	Sensitive Data	LLM02
Agent & Tool Exploitation	Integration Layer	LLM06
Training Data Poisoning	Model Level	LLM04
Supply Chain Compromise	Infrastructure	LLM03

CrowdStrike tracks 150+ prompt injection techniques across dual taxonomies: IM##### (delivery) + PT##### (manipulation)

Direct Prompt Injection & Jailbreaking

Attack Techniques

Roleplay Dynamics — 89.6% ASR

Persona-based bypass (DAN, character play)

Logic Traps — 81.4% ASR

Conditional structures, moral dilemma framing

Encoding Tricks — 76.2% ASR

Base64, ROT13, zero-width characters

Persuasion (PAP) — 92% ASR

Authority appeal, emotional manipulation

The Intelligence Paradox

More capable models are MORE susceptible to persuasion attacks

Better contextual understanding = easier to manipulate with well-crafted language

This is social engineering for AI

Keysight (Jan 2026) identified 40 persuasion techniques for jailbreaking

Key techniques:

- Authority appeal
- Emotional manipulation
- Reciprocity & rapport building
- Scarcity / urgency framing

Indirect Prompt Injection — The XSS of AI

The Attack Chain

1 **Attacker plants instructions**

Email / website / document

2 **AI assistant reads content**

Background processing

3 **AI follows hidden commands**

User never sees attack

4 **Data exfiltrated or actions taken**

MFA codes, emails, files

Real-World Incidents

Microsoft Copilot / Outlook	Hidden prompt in phishing email exfiltrated MFA codes via Graph API	2025
Reprompt Attack	Single-click data exfiltration from Copilot Personal via legitimate MS link	Jan 2026
Perplexity Comet AI	Malicious email hijacked AI assistant, exfiltrated inbox data	2025
GitHub Copilot Chat	CVE-2025-53773 — CVSS 9.6 RCE via prompt injection	2025

Agent & Tool Exploitation

MITRE ATLAS Oct 2025 Update

14 New Agentic Techniques

Context Poisoning
Manipulate agent context to influence decisions

Memory Manipulation
Alter long-term memory for persistent influence

Thread Injection
Inject malicious instructions into chat threads

Config Tampering
Modify agent config to affect all shared agents

RAG Credential Harvesting
Extract credentials from RAG databases

Tool Invocation Exfil
Leverage agent tools to exfiltrate data

Real-World Agent Exploits

Rogue MCP Server
Injected malicious code into Cursor IDE browser

CVE-2025-59944
Zero-click RCE in MCP-integrated IDEs

Prompt-to-SQL (P2SQL)
Protocol-layer injection via GitHub MCP server

DeepSeek Crisis (Jan 2026)
Exposed databases, user data, API keys — government bans worldwide

Package Hallucination
LLM recommends non-existent packages — attacker registers them with malware

MITRE ATLAS — The ATT&CK for AI

15	66	46	26	33
Tactics	Techniques	Sub-techniques	Mitigations	Case Studies

Complementary Frameworks

OWASP Top 10 for LLMs (2025)	Application-specific risks + mitigations	`LLM-specific`
OWASP Red Teaming Guide	Structured methodology for AI red teaming	`Methodology`
OWASP Top 10 Agentic (2025)	Risks specific to autonomous AI agents	`Agent risks`
NIST AI RMF + IR 8596	Governance + mapping AI risks to cybersecurity	`Governance`
MITRE SAFE-AI	Maps ATLAS threats to NIST 800-53 controls	`100 controls`
EU AI Act (Aug 2026)	Mandatory adversarial testing for high-risk AI	`Regulatory`

~70% of ATLAS mitigations map to existing security controls — your SOC knowledge applies.

The Red Teamer's Toolkit

Microsoft PyRIT

= Metasploit for AI

- Multi-turn attack orchestration
- Audio, image & math converters
- Azure Content Safety scoring
- AI Red Teaming Agent (Apr 2025)
- Integrated into Azure AI Foundry

NVIDIA Garak

= Nessus for AI

- 120+ vulnerability categories
- Plugin architecture for custom probes
- Open-source (NVIDIA)
- v0.14.0 adding agentic AI support
- Led by OWASP LLM Top 10 core member

Tool	Analogous To	Best For
Promptfoo	OWASP ZAP	CI/CD integrated testing
DeepTeam	Burp Suite	Probe-based testing
FuzzyAI	AFL / Radamsa	Fuzzing for unknown vulns
ATLAS Navigator	ATT&CK Navigator	Threat modeling

Live Demo: Red Teaming in Action

NVIDIA Garak

Vulnerability Scanner for LLMs

What We'll Run:

- Automated probe scan against a local model
- Jailbreak detection across 120+ categories
- Data leakage and hallucination probes

Key Commands:

```
garak --model_type ollama
  --model_name llama3
  --probes all
```

Watch For:

- HTML report with pass/fail by vulnerability class
- How quickly automated scans surface issues manual testing might miss

Promptfoo

Red Team & Eval Framework

What We'll Run:

- Red team evaluation with custom policies
- Prompt injection & jailbreak test suites
- Side-by-side model comparison

Key Commands:

```
promptfoo redteam init
promptfoo redteam run
promptfoo view
```

Watch For:

- Interactive web UI with vulnerability scoring
- CI/CD integration — red teaming as part of the deployment pipeline

Both tools are open-source. Garak = scanner (find vulnerabilities). Promptfoo = evaluator (test policies & compare models).

Red Team Engagement Methodology

01

Recon & Scoping

- Define system under test
- Map trust boundaries
- Extract system prompt
- Map to ATLAS techniques

02

Attack Planning

- Select attack categories
- Design multi-step chains
- Prepare poisoned artifacts
- Choose manual vs automated

03

Execution

- Manual adversarial prompts
- Automated Garak/PyRIT scans
- Multi-turn rapport attacks
- Indirect injection testing

04

Report & Remediate

- Document with ATLAS IDs
- Classify AI-specific severity
- Architectural mitigations
- Purple team handoff

Defenses — What Actually Works

Instruction Hierarchy	System > developer > user — enforce at orchestration layer
Context Isolation	Separate trusted instructions from untrusted retrieved content
Input Scanning	Pre-inference detection of jailbreak patterns (regex + ML)
Output Validation	Filter/sanitize LLM outputs before rendering or executing
Tool-Call Gating	Validate agent actions against allowlists — never auto-execute
Least Privilege	Agents get minimal permissions, scoped to specific tasks
Continuous Red Teaming	Automated scans in production, canary prompts, bug bounties

No complete defense against prompt injection exists today — it is an active arms race.

The AI Security Pipeline

Defense at every stage — from model selection to production monitoring

Pre-Deployment	Development & Testing	Runtime & Production
ModelScan ML model supply chain scanner (Protect AI)	**PyRIT** Multi-turn attack orchestration (Microsoft)	**LLM Guard** 15 input + 20 output scanners (Protect AI)
HiddenLayer AI supply chain security + tamper detection	**Garak** 120+ vuln category scanner (NVIDIA)	**NeMo Guardrails** Topic control, PII, RAG grounding (NVIDIA)
PS-Fuzz System prompt fuzzing (Prompt Security)	**DeepTeam** 40+ vulns, OWASP Top 10 aligned	**Lakera Guard** Real-time LLM protection (now Check Point)
PurpleLlama Llama Guard 4, Prompt Guard 2, CyberSecEval 4 (Meta)	**Promptfoo** CI/CD integrated red teaming (30K+ devs)	**Arthur Shield** First firewall for LLMs — 1B+ tokens monitored
Dataset Filtering Training data validation + poison detection	**ARTKIT** Multi-turn automated red teaming	**WhyLabs LangKit** Monitoring: quality, relevance, jailbreak detection

Defense in depth: no single tool is sufficient — layer scanning, guardrails, and monitoring across the entire pipeline.

Guardrails & Defense Stack

Key Frameworks & Platforms

Platform	Type	Key Capabilities	2025-26 Update
NVIDIA NeMo Guardrails	Open-Source	Topic control, PII detection, RAG grounding, jailbreak prevention, multimodal	*LangChain + Cisco AI Defense integration*
Meta PurpleLlama Suite	Open-Source	Llama Guard 4 (multimodal), LlamaFirewall, Prompt Guard 2, CyberSecEval 4	*75% latency reduction with Prompt Guard 2-22M*
Guardrails AI	Open-Source	Community-driven validators, RAIL specification, PII/toxicity/hallucination detection	*Guardrails Index benchmarks 24 guardrails*
LLM Guard	MIT License	35 scanners (15 input + 20 output), prompt injection, PII, secrets, toxicity	*Single API call deployment*
Prompt Security	Enterprise	GenAI authorization framework, desktop agent for Copilot 365, feature-level control	*Industry-first AI authorization (Mar 2025)*
HiddenLayer	Enterprise	Supply chain + runtime defense, airgapped environments, attack simulation	*US MDA SHIELD contract (Dec 2025)*

Cloud-Native AI Security

Azure AI Content Safety	Prompt Shields + multimodal filtering	Real-time
AWS Bedrock Guardrails	6 policies + Automated Reasoning	88% block / 99% acc
Google Vertex AI Safety	Jailbreak classifier + configurable filters	Gemini 2.5 Flash

2025-26 Industry Consolidation

Lakera → Check Point	Q4 2025
Robust Intelligence → Cisco	2025
CalypsoAI → F5 Networks	Jan 2026
Protect AI Guardian → Palo Alto Networks	2025

AI security market is consolidating fast — major security vendors are acquiring specialized AI defense startups.

Key Takeaways

1. **Language is the new exploit vector**
   GenAI attack surface is fundamentally different

2. **Every frontier model breaks**
   Red teaming is essential, not optional

3. **Indirect injection = XSS of AI**
   Untrusted data treated as trusted instructions

4. **Agentic AI explodes the attack surface**
   Tools + autonomy + memory = compounding risk

5. **Your cybersecurity skills transfer**
   Attacker mindset + methodology + ATLAS framework

6. **The field is wide open**
   Frameworks exist, tools exist — practitioners are scarce

Industry Voices to Follow

The researchers and leaders shaping AI security — your reading list starts here

Red Teaming & Offensive Research

Johann Rehberger — *Independent Researcher*
Indirect prompt injection pioneer; Month of AI Bugs (2025)

Ram Shankar Siva Kumar — *Microsoft AI Red Team*
Founded MS AI Red Team; PyRIT creator; author "NOT WITH A BUG"

Leon Derczynski — *NVIDIA / ITU Copenhagen*
Garak creator; OWASP LLM Top 10 core team

Kai Greshake — *NVIDIA*
Foundational indirect prompt injection paper (2023)

Sven Cattell — *AI Village / DEF CON*
Founded AI Village; largest live AI hacking events

Simon Willison — *Independent Developer*
Coined “prompt injection”; prolific AI security blogger

Pliny the Liberator — *Independent (@elder_plinius)*
Frontier model jailbreaks; TIME 100 AI 2025

AI Safety & Defense Leaders

Steve Wilson — *OWASP / Exabeam*
Project lead: LLM Top 10 & Agentic Top 10

Daniel Miessler — *Unsupervised Learning*
AI security newsletter (700K+ followers); Fabric framework

Bruce Schneier — *Harvard Kennedy School*
Security thought leader; AI risk & policy analysis

Mark Russinovich — *Microsoft Azure CTO*
Enterprise AI security strategy; Azure Deputy CISO

Elie Bursztein — *Google DeepMind*
Sec-Gemini lead; AI cybersecurity research (60+ papers)

Hyrum Anderson — *Cisco (ex-Robust Intel.)*
Cisco AI Security Framework; MITRE ATLAS contributor

Daniel Fabian — *Google*
Head of Google Red Teams; built ML Red Team from ground up

Follow their blogs, GitHub, newsletters, and talks — this is where the field moves fastest

Your Path Forward

From cybersecurity student to AI red teamer — a concrete roadmap

This Week

- Play Gandalf by Lakera (prompt injection game)
- Install Ollama + a small local model
- Run Garak scan on your local model
- Read OWASP Top 10 for LLM Apps 2025

This Month

- Complete Damn Vulnerable LLM Agent (DVLA)
- Practice on HackTheBox AI challenges
- Study MITRE ATLAS Navigator at atlas.mitre.org
- Write your first PyRIT attack orchestration

This Quarter

- Build a red team report on a real AI app
- Contribute to Garak / DeepTeam open-source
- Get OSCP or equivalent security certification
- Publish findings on a blog / LinkedIn / GitHub

Career Targets

- AI Red Team Engineer — Microsoft, Google, Meta
- ML Security Researcher — Anthropic, OpenAI
- AI Safety Analyst — CrowdStrike, Palo Alto, Cisco
- Bug Bounty — AI-specific programs now paying $25K+

The field is < 5 years old. There is no queue. Start now.

GenAI & Security Research at MMU

Pioneering work already happening at Maharshi Markandeshwar University

AI Security & Safety

Dr. Poonam Sharma — *Assoc. Prof, CSE*
Secure Protocols on Biomedical Smart-Devices; GenAI Capabilities & Limitations

Dr. Meenakshi Garg — *Trust Member / Faculty*
Cyber Security Based ASIC for Epileptic Seizure Prediction (HW-embedded security)

Dr. Sarvesh Tanwar — *Alumni / Collaborator*
DITrust Chain; Blockchain for Data Security (critical for AI provenance)

Manju Bagga — *Asst. Prof, MCA*
IoT Security, Deep Learning; Security Hazards & Countermeasures

Dr. Deepak Dudeja — *Professor, CSE*
Medical AI Security; Co-author on Cyber Security Based ASIC research

Dr. Gulbir Singh — *Ex-MMU Collaborator*
Scary Dark Side of AI; Generative AI Capabilities and Limitations

Rohit Chauhan — *MMIM (Mgmt)*
AI-powered travel planning: exploring anthropomorphism and privacy

GenAI Research & Applications

Dr. Vivek Bhatnagar — *Head, MMICT&BM*
Ethical Implications of GenAI; Exploring Creative Capacities of GenAI

Gourav Kalra — *Mech. Engineering*
Author: Campus Bot Revolution (Book on RAG & LLM implementation)

Dr. Arun Kumar Gupta — *Mech. Engineering*
Co-author: Campus Bot Revolution — Applied GenAI

Bikram Jit Singh — *Mech. Engineering*
Co-author: Campus Bot Revolution — Applied GenAI

Dr. Suneet Kumar — *Professor, CSE*
GenAI Technical Architecture; Co-author on Ethics & Creativity papers

Guarav Kumar — *Faculty, MMU*
Ethics of Artificial Intelligence in the Education Sector

Shaveta Jain — *PhD Scholar, CSE*
Efficient Deep Learning models for disease detection

Collaboration between industry red teaming and academic research strengthens both

Your Turn to Red Team

Questions & Discussion

Challenge: Try This Week

- Try Gandalf by Lakera — free prompt injection game
- Install Garak and promtfoo and scan a local model and application interacting with model
- Explore MITRE ATLAS Navigator at atlas.mitre.org
- Document what you find — think about WHY it worked

Manish Pandey | Cybersecurity Professional | linkedin.com/in/manishkp

Special thanks to Tejbir Rana Sir for facilitating this session