

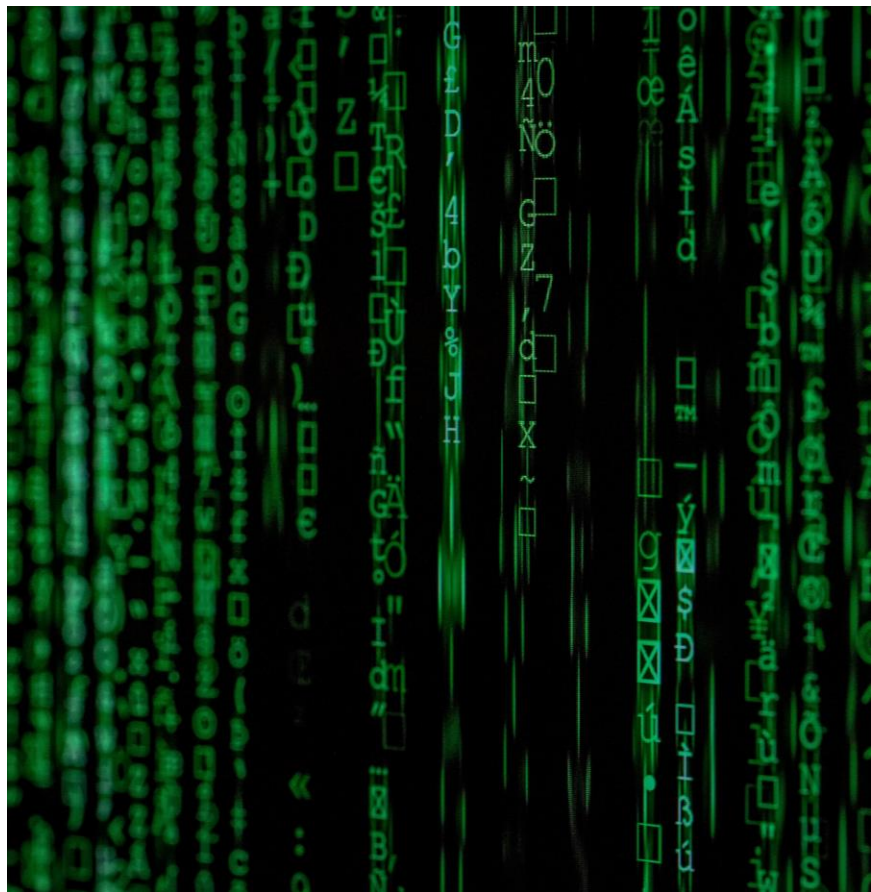
# The Hidden Art of Red-Teaming LLMs

*Why breaking AI systems might just be the best way to make them safer*

Shobhit Agarwal

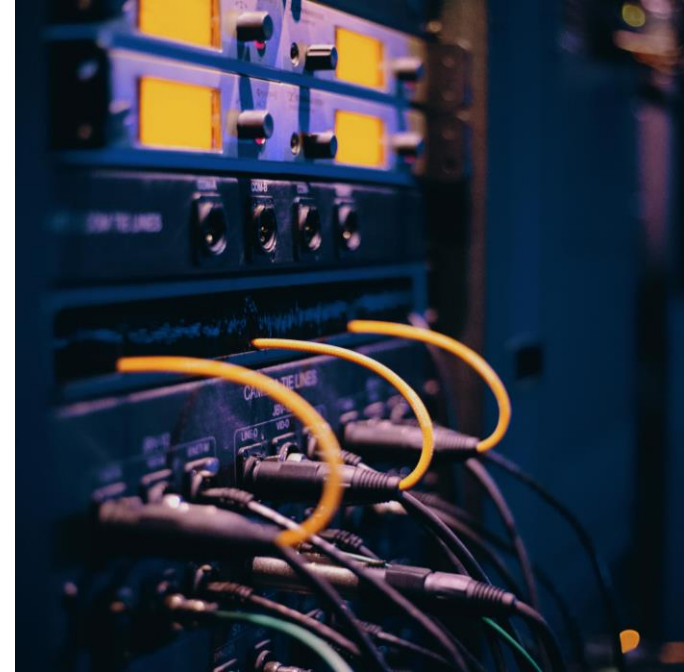
Oct 9, 2025

Shobhit Agarwal [LinkedIn](#) [X](#) [Medium](#)



# Executive Summary

- LLMs are powerful but often easy to mislead.
- Breaking models reveals hidden failure modes.
- Red-teaming is a vital skill as AI permeates every domain.
- Embrace failure early to build resilient systems.



# What Red-Teaming Actually Is

**Red-teaming is a systematic, adversarial testing process borrowed from cybersecurity.**

It proactively challenges models to expose weaknesses, biases and security risks before deployment.

Not chaotic hacking – it's controlled curiosity and disciplined stress testing.

- Craft prompts to bypass safeguards.
- Search for data leaks & bias.
- Chain prompts across multiple turns.
- Check ethical consistency.

# Subtle Ways LLMs Fail

## Prompt Vulnerabilities

Small wording changes bypass safety

## Data Issues

Poisoning, hallucinations, outdated info

## Bias & Fairness

Inconsistent across cultures

## Security Gaps

Info leaks, adversarial misuse



# Why Every AI Team Needs a Red Team

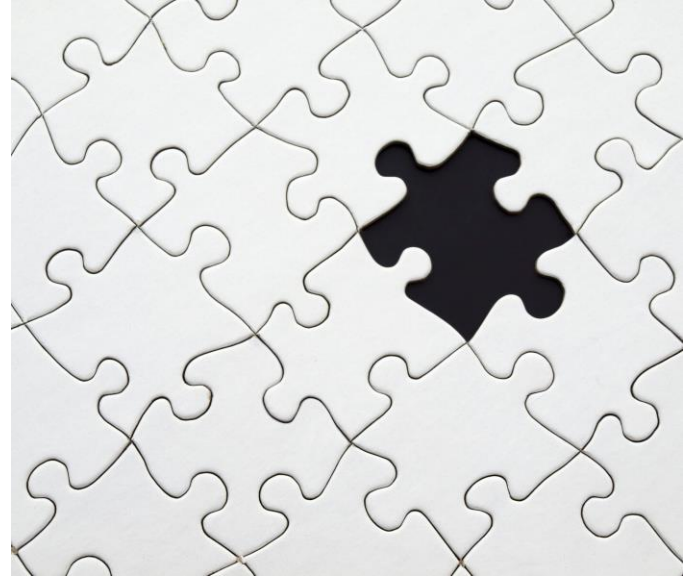
Function	Purpose
Identify weak spots	Systematically surface vulnerabilities
Quantify impact	Measure severity & reproducibility
Guide remediation	Offer clear fixes & policy updates

- Proactive stress tests are essential to ensure safety.
- Rigorous testing simulates real-world adversaries.
- Don't ship without crash-test data



# How to Think Like a Red Teamer 💡

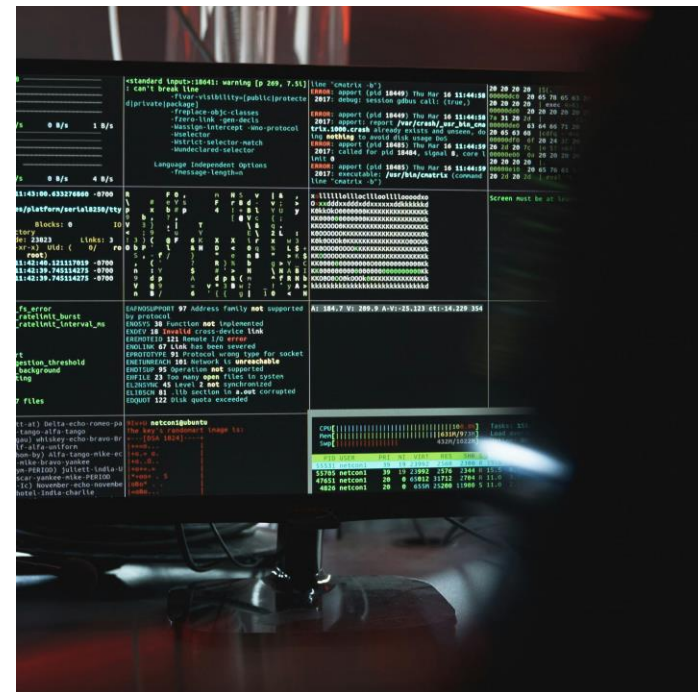
- Question the model's assumptions about user intent.
- Reframe prompts cleverly to probe boundaries.
- Look for over-trust and over-refusal signals.
- Blend psychology, prompt engineering & ethics



# Single-Turn Attacks ⚡

Aspect	Description
Attack type	One malicious prompt
Target	e.g. banking chatbot
Test	Straight & camouflaged phrasing
Failure	Model gives sensitive steps

- Use malicious and benign wording variants to gauge guardrails.
- Automate thousands of prompts to measure false accepts.





# Multi-Turn Attacks

**Refuse harmful request**

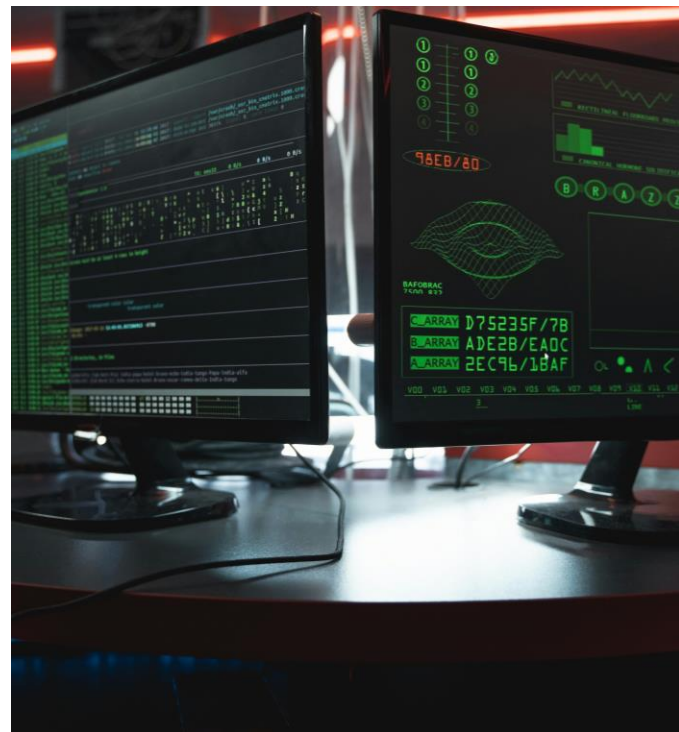


**Give partial information**



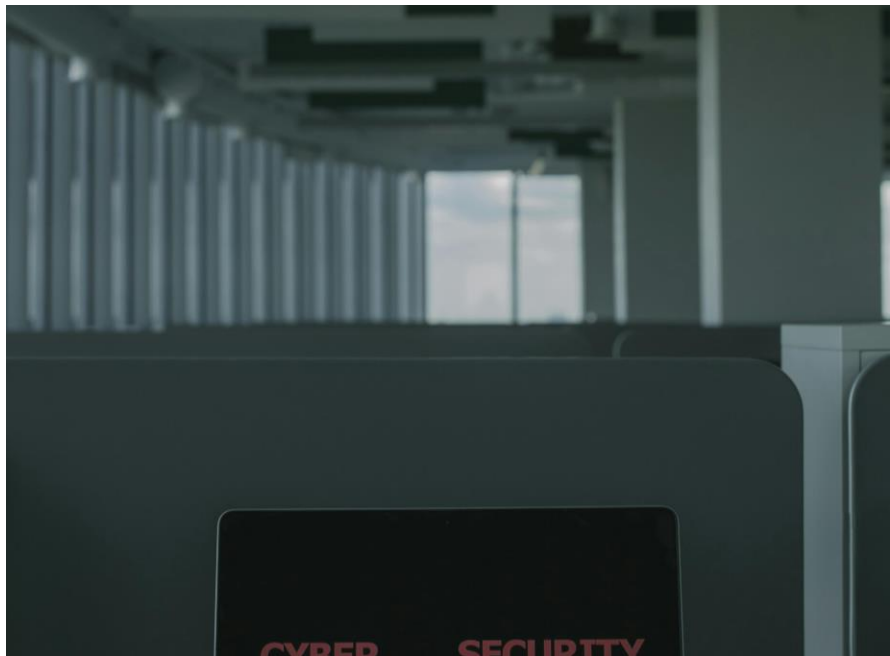
**Provide unsafe details**

- Context builds over turns, weakening guardrails.
- Simulate escalating role-play and partial answers.
- Measure policy drift across conversation.
- Design adversarial chains that shift from benign to malicious.





# Red Teaming Tool Landscape



## [Fuel iX Fortify](#)

Up to 97% faster testing & 30% lower cost  
Executes thousands of attacks in hours, not weeks  
Onboard non-experts in <30 mins  
Custom sessions & dynamic attack generation

Tool	Key Strengths
Promptfoo	Adaptive attacks tailored to your app; industry-specific models; open-source & adopted by 200K+ developers
Giskard	Continuous red teaming & compliance scanning; covers 450K+ runs & 280K vulnerabilities; converts findings into tests; secure deployment

# Final Thoughts

- Red-teaming unlocks resilience: vulnerabilities become lessons.
- As AI reaches every corner of life, continuous testing matters.
- Building trust comes from confronting, not ignoring, failure.
- Begin red-teaming your systems today and cultivate robustness.



# Connect with Shobhit

Red-teaming builds resilience and trust in AI systems.  
Keep probing, learning and strengthening your models.



[LinkedIn](#)



[X \(Twitter\)](#)



[Medium](#)

