TESTING, EVALUATION & RED-TEAMING

CYB-4203/6203: SECURE AND TRUSTWORTHY AI

Unit 9 — Monday, April 13, 2026

Dallas Elleman — Spring 2026

Interactive version at
https://dallaselleman.github.io/cyb-4203-6203-spring-2026/course_materials/presentations/revealjs/pres-18.html

TODAY'S ROADMAP

9.1 Security Testing

Labs, frameworks & real attacks

→

9.2 Evaluation

Benchmarks, metrics & evals

→

9.3 Red-Teaming

Wednesday — we break things

THE STORY SO FAR

Frameworks you already know from Unit 6

OWASP Top 10 for LLMs

LLM-specific vulnerability ranking

genai.owasp.org

MITRE ATLAS

ATT&CK-style matrix for AI threats

atlas.mitre.org

CSA MAESTRO

7-layer agentic AI security architecture

cloudsecurityalliance.org

AIUC-1

First AI agent certification standard

aiuc.com

A BRIEF HISTORY OF AI SECURITY TESTING

2004–2012
Foundations
"Can Machine Learning Be Secure?" (Barreno et al., 2006)

2013–2016
Adversarial Examples
Szegedy discovers neural nets are brittle; Goodfellow creates FGSM

2017–2018
Physical Threats
Adversarial stop signs fool self-driving cars (Eykholt et al.)

2019–2021
Frameworks
MITRE ATLAS, NIST taxonomy, OWASP begins ML work

2022–2023
The LLM Era
ChatGPT launches; prompt injection, jailbreaks go viral; EO 14110

2024–2026
Regulation
AI Safety Institutes, EU AI Act, frontier model testing required

FRONTIER LABS: THE TRANSPARENCY SPECTRUM

Anthropic	OpenAI	Google DeepMind	Meta	xAI
RSP v3.0 + ASL Levels	Preparedness Framework v2	Frontier Safety Framework v3	Purple Llama (open tools, closed process)	Risk Management Framework (late, incomplete)

Most Transparent — Least Transparent

Not all labs treat security testing the same way.

ANTHROPIC: RESPONSIBLE SCALING POLICY

ASL-1 No meaningful risk (chess AI, 2018-era LLMs)

ASL-2 Early dangerous capabilities (current Claude models through Sonnet 4)

ASL-3 Substantial misuse risk (Claude Opus 4 — activated May 2025)

ASL-4+ Not yet defined

Modeled after biosafety levels (BSL) | NNSA/DOE nuclear security partnership | Constitutional Classifiers: 3,000+ hrs red teaming, no universal jailbreak found

anthropic.com/responsible-scaling-policy

PROJECT GLASSWING

(April 7, 2026)

Thousands

of zero-day vulnerabilities found across every major OS and browser

27 yrs

oldest bug discovered — in OpenBSD

$100M+

in model credits and donations to open-source security

4 vulns

chained into one browser exploit: JIT heap spray, sandbox escape, privilege escalation

Claude Mythos Preview — deemed "too dangerous to release"

anthropic.com/glasswing red.anthropic.com/2026/mythos-preview/

OPENAI & GOOGLE DEEPMIND

OPENAI

Preparedness Framework v2

- GPT-5: 5,000+ hours red teaming
- 400+ external testers
- System cards as industry standard
- Safety Evaluations Hub

openai.com/safety/evaluations-hub

GOOGLE DEEPMIND

Frontier Safety Framework v3

- New CCLs: manipulation, deceptive alignment, shutdown resistance
- CART: 350+ exercises in 2025
- Automated Red Teaming for Gemini
- 300+ published safety papers

deepmind.google/frontier-safety-framework

META & XAI: A STUDY IN CONTRASTS

META

Open Tools, Closed Process

- Purple Llama: CyberSecEval 4, LlamaFirewall, GOAT
- More reusable safety infrastructure than any other lab
- Safety team warnings overridden for Llama 4
- Fabricated benchmark results surfaced pre-launch

XAI

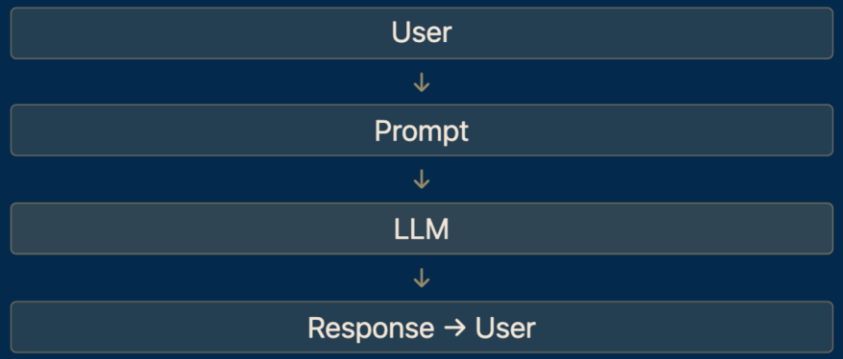
Consistently Behind

- Safety reports late or missing
- Grok 4 launched without system card
- CSAM generation incident (Jan 2026)
- Stolen code repo, leaked user conversations

Open-source tooling does not equal responsible process. Missing reports does not equal missing risks.

AGENT SECURITY: A NEW ATTACK SURFACE

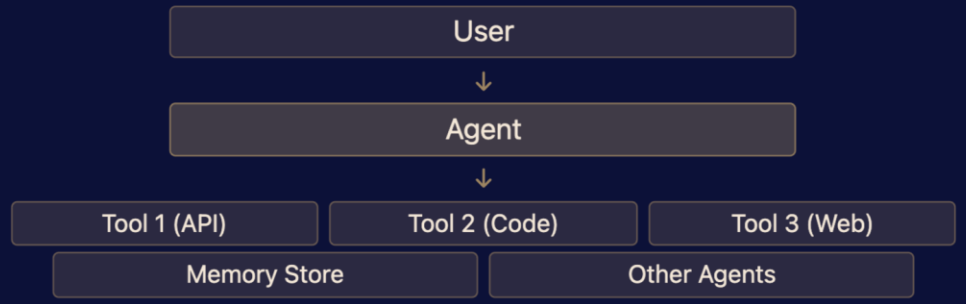
SIMPLE LLM CHAT



3 attack vectors

- Direct prompt injection
- Jailbreaking
- Training data extraction

AGENTIC AI SYSTEM



8+ attack vector categories

- Indirect prompt injection
- Tool poisoning
- Privilege escalation
- Memory poisoning
- Inter-agent manipulation
- Credential abuse
- Supply chain attacks

78% of breached agents had over-permissioned access

43% of public MCP servers contain injection flaws

DEMONSTRATED AGENT ATTACKS

Tool Poisoning

CrowdStrike, 2025 — CVE-2025-6514

- `add_numbers` tool with hidden instruction to exfiltrate SSH keys
- 84.2% success rate with auto-approval
- Real CVEs assigned

Indirect Prompt Injection

Palo Alto Unit 42 — December 2025

- Hidden instructions in ad content tricked AI ad-review system
- Attacker used multiple injection methods simultaneously
- First documented real-world IDPI

Memory Poisoning

MINJA — NeurIPS 2025 (Dong et al.)

- 95%+ success via query-only interaction
- Poison in February, exploit in April
- Works cross-session

Reference: OWASP Top 10 for Agentic Applications (Dec 2025) — genai.owasp.org

9.2: EVALUATIONS

EVALUATION FOUNDATIONS

	Predicted Malicious	Predicted Benign
Actually Malicious	TP	FN
Actually Benign	FP	TN

Precision = TP / (TP + FP)

"Of everything I flagged, how much was real?"

Recall = TP / (TP + FN)

"Of everything real, how much did I catch?"

F1 = 2 · P · R / (P + R)

"The balance between the two"

WHEN METRICS LIE: THE BASE RATE FALLACY

1. A malware classifier with **99% accuracy**. Sounds great, right?

2. But only **0.1% of files are actually malicious**.

3. **1,000,000 files scanned**
   1,000 malicious → **990 caught (TP)**, 10 missed (FN)
   999,000 benign → 989,010 correct (TN), **9,990 false alarms (FP)**

91%
of alerts are false positives

9.0%
Precision

This is why SOC analysts drown in alerts. Alert fatigue is a direct consequence of the base rate problem.

WHY TRADITIONAL METRICS BREAK FOR LLMS

No Closed Label Set

> "Write me a summary of this paper"

Output A | Output B | Output C

What is a "false positive" for an open-ended question?

Semantic Equivalence

"The cat sat on the mat"

vs.

"A feline rested atop the rug"

0% string match.
100% semantic match.

Quality is Multidimensional

A response can be correct but unhelpful,
helpful but unsafe.

We need new tools. Enter: Evals.

THE RISE OF "EVALS"

HOW WE GOT HERE

Pre-2023	Test set + accuracy = done
2023	OpenAI open-sources Evals framework — the term sticks
Now	Multi-dimensional evaluation is the standard

5 DIMENSIONS

Capability — Can it do the task?

Reliability — Does it perform consistently?

Safety — Does it refuse harmful requests?

Alignment — Does it follow instructions?

Robustness — Does it resist adversarial pressure?

LLM-as-Judge

"Use a strong model to evaluate a weaker model's output"

Known biases: position bias • verbosity bias • self-preference bias

EVAL TOOLS: THE LANDSCAPE

Promptfoo

Open-source eval CLI

- 13,200+ GitHub stars
- YAML config, visual comparison
- Built-in red teaming: 50+ vulnerability categories

We're demoing this next.

promptfoo.dev

Garak (NVIDIA)

nmap for LLMs

- Vulnerability scanner
- Probes: hallucination, data leakage, prompt injection, jailbreaks
- Maps to AI security frameworks

garak.ai

Inspect AI (UK AISI)

What governments use

- 100+ pre-built evaluations
- Tested 30+ frontier models
- Open-source, Python-based

inspect.aisi.org.uk

PROMPTFOO: HOW IT WORKS

YAML Config → promptfoo eval → Visual Comparison

```yaml
prompts:
  - "You are a helpful assistant. Answer: {{question}}"
providers:
  - openai:gpt-4o-mini
  - anthropic:messages:claude-3-5-haiku-20241022
tests:
  - vars:
      question: "What is SQL injection?"
    assert:
      - type: llm-rubric
```

Let's see it in action.

LIVE DEMO: COMPARING MODEL SAFETY TRADEOFFS

LIVE DEMO

promptfoo eval → promptfoo view

Comparing GPT-4o-mini vs Claude 3.5 Haiku on security-relevant prompts

EXAMPLE BENCHMARKS

Category	Benchmark	What It Measures
General	MMLU	57 subjects, knowledge breadth
General	HumanEval	Code generation (Python)
General	TruthfulQA	Tendency to reproduce misconceptions
Safety	BBQ	Social bias in Q&A (9 dimensions)
Safety	ToxiGen	Implicit hate speech (13 groups)
Security	HarmBench	510 behaviors, automated red teaming
Security	JailbreakBench	Jailbreak tracking (NeurIPS 2024)

These are what you see cited in system cards and model releases.

WHY YOU SHOULDN'T TRUST BENCHMARKS

Contamination

52–57%

exact match rates for GPT models on MMLU — the model memorized the test

(Deng et al., NAACL 2024)

Goodhart's Law

"When a measure becomes a target, it ceases to be a good measure."

Saturation

MMLU and HellaSwag at **95%+** — they no longer differentiate frontier models

Gaming

Fine-tune on benchmark data. Scores go up. Real capability? Unchanged.

WHO EVALUATES THE EVALUATORS?

Reference: "When Scanners Lie" (2026)

Evaluator design influences reported attack success rate

"The tools we use to measure safety have their own failure modes."

"Different evaluators produce different scores for the same model on the same attacks."

WEDNESDAY: RED-TEAMING DEEP DIVE

1. Red teaming as a discipline: history, methodology, roles
2. Structured red teaming methodologies and frameworks
3. Hands-on: red teaming exercise
4. Final project assigned: group red-teaming exercise

Come ready to break things.

KEY TAKEAWAYS

1. AI security testing has evolved from theoretical (2004) to regulatory requirement (2024+)

2. Frontier labs vary dramatically in transparency — from Anthropic's Glasswing to xAI's missing reports

3. Traditional metrics break down for LLMs — the "evals" paradigm is the new standard

4. Benchmarks are necessary but insufficient — you can't benchmark your way to safety

REFERENCES & FURTHER READING

FRONTIER LAB SAFETY

anthropic.com/glasswing
Project Glasswing

anthropic.com/responsible-scaling-policy
Anthropic RSP

red.anthropic.com
Anthropic Frontier Red Team

openai.com/safety/evaluations-hub
OpenAI Safety Evaluations

deepmind.google/safety
DeepMind Safety

AGENT SECURITY & FRAMEWORKS

genai.owasp.org
OWASP Agentic Top 10

atlas.mitre.org
MITRE ATLAS

EVAL TOOLS & RESEARCH

promptfoo.dev
Promptfoo

garak.ai
Garak (NVIDIA)

inspect.aisi.org.uk
Inspect AI (UK AISI)