GENERATIVE AI · TAUGHT WITH ANTHROPIC

# AI Evaluation and Red Teaming

Your own eval set, judges you have calibrated, and regression tests in CI. Then the other half: prompt injection, jailbreaks, and the guardrails that do not hold as well as vendors claim.

Advanced | 5 phases | 25 sessions | 38 contact hours

## Syllabus

5 phases, 25 sessions. Each phase ends in something you have built.

### 1 Why the Score Lies

Stop trusting public numbers and build a private eval set from your own failures.

6 sessions
$87 alone

**1.1 The Saturated Leaderboard**

MMLU and GPQA now sit near the ceiling, so a five-point spread ranks nothing about your product.

FREE PREVIEW

**1.2 Contamination, Measured**

Canary strings, n-gram overlap, and verbatim recall probes on the benchmark you were about to trust.

**1.3 The Leaderboard Illusion**

Private variant testing and silent deprecation, and why an Arena rank is not a procurement decision.

**1.4 Error Analysis First**

Read one hundred real traces and open-code the failures before you write a single test case.

**1.5 A Failure Taxonomy**

Group the open codes into categories a grader can score, with counts so you know what to fix first.

**1.6 Dev Split, Frozen Gate**

Two splits, one you iterate against and one only CI may run, plus items dated after the training cutoff.

By the end: **A 120-case private eval set, split into dev and a frozen gate, with a written contamination check.**

## 2 Graders and Judges

Get a scorer whose numbers survive someone asking how you know.

5 sessions
$87 alone

**2.1 Deterministic Graders First**
Exact match, schema validation, and unit tests, plus reshaping a task so code can score it at all.

**2.2 LLM as Judge**
Pointwise, pairwise, and binary rubrics, and why a one-to-five Likert judge returns mostly noise.

**2.3 Measuring Judge Bias**
Quantify position, verbosity, self-preference, and same-family preference leakage on your own set.

**2.4 Humans and Agreement**
Two annotators, Cohen's kappa, and the rubric rewrite that pulls a 0.41 up to something usable.

**2.5 Aligning the Judge**
Pick a threshold from TPR and FPR against human labels, then catch criteria drift when the rubric moves.

By the end: **A judge aligned to two human annotators, reported with a confusion matrix and a kappa you can defend.**

## 3 Agents, CI and Production

Score whole agent runs and wire the result into the pipeline that ships them.

5 sessions
$87 alone

**3.1 Trajectory Versus Outcome**
Score the final state and the path taken, and catch the run that passed for the wrong reason.

**3.2 Tool Call Accuracy**
Argument-level scoring, wrong-tool rates, and pass^k across repeated runs of the same task.

**3.3 Evals in CI**
A promptfoo and pytest gate on every prompt and model bump, under a fixed token and wall-clock budget.

**3.4 Migrating a Model**
Re-baseline when the model ID changes, and find the regression no public leaderboard would show you.

**3.5 Online Evaluation**
Sample live traces into the eval set, and read thumbs-down base rates against guardrail metrics.

By the end: **A CI gate that blocks a merge on eval regression, plus an online scorer running on sampled live traffic.**

## 4 Red Teaming

Attack your own system with a stated budget, then fix it somewhere the model cannot be talked out of.

6 sessions
$87 alone

**4.1 Attack Budget**
Single-shot success rates are marketing, best-of-N at ten thousand tries is the real threat model.

**4.2 Jailbreaks in Practice**
Many-shot, crescendo, and obfuscation, run through garak and PyRIT against your own endpoint.

**4.3 Indirect Prompt Injection**
Attacker text arriving through retrieval, email, tool output, and MCP tool descriptions you approved.

**4.4 AgentDojo, Honestly**
Run the benchmark, then separate a real refusal from an agent too weak to finish the attack.

**4.5 Defence by Architecture**
Dual-LLM isolation, CaMeL-style provenance, typed tools, and authorization enforced outside the model.

**4.6 Limits of Guardrails**
Measure a guard model's recall at your false-positive budget, then adapt an attack straight past it.

By the end: **A red-team report with attack budgets, success rates before and after, and one architectural fix that held.**

## 5 Evidence and Obligations

Turn the testing you did into the documents a regulator or a customer will ask for.

3 sessions
$87 alone

**5.1 What Is Actually Binding**
GPAI duties and Article 50 transparency are live, while the high-risk dates moved to December 2027.

**5.2 The Article 15 Pack**
Poisoning, evasion, confidentiality, and injection tests written up so an auditor can rerun them.

**5.3 Cards and Incidents**
A model card that indexes evidence, and a runbook with the fifteen, ten, and two day clocks.

By the end: **A signed evidence pack: model card, attack-class coverage table, and an incident runbook with named owners.**

## Tools you will use

Anthropic Claude API | Inspect AI 0.3 | promptfoo 0.121 | DeepEval 4 | Langfuse 4 | Braintrust | NVIDIA garak 0.15 | Microsoft PyRIT 0.14 | AgentDojo | OpenTelemetry GenAI semconv

## What you will build

1. **The Private Eval Set**
   120 real failures, a frozen gate, and a judge aligned to human labels
2. **The CI Gate**
   Trajectory scoring, a merge blocked on regression, scorers on live traffic
3. **Red Team and Evidence Pack**
   Attack budgets, rates before and after, a card an auditor can rerun

## Registration

| | |
|---|---|
| Phase 1: Why the Score Lies | $87 |
| Phase 2: Graders and Judges | $87 |
| Phase 3: Agents, CI and Production | $87 |
| Phase 4: Red Teaming | $87 |
| Phase 5: Evidence and Obligations | $87 |
| **Whole track, all 5 phases** | **$360** |

Buying every phase separately costs $435, so the whole track saves $75.

No payment is taken online. Registering creates an order and payment instructions are sent with it. Access opens once payment is confirmed.