Cloud Contraptions
TRAINING & CONSULTING

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

Guardrails for LLM Applications: Safety, Security, and Validation

Class Duration

14 hours of live training delivered over 2 days.

Student Prerequisites

- Professional software development experience
- Familiarity with LLM API usage and basic LLM application architecture

Target Audience

Software engineers and platform teams building production LLM applications and agents who need to constrain inputs and outputs, defend against prompt injection, and enforce policy. Relevant for teams shipping customer-facing assistants, agents with tool access, or LLM features that touch sensitive data. For the broader pipeline-security and leadership-policy companions, see Securing AI-Assisted Development Pipelines and Responsible AI for Engineering Leaders.

Description

A capable model is not a safe product. This two-day course teaches guardrails as a defense-in-depth discipline for LLM applications: validating and constraining inputs and outputs, moderating content, defending against prompt injection and jailbreaks, handling sensitive data, and keeping agents within scope. Participants map real risks to the OWASP Top 10 for LLM Applications and then build layered controls using the current tooling, including NeMo Guardrails with Colang, Guardrails AI validators, Llama Guard, Azure AI Content Safety, and gateway-level enforcement. The course pays particular attention to agents and tool use, where excessive agency turns a bad output into a harmful action, and closes with red-teaming, testing, and monitoring so guardrails keep working as the application evolves.

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

Learning Outcomes

- Explain the guardrails landscape and where each control belongs in a defense-in-depth architecture.
- Map application risks to the OWASP Top 10 for LLM Applications and select controls accordingly.
- Validate and constrain model outputs with schema, type, and content validators.
- Detect and mitigate prompt injection and jailbreaks across direct and indirect vectors.
- Apply content moderation and safety classification with tools such as Azure AI Content Safety and Llama Guard.
- Build conversational rails for topic and scope control with NeMo Guardrails and Colang.
- Handle sensitive data with PII detection, redaction, and secret-leakage prevention.
- Test, red-team, and monitor guardrails over the life of the application.

Training Materials

Comprehensive courseware is distributed online at the start of class. All students receive a downloadable MP4 recording of the training.

Software Requirements

Python 3.12+, API keys for at least one frontier model, accounts or access for guardrails tooling (NeMo Guardrails, Guardrails AI, and Azure AI Content Safety; free tiers are acceptable), and Git.

Training Topics

The Guardrails Landscape in 2026

- Why a capable model is not a safe product
- Defense in depth: input, output, and conversation rails
- The tooling landscape: NeMo Guardrails, Guardrails AI, Llama Guard, LLM Guard, Azure AI Content Safety
- Gateway-level versus application-level enforcement

OWASP LLM Top 10 and Threat Modeling

- The OWASP Top 10 for LLM Applications

- Prompt injection, sensitive information disclosure, and excessive agency
- Threat modeling an LLM feature
- Choosing controls by risk and impact

Input Validation and Sanitization

- Validating and normalizing user input
- Detecting malicious and adversarial prompts
- Allowlists, denylists, and intent checks
- Handling untrusted content from documents and tools

Output Validation and Structured Constraints

- Schema and type validation of model output
- Composable validators and quality constraints
- Re-asking and repair on validation failure
- Blocking unsafe or off-policy responses

Prompt Injection and Jailbreak Defense

- Direct and indirect prompt injection
- Jailbreak detection and mitigation
- Isolating system instructions from untrusted content
- Defending retrieval and tool inputs

Content Moderation and Safety Classification

- Content safety categories and severity
- Azure AI Content Safety and Llama Guard
- Toxicity, self-harm, and policy-violation detection
- Tuning thresholds to reduce false positives

Conversational Rails and Scope Control

- Topic and scope boundaries with NeMo Guardrails and Colang
- Keeping assistants on task
- Refusal and safe-completion patterns
- Fact-checking and hallucination rails

Sensitive Data and PII Handling

- PII detection and redaction
- Secret and credential leakage prevention
- Data minimization in prompts and logs

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

- Regulated data considerations

Guardrails for Agents and Tool Use

- Excessive agency and the blast radius of actions
- Approval gates and capability scoping
- Constraining tool calls and side effects
- Sandboxing and least-privilege execution

Testing, Red-Teaming, and Monitoring

- Red-teaming prompts and adversarial test suites
- Regression testing guardrails in CI
- Monitoring blocks, bypasses, and drift
- Incident response for guardrail failures

