

Generative AI in Production

Traditional MLOps is a set of practices to productionize traditional ML systems for enterprise applications. Generative AI raises new challenges in managing and productionizing applications at scale. The field of generative AI operations seeks to address these new challenges. In this course, you learn about the challenges that arise when deploying and productionizing generative AI-powered applications. You learn how to secure your generative AI-powered applications. Finally, you will discuss best practices for logging and monitoring your generative AI-powered applications in production.

**DURATION**

1 day

**LEVEL**

300 - Advanced

**FORMAT**

Instructor-led

What you'll learn

- Productionize LLM-powered applications
- Secure generative AI applications
- Implement logging and monitoring for LLM-powered applications

Overview	5 modules · 4 labs
Who this course is for	• Developers, DevOps engineers and machine learning engineers who wish to operationalize GenAI-based applications
Products	• Agent Platform • Agent Runtime on Gemini Enterprise Agent Platform • Model Armor • Gemini Enterprise Agent Platform Inference • Model Monitoring on Gemini Enterprise Agent Platform • Agent Studio on Gemini Enterprise Agent Platform
Prerequisites	• Completion of the 'Application Development with LLMs on Google Cloud' or equivalent knowledge.

Module 01

Introduction to Generative AI in Production

Topics	• Generative AI operations • Traditional MLOps vs. GenAIOps • Components of an LLM system • RAG/ReAct architecture
Objectives	• Understand generative AI operations • Compare traditional MLOps and GenAIOps • Analyze the components of an LLM system • Define and compare RAG and ReAct

Module 02

Generative AI Application Deployment

Topics	• Application deployment options • Deployment, packaging, and versioning
Objectives	• Evaluate application deployment options • Deploy, package, and version apps

Module 03

Productionizing Generative AI

Topics	• Maintenance and updates • Testing and evaluation • CI/CD pipelines for gen AI-powered apps
Objectives	• Maintain and update LLM models • Test and evaluate gen AI-powered apps • Deploy CI/CD pipelines for gen AI-powered apps
Activities	• Lab: Tracking Versions of Generative AI Applications

Module 04

Securing Generative AI Applications

Topics	• Security challenges • Prompt security • Sensitive Data Protection and DLP API • Model Armor
Objectives	• Identify security challenges for gen AI applications • Understand prompt security issues • Apply sensitive data protection and DLP API • Implement Model Armor

Module
05

Observability for Production LLM Systems

Topics

- Cloud Operations
- Cloud Logging
- Monitoring
- Cloud Trace
- Agent Analytics and AgentOps
- Putting it all together

Objectives

- Describe the purpose and capabilities of Google Cloud Observability
- Explain the purpose of Cloud Monitoring
- Explain the purpose of Cloud Logging
- Explain the purpose of Cloud Trace