Google Cloud

Generative AI in Production

In this course, you learn about the different challenges that arise when productionizing generative AI-powered applications versus traditional ML. You will learn how to manage experimentation and tuning of your LLMs, then you will discuss how to deploy, test, and maintain your LLM-powered applications. Finally, you will discuss best practices for logging and monitoring your LLM-powered applications in production.

DURATION
1 day

LEVEL
Advanced

FORMAT
Instructor-led

What you'll learn

- Describe the challenges in productionizing applications using generative AI.
- Manage experimentation and evaluation for LLM-powered applications.
- Productionize LLM-powered applications.
- Implement logging and monitoring for LLM-powered applications.

Overview

4 modules · 6 labs

Who this course is for

Developers and machine learning engineers who wish to operationalize Gen AI-based applications

Products

- Vertex AI
- Vertex AI Pipelines
- Vertex AI Evaluation
- Vertex AI Studio
- Vertex AI Gemini API
- Gemini

Prerequisite

Completion of "Introduction to Developer Efficiency on Google Cloud" or equivalent knowledge.

Module 01 Introduction to Generative AI in Production

Topics

- AI System Demo: Coffee on Wheels
- Traditional MLOps vs. GenAIOps
- Generative AI Operations
- Components of an LLM System

Objectives

- Understand generative AI operations
- Compare traditional MLOps and GenAIOps
- Analyze the components of an LLM system

Module 02 Managing Experimentation

Topics

- Datasets and Prompt Engineering
- RAG and ReACT Architecture
- LLM Model Evaluation (metrics and framework)
- Tracking Experiments

Objectives

- Experiment with datasets and prompt engineering.
- Utilize RAG and ReACT architecture.
- Evaluate LLM models.
- Track experiments.

Activities

- Lab: Unit Testing Generative AI Applications
- Optional Lab: Generative AI with Vertex AI: Prompt Design

Module 03 Productionizing Generative AI

Topics

- Deployment, packaging, and versioning (GenAIOps)
- Testing LLM systems (unit and integration)
- Maintenance and updates (operations)
- Prompt security and migration

Objectives

- Deploy, package, and version models
- Test LLM systems
- Maintain and update LLM models
- Manage prompt security and migration

Activities

- Lab: Vertex AI Pipelines: Qwik Start
- Lab: Safeguarding with Vertex AI Gemini API

Module 04 Logging and Monitoring for Production LLM Systems

Topics

- Cloud Logging
- Prompt versioning, evaluation, and generalization
- Monitoring for evaluation-serving skew
- Continuous validation

Objectives

- Utilize Cloud Logging
- Version, evaluate, and generalize prompts
- Monitor for evaluation-serving skew
- Utilize continuous validation

Activities

- Lab: Vertex AI: Gemini Evaluations Playbook
- Optional Lab: Supervised Fine Tuning with Gemini for Question and Answering