APPLIED AND GENERATIVE AI

Deploying and Scaling Generative AI Applications

Level: Advanced • 2 days (expandable to 3) • Virtual, In-person

Overview

A generative AI feature that works on your laptop is not the same as one that serves real users reliably and affordably. LLM applications bring their own production challenges: token costs that scale with use, latency that varies with every request, and outputs that are never quite deterministic. This course is about crossing that gap, deploying and scaling LLM applications that hold up under real load and real budgets.

This is a hands-on, advanced course. It starts where production always should, with the question of what "good" means: the axes of latency, cost, throughput, reliability, and safety, and how LLM apps differ from ordinary services on each. Only once we can set targets do we deploy, first to a managed platform and then, where it makes sense, by serving an open model ourselves. From there we work the problems that actually bite in production, in order: performance and cost, then scaling under load, then reliability and operations. Rather than tour every piece of infrastructure, we go deep on the concerns that determine whether an LLM app survives contact with users. Every module ends with hands-on work and builds on the one before.

Who Should Attend

- Engineers taking LLM applications from prototype to production
- Platform and DevOps engineers supporting generative AI workloads
- Technical leads responsible for the cost and reliability of AI features

Prerequisites

- Comfortable building LLM applications, for example from *Building Generative AI Applications* or equivalent experience
- Familiar with deploying and running a service in the cloud
- Comfortable with the command line and containers

What You Will Learn

- Set production targets for latency, cost, throughput, reliability, and safety
- Choose among hosted APIs, managed platforms, and self-hosted open models
- Deploy an LLM application to a managed platform
- Reduce latency and cost with streaming, caching, and right-sized models
- Scale under load with concurrency, rate limits, and autoscaling
- Make an LLM application reliable and observable in production

Course Outline

Day one: from app to deployed service

- What "Production" Means for LLM Applications
  - The axes that matter: latency, cost, throughput, reliability, and safety

  - How LLM apps differ: token cost, variable latency, and nondeterminism
  - Setting targets before you deploy
  - Lab: define production targets for a sample application
- Deployment Options
  - Hosted model APIs versus self-hosted open models
  - Managed platforms such as Azure AI Foundry, and where local deployment fits
  - Choosing an option against your constraints
  - Lab: deploy an LLM application to a managed platform
- Serving Open Models Yourself
  - When self-hosting is worth it
  - The moving parts: inference server, hardware, and model size
  - Lab: stand up a local or self-hosted open model behind an API

Day two: performance, scale, and reliability

- Performance and Cost
  - Latency: streaming, caching, and the size of the prompt and context
  - Cost: token accounting, right-sizing models, and caching
  - Lab: cut latency and cost on a working application, and measure the change
- Scaling Under Load
  - Concurrency, rate limits, and backpressure
  - Queueing and autoscaling
  - Lab: load-test your application and handle a traffic spike
- Reliability and Operations
  - Fallbacks, retries, timeouts, and graceful degradation
  - Observability for LLM apps: logging, tracing, and metrics
  - Runtime guardrails and ongoing monitoring (see *Evaluating and Monitoring Generative AI Applications*)
  - Lab: add fallbacks and observability, then break and recover the application

Extended Version

The three-day version keeps the same gradient and adds production depth:

- Multi-region and high-availability deployment
- Serving fine-tuned and multiple models behind one interface
- Cost governance and budgets across an organization
- A capstone that deploys, load-tests, hardens, and instruments an LLM application end to end