Course description

In this course, you will learn about the fundamental concepts, methods, and strategies for deploying, managing, and scaling generative AI models in real-world production environments. You will gain a solid understanding of the unique challenges and opportunities that arise when moving generative AI from development to operation. Additionally, you will learn about practical insights into MLOps principles, architectural patterns, monitoring, and responsible AI practices essential for successful production deployments. By the end of the course, you will explore how to plan and discuss the implementation of robust and secure generative AI solutions within your organization.

- Course level: Fundamental
- Duration: 1 day

Activities

This course includes presentations, hands-on labs, demonstrations, and group exercises.

Course objectives

In this course, you will learn to:

- Summarize the lifecycle and key challenges of deploying generative AI models in production

- Discuss appropriate architectural patterns and deployment strategies for generative AI applications
- Describe how to effectively monitor and evaluate generative AI model performance and quality in live environments
- Recognize the importance of MLOps principles and CI/CD for generative AI solutions
- Explain how to integrate responsible AI, security, and cost management practices into production deployments

Intended audience

This course is intended for those with limited prior knowledge of generative AI production:

- Machine Learning Engineers
- MLOps Engineers
- Data Scientists
- Solution Architects
- Technical Project Managers
- IT Operations Professionals

Prerequisites

We recommend that attendees of this course have:

- None

Course outline

Module 1: Introduction to Generative AI in Production

- Generative AI recap and its production relevance
- Understanding the Generative AI lifecycle: From concept to deployment
- Key challenges in production: Scalability, latency, cost, and quality
- Overview of MLOps for Generative AI
- Demo: A simple deployed generative AI application

Module 2: Foundational Models and Deployment Options

- Overview of different types of foundation models (LLMs, image, code)
- Exploring model sizes, capabilities, and trade-offs for production
- Deployment choices: Cloud APIs vs. self-hosting vs. managed services

- Introduction to cloud-based Generative AI services and model hubs (e.g., Amazon Bedrock, Hugging Face)
- Explore available models and their typical use cases

Module 3: Prompt Engineering for Production Readiness

- Review of prompt engineering fundamentals
- Designing robust and reliable prompts for consistent production outputs
- Strategies for prompt version control and management
- Advanced prompting techniques for production (e.g., chain-of-thought, self- consistency)
- Model settings and parameters for production optimization
- Hands-on Lab: Designing and refining production-ready prompts

Module 4: Architectural Considerations for GenAI Applications

- Common architectural patterns for Generative AI applications (e.g., RAG, agents)
- Integrating generative AI models into existing system architectures
- Data flow and processing pipelines for GenAI (e.g., vector databases, caching)
- Principles for building scalable and resilient GenAI backends
- Hands-on Lab: Building a simplified Retrieval Augmented Generation (RAG) component

Module 5: Monitoring, Evaluation, and Cost Management

- The necessity of monitoring in Generative AI production (model drift, data drift, prompt drift)
- Key performance metrics for Generative AI models (latency, throughput, quality scores, hallucination rates)
- Introduction to A/B testing and experimentation for prompt and model variations
- Understanding cost implications and basic optimization strategies for GenAI APIs and inference
- Hands-on Lab: Basic monitoring setup for a deployed Generative AI endpoint

Module 6: MLOps Principles for Generative AI

- Introduction to MLOps and its application to Generative AI
- Automating deployment, updates, and rollbacks for models and prompts
- Version control for models, data, prompts, and application code

- CI/CD pipelines for Generative AI applications
- Model registries and lifecycle management for production models
- Demo: A simplified Generative AI MLOps pipeline

Module 7: Responsible AI and Security in Production

- Review of responsible AI principles specific to Generative AI in production
- Implementing guardrails, content moderation, and output filtering
- Bias detection and mitigation strategies for Generative AI outputs
- Data privacy, compliance, and security considerations for GenAI applications
- Protecting against adverse prompts and input/output vulnerabilities
- Hands-on Lab: Implementing basic content filtering and guardrails

Module 8: Course Wrap-up: Next Steps and Advanced Topics

- Summary of key learnings and best practices for Generative AI in production
- Future trends and emerging technologies in production GenAI (e.g., multi-modal, personalization)
- Advanced topics: distributed inference, edge deployment, federated learning
- Resources for continuous learning and community engagement
- Capstone Exercise: Planning a Generative AI production rollout within an organization

cloudthat

Your course overview ends here.

But your cloud journey doesn't!

Unlock advanced training and certification pathways with **CloudThat.**

About CloudThat

Established in 2012, CloudThat is an award-winning company and the first in India to offer cloud training and consulting services for individuals and enterprises worldwide. Recently, it won Google Cloud's New Training Partner of the Year Award for 2025, becoming the first company in the world in 2025 to hold awards from all three major cloud giants: AWS, Microsoft, and Google. CloudThat notably won consecutive AWS Training Partner of the Year (APJ) awards in 2023 and 2024 and the Microsoft Training Services Partner of the Year Award in 2024, bringing its total award count to an impressive 12 awards in the last 8 years.

CloudThat specializes in Cloud Migration, Data Platforms, DevOps, IoT, and cutting-edge technologies like Gen AI & AI/ML. As a Microsoft Solutions Partner, AWS Premier Tier Training Partner, Google Cloud Platform Partner, and collaborator with leading organizations like HPE, Oracle, Nvidia and Databricks, CloudThat has trained over 850,000 professionals across 600+ cloud certifications, empowering students and professionals worldwide to advance their skills and careers.

To know more about our cloud certification training, email at sales@cloudthat.com or call us at +91-8880002200 or +1-8555588830 for those from outside India.