

**Comprehensive Generative AI Engineering for LLMOps**

**Course Number:** AI-128WA  
**Duration:** 5 days

**Overview**

This comprehensive Generative AI (GenAI) course is for DevOps and ITOps professionals who want to master the deployment, management, and scaling of Generative AI and Large Language Model (LLM) applications. The course covers various topics, from the foundations of LLMs to advanced deployment strategies and best practices. Participants will gain hands-on experience with popular tools and frameworks, including Docker, Kubernetes, and cloud platforms in an LLM environment.

**Prerequisites**

* Practical Python programming and scripting for automation tasks (6+ months)
  + API call access and event stream handling
  + Exception handling, debugging, testing, and logging
* Experience with containerization technologies (e.g., Docker) and orchestration platforms (e.g., Kubernetes)
* Familiarity with CI/CD pipelines and tools, such as Jenkins, GitLab, or GitHub Actions
* Knowledge of cloud platforms (e.g., AWS, GCP, Azure) and their services
* Experience with monitoring and logging tools, such as Prometheus, Grafana, and ELK stack (Elasticsearch, Logstash, Kibana) is recommended but not required
* Machine Learning concepts recommended - classification, regression, clustering

**Materials**

All Generative AI training students receive comprehensive courseware.

**Software Needed on Each Student PC**

All attendees must have a modern web browser and an Internet connection.

**Objectives**

* Gain a solid understanding of Large Language Models (LLMs) and their unique applications to streamline operations workflows
* Master prompt engineering techniques to maximize the effectiveness of LLMs in addressing IT and operational challenges, including log analysis, alert generation, and incident response
* Integrate LLMs seamlessly into existing monitoring, alerting, and automation tools for improved efficiency and decision-making
* Develop practical, real-world use cases that leverage LLMs to enhance operational processes and resolve production outages
* Confidently deploy and manage open-source LLMs, including Llama 3, in production environments for optimal performance and scalability
* Package LLM applications into containers using Docker for efficient deployment and portability
* Orchestrate LLM containers with Kubernetes for scalable, resilient, and automated management
* Implement robust strategies to scale LLM applications to handle varying workloads and ensure optimal performance
* Monitor and troubleshoot LLM applications effectively, tracking key performance metrics and addressing potential issues proactively
* Implement automated testing and robust security measures to ensure the reliability, safety, and compliance of LLM applications
* Optimize resource usage and costs in cloud-based LLM deployments, ensuring cost-effectiveness and sustainability

**Outline**

* Introduction
* LLM Fundamentals for Ops
  + Introduction to Generative AI and LLMs for Operations Workflows
  + LLM Architecture and Deployment Considerations
    - Implications of LLM architecture on deployment, scaling, and resource management
* Prompt Engineering for Ops
  + Introduction to Prompt Engineering
    - Techniques for creating effective prompts
    - Best practices for prompt design and optimization
  + Developing prompts for IT and traditional Ops tasks
    - Log analysis
    - Alert generation
    - Incident response
  + Improving response to production outages and IT challenges with PE
* LLM Integration for Ops
  + Overview of key LLM APIs and libraries
    - OpenAI API
    - HuggingFace Transformers
  + Strategies for integrating LLMs into monitoring, alerting, and automation tools
    - Use Case Development
    - Real-World Case Studies
  + Building an LLM-powered monitoring and alerting system
* Deployment and Management of Open-Source LLMs
  + Introduction to Open-Source LLMs
    - Advantages and limitations in production environments
  + Best practices for deploying and managing open-source LLMs
  + Techniques for managing LLM infrastructure, scaling, and performance
  + Setting up Lllama 3 from HuggingFace
* Containerization and Orchestration
  + Containerizing LLM applications using Docker
  + Orchestrating LLM containers using Kubernetes
  + Deploying an LLM application using Docker and Kubernetes
* Scaling LLM Applications
  + Strategies for horizontal and vertical scaling
  + Load balancing and auto-scaling techniques
  + Implementing auto-scaling for an LLM application
* Monitoring and Troubleshooting
  + Key performance metrics for LLM applications
  + Automated Testing for LLMOps
    - Differences of LLMOps testing and traditional software testing
    - Evaluation using CI/CD Tools
  + Evaluating LLM problems like hallucinations, data drift, unethical/harmful outputs
  + Monitoring tools and techniques (e.g., Weights and Biases, CircleCI)
    - Setting up monitoring for an LLM application
    - Creating dashboards and alerts for key metrics
* Security, Compliance, and Cost Optimization
  + Securing LLM application infrastructure and data
  + Ensuring compliance with relevant regulations and standards
  + Strategies for optimizing resource usage and costs in cloud-based LLM deployments
* Conclusion