

**Introduction to Generative AI Engineering for LLMOps**

**Course Number:** AI-124WA
**Duration:** 3 days

**Overview**

This Introduction to Generative AI (GenAI) course teaches DevOps and ITOps professionals to master the deployment, management, and scaling of GenAI and Large Language Model (LLM) applications.

**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 for successful operations workflows
* Master prompt engineering techniques to maximize the effectiveness of LLMs in addressing IT and operational challenges
* Integrate LLMs seamlessly into your existing monitoring, alerting, and automation tools for improved efficiency and decision-making
* Develop practical, real-world use cases that leverage LLMs to enhance log analysis, alert generation, incident response, and more
* Confidently deploy and manage open-source LLMs, including Llama 3, in production environments for optimal performance and scalability

**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
* Conclusion