

**Serving and Deploying Enterprise LLM Applications**

**Course Number:** AI-138WA  
**Duration:** 4 days

**Overview**

This advanced Large Language Model (LLM) training is for Ops professionals who want to master deploying, managing, and scaling sophisticated LLM-based applications in enterprise environments. The course covers advanced topics such as scalable model serving infrastructures, monitoring and troubleshooting techniques, Agentic RAG deployment, and CI/CD and DevOps practices for LLM-based applications.

**Prerequisites**

* Practical programming skills in Python and familiarity with LLM concepts and frameworks (3+ Months LLM, 6+ Months Python and Machine Learning)
  + LLM Access via API, Open Source Libraries (HuggingFace)
  + LLM Application development experience (RAG, Chatbots, etc)
* Strong understanding of containerization, orchestration, and cloud computing concepts
* Experience with monitoring, logging, and troubleshooting of production systems
* Familiarity with DevOps practices and CI/CD pipelines
  + MLOps knowledge preferred but not required

**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**

* Design and implement scalable model serving infrastructures for LLM-based applications, leveraging Kubernetes and serverless technologies for optimal performance and high availability
* Optimize model serving performance and cost-efficiency by implementing advanced techniques like caching, compression, and quantization and leveraging spot instances and reserved capacity
* Implement comprehensive monitoring and logging for LLM-based applications, setting up distributed tracing, metrics collection, and log aggregation. Utilize advanced dashboards, alerts, and automated troubleshooting for proactive issue resolution
* Deploy and manage agentic RAG architectures at scale in production environments, ensuring scalability, fault tolerance, and optimized performance through monitoring and resource utilization
* Streamline LLM-based application deployments with advanced CI/CD pipelines, integrating automated testing, staging, and production deployments while leveraging GitOps and infrastructure-as-code practices for efficient collaboration

**Outline**

* Advanced Model Serving Infrastructure and Scalability
  + Designing and implementing scalable model serving infrastructures for LLM-based applications
    - Leveraging Kubernetes and serverless technologies for auto-scaling and high availability
    - Implementing multi-region and multi-cloud deployment strategies for scale
  + Optimizing model serving performance and cost-efficiency
    - Implementing advanced caching, compression, and quantization techniques for model serving
    - Leveraging spot instances, reserved capacity, and other cost optimization strategies
  + Implementing a scalable and cost-efficient model serving infrastructure for an LLM-based application
* Monitoring, Logging, and Troubleshooting for LLM-Based Applications
  + Implementing advanced monitoring and logging techniques for LLM-based applications
    - Setting up distributed tracing, metrics collection, and log aggregation for LLM-based applications
    - Implementing advanced monitoring dashboards and alerts for key performance and quality metrics
  + Troubleshooting and root cause analysis for LLM-based application issues
    - Leveraging advanced debugging, profiling, and visualization tools for identifying performance bottlenecks and errors
    - Implementing automated anomaly detection and incident management workflows for LLM-based applications
  + Setting up comprehensive monitoring, logging, and troubleshooting for an LLM-based application
    - Configuring distributed tracing, metrics collection, and log aggregation
    - Implementing monitoring dashboards, alerts, and automated troubleshooting
* Deploying and Managing Agentic RAG Architectures at Scale
  + Deploying and managing Agentic RAG architectures in production environments
    - Designing and implementing scalable and fault-tolerant Agentic RAG deployment architectures
    - Leveraging containerization, orchestration, and serverless technologies for Agentic RAG deployment
  + Monitoring and optimizing Agentic RAG performance and resource utilization
    - Implementing advanced monitoring and profiling techniques for Agentic RAG components
    - Optimizing Agentic RAG deployments for cost-efficiency and performance at scale
  + Deploying and managing an Agentic RAG architecture in a production environment
* CI/CD and DevOps Practices for LLM-Based Application Deployments
  + Implementing advanced CI/CD pipelines and workflows for LLM-based application deployments
    - Designing and implementing end-to-end CI/CD pipelines with automated testing, staging, and production deployments
    - Leveraging GitOps and infrastructure-as-code practices for declarative and version-controlled deployments
  + Adopting DevOps best practices for collaborative and efficient LLM-based application development and deployment
    - Implementing agile development methodologies and continuous feedback loops for LLM-based applications
    - Establishing effective collaboration and communication channels between development, ops, and data science teams
  + Implementing a CI/CD pipeline and DevOps practices for an LLM-based application deployment
    - Designing and implementing an end-to-end CI/CD pipeline with automated testing and deployment stages
* Conclusion