

**Developing Advanced LLM Applications**

**Course Number:** AI-136WA  
**Duration:** 4 days

**Overview**

This advanced Generative AI training is designed for developers who want to explore enterprise-grade Large Language Model (LLM) architectures and design patterns. This course covers chatbot architectures, Agentic RAG, LLM-powered agents, and model serving and deployment techniques. Participants learn how to design and implement advanced LLM-based applications using cutting-edge technologies and frameworks.

**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)
* Familiarity with deep learning concepts and frameworks (e.g., TensorFlow, PyTorch)
* Experience with software development practices, system design, and enterprise application architecture recommended
* CI/CD Pipelines and monitoring for traditional ML models (MLOps) recommended

**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 advanced chatbot architectures with LLMs for personalized, context-aware interactions
* Integrate chatbots seamlessly into enterprise systems  for streamlined workflows
* Master agentic RAG architectures and techniques to build powerful systems capable of multi-hop reasoning and graph-based knowledge representation
* Build and orchestrate LLM-powered agents for autonomous decision-making and complex task completion in enterprise environments
* Deploy and manage LLM-based applications with advanced model serving techniques, ensuring scalability, cost-efficiency, and continuous improvement through CI/CD

**Outline**

* Deep Dive into Enterprise-Grade Chatbot Architectures
  + Designing and implementing advanced chatbot architectures using LLMs
    - Leveraging multi-turn conversation management and context tracking techniques
    - Implementing personalized and adaptive chatbot interactions based on user profiles
  + Integrating chatbots with enterprise systems and workflows
    - Strategies for integrating chatbots with CRM (customer relationship management), ERP (enterprise resource planning), and other enterprise applications
    - Implementing secure authentication and authorization mechanisms for chatbot interactions
  + Building an enterprise-grade chatbot using advanced LLM architectures
    - Designing and implementing a multi-turn, context-aware chatbot architecture
    - Integrating the chatbot with enterprise systems and implementing security measures
* Advanced Agentic RAG Architectures and Techniques
  + Exploring advanced Agentic RAG architectures and design patterns
    - Implementing multi-hop reasoning and iterative query refinement techniques in RAG
    - Leveraging graph-based knowledge representations and reasoning in Agentic RAG
  + Optimizing Agentic RAG performance and scalability
    - Implementing distributed retrieval and generation techniques for large-scale Agentic RAG
    - Leveraging caching, pruning, and other optimization techniques for efficient Agentic RAG inference
  + Implementing an advanced Agentic RAG architecture for a specific use case
    - Designing and implementing a multi-hop Agentic RAG architecture with graph-based reasoning
    - Optimizing the Agentic RAG implementation for performance and scalability
* Designing and Implementing LLM-Powered Agents and Workflows
  + Designing LLM-powered agents for autonomous decision-making and task completion
    - Implementing goal-oriented and adaptive agent architectures using LLMs
    - Leveraging reinforcement learning and planning techniques for agent decision-making
  + Orchestrating multi-agent workflows and interactions in enterprise environments
    - Designing and implementing multi-agent communication and coordination protocols
    - Implementing fault-tolerant and scalable multi-agent workflows using serverless architectures
  + Building an LLM-powered agent-based workflow for a specific enterprise use case
    - Designing and implementing a goal-oriented, adaptive agent architecture using LLMs
    - Orchestrating a multi-agent workflow using serverless technologies and coordination protocols
* Advanced Model Serving and Deployment Techniques
  + Exploring advanced model serving architectures and design patterns
    - Implementing model versioning, A/B testing
    - Leveraging serverless and edge computing for low-latency and cost-efficient model serving
  + Implementing CI/CD pipelines for automated model deployment and monitoring
    - Designing and implementing end-to-end CI/CD pipelines for LLM-based applications
    - Integrating model performance monitoring and drift detection into CI/CD workflows
  + Implementing an advanced model serving architecture with CI/CD for an LLM-based application
    - Designing and implementing a serverless model serving architecture with versioning and A/B testing
  + Setting up a CI/CD pipeline for automated model deployment and monitoring
* Conclusion