

**Designing and Implementing Enterprise-Grade ML Applications**

**Course Number:** AI-140WA  
**Duration:** 4 days

**Overview**

This advanced Machine Learning (ML) course is designed for Data Science and ML professionals who want to master designing and implementing enterprise-grade ML applications. Attendees learn how to evaluate advanced LLM architectures and dive into advanced topics, such as fine-tuning and quantization techniques, LLM-powered recommender systems, model evaluation, and debugging, as well as ethical considerations and responsible AI practices for enterprise-grade LLMs.

**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 (OpenAI), Open Source Libraries (HuggingFace)
  + LLM Application development experience (RAG, Chatbots, etc)
* Strong practical understanding of ML concepts, algorithms, and evaluation
  + Supervised Learning, Unsupervised Learning, and respective algorithms
* Statistics, Probability, and Linear Algebra (vectors) foundations
* Experience with at least one deep learning framework (e.g., TensorFlow, PyTorch)

**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**

* Produce high-performing, domain-specific LLMs through advanced fine-tuning techniques
* Deploy efficient LLM models in resource-constrained environments through effective model compression
* Develop LLM-powered recommender systems that deliver personalized, context-aware user experiences
* Quantify LLM-based application performance, identifying areas for improvement and optimization
* Diagnose and enhance LLM models through in-depth interpretation and robust debugging techniques
* Build fair and unbiased LLM-based applications through advanced bias mitigation strategies
* Ensure transparency, accountability, and explainability in LLM-based applications, adhering to responsible AI principles

**Outline**

* Advanced Fine-Tuning and Quantization Techniques for LLMs
  + Exploring advanced fine-tuning techniques and architectures for domain-specific LLM adaptation
    - Implementing multi-task, meta-learning, and transfer learning techniques for LLM fine-tuning
    - Leveraging domain-specific pre-training and intermediate fine-tuning for improved LLM performance
  + Quantization and compression techniques for efficient LLM fine-tuning and deployment
    - Implementing post-training quantization and pruning techniques for LLM model compression
    - Exploring quantization-aware training and other techniques for efficient LLM fine-tuning
  + Implementing advanced fine-tuning and quantization techniques for a domain-specific LLM
    - Designing and implementing a multi-task fine-tuning architecture with domain-specific pre-training
    - Applying quantization and pruning techniques for fine-tuning
* Designing and Implementing LLM-Powered Recommender Systems
  + Exploring advanced architectures and techniques for LLM-powered recommender systems
    - Leveraging LLMs for multi-modal and context-aware recommendation generation
    - Implementing hybrid recommender architectures combining LLMs with collaborative and content-based filtering
  + Evaluating and optimizing LLM-powered recommender system performance
    - Designing and conducting offline and online evaluation studies for LLM-powered recommender systems
    - Implementing advanced evaluation metrics and techniques for assessing recommendation quality and diversity
  + Hands-on: Building an LLM-powered recommender system for a specific domain
* Advanced Model Evaluation, Interpretation, and Debugging Techniques
  + Implementing advanced evaluation and benchmarking techniques for LLM-based applications
    - Designing and conducting comprehensive evaluation studies with domain-specific metrics and datasets
    - Leveraging advanced evaluation frameworks and platforms for automated and reproducible evaluation
  + Model interpretation and debugging techniques for understanding LLM behavior and failures
    - Implementing advanced model interpretation techniques, such as attention visualization and probing
    - Leveraging debugging techniques, such as counterfactual analysis and influence functions, for identifying and mitigating LLM failures
  + Conducting an advanced evaluation and debugging study for an LLM-based application
    - Designing and implementing a comprehensive evaluation study with domain-specific metrics and datasets
    - Applying model interpretation and debugging techniques for LLMs
* Ethical Considerations and Responsible AI Practices for Enterprise-Grade LLMs
  + Implementing advanced techniques for mitigating biases and ensuring fairness in LLM-based applications
    - Leveraging advanced bias detection and mitigation techniques, such as adversarial debiasing and fairness constraints
    - Designing and conducting fairness audits and assessments for LLM-based applications
  + Ensuring transparency, accountability, and explainability in LLM-based decision-making
    - Implementing advanced explainability techniques, such as counterfactual explanations and feature importance
    - Designing and implementing governance frameworks and processes for responsible LLM deployment and monitoring
  + Conducting an ethical assessment and implementing responsible AI practices for an LLM-based application