LLMOps: Systems, Infrastructure, and Operations for Large Language Models

Course ID: PA6400

Pextra Academy™

Professional Training in Cloud Computing & Infrastructure Engineering

Program Overview

This advanced program explores the emerging discipline of LLMOps — the operational practices, systems engineering, and infrastructure required to deploy and manage Large Language Models (LLMs) at scale.

At the intersection of machine learning, distributed systems, and cloud infrastructure, the course covers the full LLM lifecycle: foundational architectures, fine-tuning, deployment, monitoring, optimization, and responsible operations in production environments.

Students gain hands-on experience with public cloud platforms (AWS) and the enterprise-grade private cloud environment (Pextra CloudEnvironment®), while working on real-world scenarios including Retrieval-Augmented Generation (RAG), PromptOps, safety guardrails, CI/CD pipelines, and observability.

The program culminates in a mentored capstone group project managed through weekly agile scrums.

Learning Outcomes

Upon successful completion of this program, students will be able to:

- ✔ Design and manage the full lifecycle of large language models, including architecture, fine-tuning, deployment, monitoring, and maintenance
- ✔ Differentiate LLMOps from traditional MLOps and apply specialized practices in public, private, and hybrid cloud environments
- ✔ Implement continuous integration and deployment (CI/CD) pipelines for reliable LLM-based applications

- ✔ Leverage AWS and private cloud services to build, deploy, and maintain scalable LLM solutions
- ✔ Integrate retrieval-augmented generation (RAG) techniques and vector databases to enhance contextual performance
- ✔ Apply prompt engineering (PromptOps) and implement safety guardrails for ethical and secure LLM usage
- ✔ Monitor LLM performance, detect drift or degradation, and troubleshoot production issues using observability tools
- ✔ Conduct agile project management through scrums and deliver production-ready LLM systems

---

Technology Stack

- AWS Academy (Machine Learning Foundations, Machine Learning for Natural Language Processing, Learner Labs)
- Pextra CloudEnvironment® (Enterprise Private Cloud with GPU support where available)
- GitHub for version control and CI/CD workflows
- AWS SageMaker, Lex, and related ML/NLP services
- Tools for RAG, vector databases, prompt engineering, and observability

---

Curriculum Highlights

The program spans approximately **52 instructional hours** (lecture + guided labs + project work).

Module	Description	Hours
Foundations of AI & LLMs	Core concepts, architectures, and differences from traditional models	6
LLMOps Fundamentals	LLM lifecycle, LLMOps vs MLOps, operational challenges	6
PromptOps & Interaction Management	Prompt engineering, safety, ethics, and guardrails	6
Deployment & Orchestration	Containerization, platforms, scaling, and inference optimization	6
Retrieval-Augmented Generation (RAG)	Vector databases, contextual enhancement, and enterprise integration	6
Monitoring & Observability	Performance monitoring, drift detection, and troubleshooting	6
Security, Robustness & Evaluation	Security practices, bias mitigation, and production evaluation	6
Emerging Trends & Capstone Project	Advanced tooling, case studies, and final project development	10
Total		**52**

Note: Hours include lecture, discussion, guided labs, and project scrums. Additional self-paced work is expected.

Textbook & Learning Resources

Required/Recommended Textbook:

LLMOps: Foundations, Deployment, and Responsible Operations of Large Language Models
Raja Alomari, PhD & Ryan Alomari
Pextra Academy™
ISBN: 979-8999628015

Available on Amazon:
`https://www.amazon.com/LLMOps-Foundations-Deployment-Responsible-Operations/dp/B0FLYWJPDL`

This textbook serves as the primary reference, with all course slides aligned to its content. A physical copy will be made available in the library where applicable.

Assessment Model

Assessment Component	Weight
Hands-on Labs (AWS) or Assignments	25%
Quizzes	10%
Midterm Examination	15%
Final Examination	25%
Capstone Project (including scrums & demo)	25%

All assignments must be presented in class. An optional deeper project component is available for additional portfolio credit.

Suggested Prerequisites

This program is designed for motivated learners with a technical background. Recommended preparation includes:

- Familiarity with machine learning or AI concepts
- Basic understanding of cloud computing and distributed systems
- Experience with Python programming
- Interest in large-scale AI infrastructure and operations

No strict prerequisites are enforced, but prior exposure to LLMs or MLOps is highly beneficial.

Program Experience

Students benefit from:

- Extensive hands-on labs using AWS Academy and real LLM-related services
- Access to both public (AWS) and private (Pextra CloudEnvironment®) cloud environments
- Weekly agile scrums and a mentored group capstone project

- Practical exercises in RAG, PromptOps, guardrails, and production monitoring
- Direct exposure to industry-standard tools and responsible AI practices

---

Instructor Note:

For access to instructional materials, please reach out using the form at http://pextra.academy/contact.