Training Course on Large Language Models Deployment and Scaling Strategies

Professional Development Training Course Outline

Category	Data Science	**Duration**	5 Days
Base Fee	$1,500 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

The rapid evolution of Large Language Models (LLMs) has ushered in a new era of AI capabilities, transforming industries and unlocking unprecedented innovation. However, transitioning these powerful AI prototypes from experimental stages to robust, scalable, and secure production environments presents a significant challenge for organizations worldwide. This training course delves deep into the critical methodologies and advanced strategies required to successfully **productionize LLMs**, ensuring their seamless integration, optimal performance, and sustained value within complex enterprise architectures. Participants will gain actionable insights and hands-on expertise in addressing the unique complexities of **LLM lifecycle management**, from infrastructure considerations and model optimization to continuous monitoring and governance, empowering them to drive tangible business outcomes with AI.

Training Course on LLM Deployment & Scaling Strategies: Productionizing Large Language Models is designed to equip machine learning engineers, data scientists, and AI architects with the essential skills to navigate the intricate landscape of **enterprise LLM deployment**. We will explore cutting-edge techniques for **efficient inference**, **cost optimization**, and building resilient **LLM pipelines**, moving beyond theoretical understanding to practical application. Through real-world case studies and interactive exercises, attendees will master the art of **scaling AI infrastructure** for LLMs, implementing robust **MLOps practices** for AI, and ensuring **responsible AI deployment**, ultimately accelerating their organization's journey towards pervasive and impactful AI adoption.

Programme Curriculum

Training Course on LLM Deployment & Scaling Strategies: Productionizing Large Language Models

Introduction

The rapid evolution of Large Language Models (LLMs) has ushered in a new era of AI capabilities, transforming industries and unlocking unprecedented innovation. However, transitioning these powerful AI prototypes from experimental stages to robust, scalable, and secure production environments presents a significant challenge for organizations worldwide. This training course delves deep into the critical methodologies and advanced strategies required to successfully **productionize LLMs**, ensuring their seamless integration, optimal performance, and sustained value within complex enterprise architectures. Participants will gain actionable insights and hands-on expertise in addressing the unique complexities of **LLM lifecycle management**, from infrastructure considerations and model optimization to continuous monitoring and governance, empowering them to drive tangible business outcomes with AI.

Training Course on LLM Deployment & Scaling Strategies: Productionizing Large Language Models is designed to equip machine learning engineers, data scientists, and AI architects with the essential skills to navigate the intricate landscape of **enterprise LLM deployment**. We will explore cutting-edge techniques for **efficient inference**, **cost optimization**, and building resilient **LLM pipelines**, moving beyond theoretical understanding to practical application. Through real-world case studies and interactive exercises, attendees will master the art of **scaling AI infrastructure** for LLMs, implementing robust **MLOps practices** for AI, and ensuring **responsible AI deployment**, ultimately accelerating their organization's journey towards pervasive and impactful AI adoption.

Course Duration

5 days

Course Objectives

Upon completion of this training, participants will be able to:

1. Architect and design scalable LLM systems for production environments.
2. Implement efficient LLM inference techniques, including model quantization and distillation.
3. Apply cost optimization strategies for cloud-based LLM deployment.
4. Develop robust MLOps pipelines for continuous LLM integration and delivery.
5. Master prompt engineering and Retrieval-Augmented Generation (RAG) for enhanced LLM performance.
6. Understand and mitigate LLM security risks and ensure data privacy.
7. Deploy and manage LLMs on various cloud platforms (AWS, Azure, GCP) and on-premise infrastructure.
8. Implement real-time LLM monitoring and performance tracking.
9. Optimize GPU utilization and resource allocation for LLM workloads.
10. Evaluate and select appropriate open-source vs. proprietary LLM solutions.
11. Design and build agentic AI systems leveraging LLMs.
12. Ensure responsible AI practices and address ethical considerations in LLM deployment.
13. Troubleshoot common challenges in large-scale LLM operations.

Organizational Benefits

- Rapidly deploy and integrate LLMs into business processes, driving faster innovation.
- Optimize resource utilization and leverage cost-effective deployment strategies, leading to significant savings.
- Improve model accuracy, latency, and throughput for superior user experiences and business outcomes.

- Implement robust security measures and responsible AI practices, reducing compliance and ethical risks.
- Establish agile MLOps workflows for continuous LLM improvement and rapid iteration.
- Leverage cutting-edge LLM capabilities to develop differentiated products and services.
- Cultivate internal expertise in a critical and rapidly evolving AI domain.
- Enable data-driven insights and automation across the organization.

Target Audience

- Machine Learning Engineers
- Data Scientists
- AI Architects
- Software Developers with an interest in AI/ML
- DevOps Engineers specializing in AI/ML infrastructure
- CTOs and Technical Leads overseeing AI initiatives
- Product Managers working on AI-powered products
- Researchers and Academics in AI/NLP

Course Outline

Module 1: Foundations of LLM Productionization

- Understanding the LLM Lifecycle: From Research to Production
- Key Challenges in Productionizing LLMs: Scale, Cost, Latency, Data
- Overview of LLM Architectures and their Deployment Implications
- Introduction to MLOps Principles for Large Language Models
- Case Study: Scaling a Customer Service Chatbot from Prototype to Production

Module 2: LLM Optimization Techniques for Efficient Inference

- Model Quantization: Reducing Precision for Faster Inference
- Model Distillation: Creating Smaller, Faster Models from Larger Ones
- Parameter-Efficient Fine-Tuning (PEFT): Adapting LLMs with Minimal Resources
- Batching and Parallelism Strategies for Throughput Optimization
- Case Study: Optimizing a 70B Parameter Model for Edge Deployment

Module 3: Infrastructure and Cloud Deployment Strategies

- Choosing the Right Hardware: GPUs, TPUs, and Specialized Accelerators
- Cloud Deployment Options: AWS SageMaker, Azure ML, Google Cloud AI Platform
- Containerization with Docker and Orchestration with Kubernetes for LLMs
- Serverless Deployment and Managed Services for Cost-Effectiveness
- Case Study: Deploying a Multi-Tenant LLM Application on Kubernetes

Module 4: Advanced LLM Serving and API Management

- Building Robust LLM APIs with FastAPI/Flask and gRPC
- Load Balancing and Auto-Scaling for High Availability
- Caching Strategies for LLM Responses and Embeddings
- API Security, Authentication, and Authorization for LLM Endpoints
- Case Study: Implementing a High-Throughput LLM Inference Service with API Gateway

Module 5: Fine-tuning and Customizing LLMs

- Data Preparation and Curation for Domain-Specific Fine-tuning
- Transfer Learning and Adapting Pre-trained LLMs to New Tasks
- Instruction Tuning and Reinforcement Learning from Human Feedback (RLHF)
- Evaluating Fine-tuned Models: Metrics and Best Practices
- Case Study: Fine-tuning an LLM for Legal Document Summarization

Module 6: Retrieval-Augmented Generation (RAG) and Knowledge Integration

- Understanding RAG Architecture: Vector Databases, Embeddings, and Retrieval
- Building RAG Pipelines for Grounded and Up-to-Date Responses
- Integrating External Knowledge Sources: Databases, APIs, and Documents
- Challenges and Best Practices in RAG Implementation
- Case Study: Developing an Enterprise Knowledge Management System with RAG

Module 7: LLM Monitoring, Observability, and A/B Testing

- Key Metrics for LLM Performance Monitoring: Latency, Throughput, Quality
- Setting up Logging, Tracing, and Alerting for LLM Systems
- Detecting Model Drift and Anomalies in Production
- A/B Testing and Canary Deployments for LLM Updates
- Case Study: Establishing a Comprehensive Monitoring Dashboard for an LLM Application

Module 8: Responsible AI, Security, and Governance

- Ethical Considerations in LLM Deployment: Bias, Fairness, and Transparency
- Data Privacy and Compliance (GDPR, HIPAA) for LLM Applications
- Mitigating Hallucinations and Adversarial Attacks
- Establishing AI Governance Frameworks and Best Practices
- Case Study: Implementing Guardrails for a Public-Facing LLM Chatbot

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

e. One-year post-training support Consultation and Coaching provided after the course.

f. Payment should be done at least a week before commence of the training, to Fineskill Training Center account, as indicated in the invoice so as to enable us prepare better for you.

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	27 Nov 2026	Online	$1,000
23 Nov 2026	27 Nov 2026	Physical	$1,500
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$6,000
23 Nov 2026	27 Nov 2026	Physical	$4,500
23 Nov 2026	27 Nov 2026	Physical	$6,000
23 Nov 2026	27 Nov 2026	Physical	$6,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com