Training Course on Open-Source Large Language Modelss

Professional Development Training Course Outline

Category	Data Science	**Duration**	10 Days
Base Fee	$3,000 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

The landscape of Artificial Intelligence is being rapidly reshaped by Large Language Models (LLMs), moving from proprietary black-box solutions to a vibrant ecosystem of open-source alternatives. This paradigm shift empowers organizations with unparalleled control, customization, and cost-efficiency, fostering innovation and reducing vendor lock-in. Mastering the deployment and fine-tuning of open-source LLMs like Meta's Llama and Mistral is no longer an optional skill but a critical competency for businesses aiming to leverage cutting-edge AI ethically and effectively within their own infrastructure.

Training Course on Open-Source LLMs: Deployment & Customization – Working with Llama, Mistral, and Derivatives delves deep into the practical aspects of working with state-of-the-art open-source LLMs. Participants will gain hands-on experience with popular models and their derivatives, learning essential techniques for local deployment, efficient fine-tuning, performance optimization, and secure integration into existing systems. By the end of this program, professionals will be equipped to build, customize, and manage powerful, domain-specific AI applications, unlocking transformative opportunities across various industries.

Programme Curriculum

Training Course on Open-Source LLMs: Deployment & Customization - Working with Llama, Mistral, and Derivatives

Introduction

The landscape of Artificial Intelligence is being rapidly reshaped by Large Language Models (LLMs), moving from proprietary black-box solutions to a vibrant ecosystem of open-source alternatives. This paradigm shift empowers organizations with unparalleled control, customization, and cost-efficiency, fostering innovation and reducing vendor lock-in. Mastering the deployment and fine-tuning of open-source LLMs like Meta's Llama and Mistral is no longer an optional skill

but a critical competency for businesses aiming to leverage cutting-edge AI ethically and effectively within their own infrastructure.

Training Course on Open-Source LLMs: Deployment & Customization – Working with Llama, Mistral, and Derivatives delves deep into the practical aspects of working with state-of-the-art open-source LLMs. Participants will gain hands-on experience with popular models and their derivatives, learning essential techniques for local deployment, efficient fine-tuning, performance optimization, and secure integration into existing systems. By the end of this program, professionals will be equipped to build, customize, and manage powerful, domain-specific AI applications, unlocking transformative opportunities across various industries.

Course Duration

10 days

Course Objectives

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

1. Understand the architecture and core principles of Transformer-based LLMs, including Llama and Mistral.
2. Evaluate the advantages and disadvantages of open-source LLMs versus proprietary solutions for enterprise AI.
3. Set up and manage diverse development environments for LLM experimentation and deployment.
4. Perform efficient data preparation and tokenization for LLM training and fine-tuning datasets.
5. Implement advanced fine-tuning techniques like LoRA and QLoRA for domain-specific customization.
6. Optimize LLM inference for performance, latency, and cost-efficiency using quantization and pruning.
7. Deploy open-source LLMs on various platforms, including on-premises infrastructure and cloud environments.
8. Integrate LLMs with existing applications and workflows using APIs and frameworks like LangChain and LlamaIndex.
9. Monitor LLM performance, resource utilization, and drift detection in production environments.
10. Address ethical considerations and responsible AI practices in LLM development and deployment.
11. Troubleshoot common issues encountered during LLM deployment and customization.
12. Leverage community resources and best practices for ongoing LLM innovation and support.
13. Develop practical, real-world LLM applications through hands-on case studies and project-based learning.

Organizational Benefits

- Significant reduction in licensing and API usage fees compared to proprietary models.
- Enhanced control over sensitive data by deploying models on-premises, ensuring compliance with internal and regulatory standards.
- Ability to fine-tune models on proprietary data for specialized use cases, gaining a competitive edge.
- Freedom from reliance on a single vendor's roadmap, allowing for agility and adoption of new advancements.

- Full visibility into model architecture and behavior, crucial for regulated industries and explainable AI.
- Access to a vibrant open-source community, fostering continuous improvement, new tools, and faster development cycles.
- Upskilling internal teams to build and manage cutting-edge AI solutions, fostering a culture of innovation.

Target Audience

1. Machine Learning Engineers and Data Scientists
2. AI/ML Developers
3. DevOps Engineers
4. Solutions Architects
5. Researchers and Academics
6. Technical Leads and Team Managers
7. Software Engineers
8. IT Professionals

Course Outline

Module 1: Introduction to Large Language Models and Open-Source Ecosystem

- Understanding the evolution and impact of LLMs in AI.
- Core concepts: Attention mechanism, Transformers, pre-training, and fine-tuning.
- Overview of the open-source LLM landscape: Llama, Mistral, Gemma, Falcon, and more.
- Distinguishing open-source from proprietary LLMs: advantages and trade-offs.
- **Case Study:** Analyzing the open-source community's impact on LLM development (e.g., Hugging Face contributions, Model Cards).

Module 2: Setting up Your LLM Development Environment

- Hardware requirements and considerations for LLM deployment (GPUs, RAM).
- Software setup: Python, PyTorch/TensorFlow, Hugging Face Transformers.
- Leveraging virtual environments and containerization (Docker) for reproducibility.
- Introduction to cloud platforms for LLM development (AWS SageMaker, Google Cloud AI Platform).
- **Case Study:** Configuring an optimized local environment for Llama 3 inference on a consumer GPU.

Module 3: Deep Dive into Llama Architecture and Usage

- Understanding the Llama family of models (Llama 2, Llama 3): variants and capabilities.
- Exploring Llama's tokenization process and vocabulary.
- Loading and interacting with Llama models using the Hugging Face library.
- Basic prompt engineering techniques for Llama models.
- **Case Study:** Generating creative content or code snippets using a pre-trained Llama 2 model.

Module 4: Deep Dive into Mistral Architecture and Usage

- Understanding the Mistral family of models (Mistral 7B, Mixtral 8x7B): efficiency and performance.
- Exploring Mistral's optimized architecture for faster inference.
- Loading and interacting with Mistral models in different frameworks.
- Prompt engineering strategies specifically tailored for Mistral's strengths.

- **Case Study:** Building a rapid response chatbot using Mistral 7B for customer support.

Module 5: Data Preparation and Preprocessing for LLMs

- Sourcing and curating high-quality datasets for fine-tuning.
- Text cleaning, normalization, and handling special characters.
- Advanced tokenization techniques (Byte Pair Encoding, SentencePiece).
- Dataset formatting and preparation for popular LLM frameworks.
- **Case Study:** Preprocessing a dataset of legal documents for a legal domain-specific LLM.

Module 6: Fundamentals of LLM Fine-tuning

- Why fine-tune? Adapting pre-trained models to specific tasks or domains.
- Types of fine-tuning: Full fine-tuning, parameter-efficient fine-tuning (PEFT).
- Choosing the right base model for your fine-tuning task.
- Understanding loss functions, optimizers, and training schedules for LLMs.
- **Case Study:** Fine-tuning a Llama model for sentiment analysis on social media data.

Module 7: Parameter-Efficient Fine-tuning (PEFT) with LoRA and QLoRA

- Introduction to PEFT: Reducing computational resources and training time.
- Detailed explanation of LoRA (Low-Rank Adaptation) and its implementation.
- Understanding QLoRA for quantizing and fine-tuning large models with limited memory.
- Practical application of LoRA/QLoRA on Llama and Mistral derivatives.
- **Case Study:** Adapting a Llama 3 model to summarize internal company reports using QLoRA on a single GPU.

Module 8: LLM Performance Optimization Techniques

- Model quantization: Reducing model size and accelerating inference (e.g., INT8, FP16).
- Model pruning and distillation for efficiency.
- Batching and parallelization strategies for faster inference.
- Hardware acceleration: Leveraging GPUs and specialized AI chips.
- **Case Study:** Optimizing a deployed Mistral model for real-time inference on edge devices for a voice assistant application.

Module 9: On-Premises LLM Deployment Strategies

- Containerizing LLMs with Docker for consistent deployment.
- Orchestrating LLM deployments with Kubernetes and Helm.
- Setting up inference servers (e.g., vLLM, Text Generation Inference).
- Managing model versions and rollbacks.
- **Case Study:** Deploying a custom Llama 2 model for an internal knowledge base system on a private cloud environment using Kubernetes.

Module 10: Cloud-Native LLM Deployment

- Deploying open-source LLMs on major cloud platforms (AWS, Azure, GCP).
- Utilizing managed services for LLM hosting and inference.
- Scalability considerations for LLM applications in the cloud.
- Cost management and resource allocation for cloud LLM deployments.
- **Case Study:** Building a scalable Llama-powered content generation service on AWS Lambda or Azure Functions.

Module 11: Integrating LLMs with Applications (LangChain & LlamaIndex)

- Introduction to LLM orchestration frameworks: LangChain and LlamaIndex.
- Building intelligent agents with LangChain for complex tasks.
- Retrieval-Augmented Generation (RAG) with LlamaIndex for factual accuracy.
- Connecting LLMs to external tools and APIs.
- **Case Study:** Developing a question-answering system for an e-commerce website by integrating a Mistral model with a product database using RAG.

Module 12: Monitoring, Logging, and MLOps for LLMs

- Key metrics for monitoring LLM performance (latency, throughput, error rates).
- Implementing logging and tracing for LLM applications.
- Detecting model drift and performance degradation.
- Setting up continuous integration/continuous delivery (CI/CD) pipelines for LLMs.
- **Case Study:** Establishing a monitoring dashboard for a Llama-based legal assistant to track accuracy and user feedback.

Module 13: Responsible AI and Ethical Considerations in LLMs

- Understanding bias in LLMs and mitigation strategies.
- Ensuring fairness, transparency, and accountability in AI applications.
- Data privacy and security best practices for sensitive information processed by LLMs.
- Legal and ethical implications of generative AI.
- **Case Study:** Analyzing potential biases in a Llama-generated medical diagnosis tool and proposing mitigation strategies.

Module 14: Advanced Topics and Future Trends in Open-Source LLMs

- Multimodal LLMs: Integrating text, image, and other data types.
- Exploring emerging open-source models and research advancements.
- Edge deployment of LLMs for low-resource environments.
- Federated learning for LLMs.
- **Case Study:** Discussing the future of open-source LLMs in personalized learning platforms.

Module 15: Capstone Project: End-to-End Open-Source LLM Solution

- Participants work in teams to design, deploy, and customize an open-source LLM solution.
- Problem identification and solution design.
- Data collection, fine-tuning, and model optimization.
- Deployment, integration, and performance evaluation.
- Presentation of projects and peer review.
- **Case Study:** Developing and deploying a customized Llama 3 model for internal HR query resolution, including fine-tuning on company policies and integrating with an internal portal.

Training Methodology

This training course employs a highly interactive and practical methodology, blending theoretical foundations with extensive hands-on exercises and real-world case studies.

- **Instructor-Led Sessions:** Expert-led discussions, lectures, and interactive Q&A.
- **Hands-on Labs:** Practical coding exercises and guided implementations using cloud environments and local setups.
- **Case Studies:** In-depth analysis and practical application of LLMs to solve real-world business problems.

- **Group Discussions:** Collaborative problem-solving and knowledge sharing among participants.
- **Project-Based Learning:** A culminating project to apply learned skills to a comprehensive LLM deployment scenario.
- **Live Demos:** Demonstrations of LLM capabilities, deployment strategies, and optimization techniques.
- **Resource Sharing:** Access to curated code repositories, documentation, and relevant research papers.

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.

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
21 Sep 2026	02 Oct 2026	Online	$2,000
21 Sep 2026	02 Oct 2026	Physical	$3,000
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0

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

Register Online Now

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