KOENIG
step forward

KOENIG SOLUTIONS · COURSE OUTLINE

LLMs: LoRA Fine-Tuning & Llama.cpp Deployment

Learn to fine-tune LLMs with LoRA and QLoRA, prepare sensitive data responsibly, and deploy models using Llama.cpp and APIs.

40 Hours	5	20	4
DURATION	MODULES	TOPICS	HANDS-ON LABS

COURSE OVERVIEW

Unlock the power of Large Language Models (LLMs) in this comprehensive course designed for AI enthusiasts and professionals. From foundational knowledge to advanced deployment techniques, this course offers a step-by-step guide through the intricacies of LLMs, ethical considerations, and real-world applications.

WHO THIS COURSE IS FOR

- **Data Scientists and Machine Learning Engineers**
- **AI Researchers and Academics**
- **Software Developers**
- **IT Professionals**
- **Business Analysts and Decision Makers**
- **Students and Enthusiasts**

TOOLS AND REQUIREMENTS

- All demonstrations and labs are performed on the Free Tier of Google Colab (T4/P100 GPU). For larger models (7B+ parameters) or longer training, Colab Pro or local hardware with 24GB+ VRAM is recommended.

LEARNING OBJECTIVES

- Explain the anatomy of pre-trained LLMs and tokenization
- Compare fine-tuning methods including Full, LoRA and QLoRA
- Apply best practices for fine-tuning on sensitive data
- Prepare and anonymise data for fine-tuning tasks
- Describe responsible AI and ethical considerations
- Deploy models locally using Llama.cpp and GGUF formats
- Configure API deployment with FastAPI, Gradio or Streamlit
- Plan scaling and security for real-world LLM applications

COURSE SYLLABUS

5 Modules, 20 Topics

01

Lab

Introduction to LLMs and Setup

TOPICS

- What are Large Language Models (LLMs)?
- **Overview of Open Source LLMs:** DeepSeek, Llama, GPT-J
- Local Deployment vs Cloud Deployment
- Hardware Requirements & Options
- Setting Up Development Environment (Colab Free Tier, Local Machine optional)

02

Lab

Understanding and Adapting LLMs

TOPICS

Anatomy of pre-trained LLMs · Tokenization and Embedding Mechanisms · Fine-tuning Methods (Full, LoRA, QLoRA) · Best Practices for Sensitive Data Fine-tuning

03

Lab

Data Preparation & Responsible AI

TOPICS

Collecting and Preparing Sensitive Data · Data Cleaning & Tokenization · Anonymization & Privacy · Responsible AI & Ethical Considerations

04

Deployment and Optimization

TOPICS

Exporting and Running Models Locally · Llama.cpp / GGUF Deployment (CPU-based) · Model Optimization for Inference · API Deployment (FastAPI / Gradio / Streamlit)

05

Lab

Scaling, Security & Real-world Applications

TOPICS

- Scaling (Colab Pro, Local GPU, Cloud GPUs)
- Security Best Practices
- **Real-world Use Cases:** Healthcare, Finance, Legal

HANDS-ON LABS

Practise What You Learn

#	MODULE	ACTIVITY
01	**Introduction to LLMs and Setup**	Set Up Google Colab Environment for LLMs
02	**Understanding and Adapting LLMs**	Fine-tune DeepSeek Coder 6.7B (Quantized) on a Small Dataset
03	**Data Preparation & Responsible AI**	Preprocess and Prepare Sensitive Data for Fine-tuning
05	**Scaling, Security & Real-world Applications**	Deploy a Secure AI Agent Powered by DeepSeek Model

WHAT IS INCLUDED

Everything You Need to Succeed

- Koenig course material for every module
- Delivery by certified instructors and subject matter experts
- Hands-on exercises throughout the course
- Course completion certificate from Koenig

ABOUT KOENIG

Who We Are

Established in 1993, Koenig Solutions is a globally recognised training provider with 35+ technology partnerships, 5,000+ training programmes and a network of certified instructors and subject matter experts, supporting individuals and enterprise teams worldwide.

FLEXIBLE DELIVERY

Learn the Way That Suits You

Classroom Training | Live Online Classes (ILO) | Fly-Me-A-Trainer (FMAT) | Flexi (Self-Paced) | 1-on-1 Training

Ready to Get Started?

Talk to a learning advisor about schedules, group training and customised delivery.

Visit koenig-solutions.com | info@koenig-solutions.com