Faysal Aziz

Deploying DeepSeek and other LLMs Locally or On Premise

A Hands-On Guide for Advanced Users - 1 day

FA-0163 · 1 day

Agentic & Generative AI · Automation, RPA & Power Platform

Introduction

Ready to break free from the cloud and run AI on your own terms? Welcome to your hands-on guide to becoming an AI power user!

Ever wanted to have a ChatGPT-like assistant that's completely under your control? In this deep dive, we'll show you how to bring LLMs like DeepSeek right to your own machine. No more worrying about data privacy or usage limits – you're about to join the ranks of developers who run their AI locally.

Perfect for Python ninjas and Linux enthusiasts, this course goes beyond the basics. You'll learn the secret sauce behind making LLMs truly useful: from supercharging them with RAG for better context, to creating custom AI agents that automate your workflow. We'll get our hands dirty with tools like Hugging Face and Ollama, turning complex concepts into practical solutions that you can actually use.

Whether you're looking to build a secure AI setup for your organization or just want the freedom to tinker with these powerful models, you'll walk away with the skills to make it happen. No marketing fluff – just real, applicable knowledge for bringing advanced AI capabilities to your own hardware.

Learning outcomes

- Understand the architecture and deployment processes of LLMs, with a focus on DeepSeek.
- Set up and configure LLMs on Linux Ubuntu systems using Python and Hugging Face.
- Implement Retrieval-Augmented Generation (RAG) to enhance model performance.
- Apply transfer learning techniques to customize models for specific tasks.
- Explore agent-based AI automation for dynamic decision-making processes.

Prerequisites

- Extensive experience with Python programming.
- Proficient understanding of Linux or POSIX operating systems.
- Access to hardware equipped with a GPU boasting at least 20,000 CUDA cores.

Training outline

1. Introduction to Large Language Models (LLMs)

- Overview of LLMs and their significance in modern AI applications.
- Benefits of local deployment versus cloud-based solutions.

2. DeepSeek: A Comprehensive Examination

- Introduction to DeepSeek and its R1 reasoning model.
- Comparative analysis: DeepSeek's R1 reasoning versus other models.
- Use cases and performance benchmarks.

3. Setting Up the Environment

- Preparing Linux Ubuntu for LLM deployment.
- Installing and configuring Python and essential libraries.
- Leveraging Hugging Face for model management and deployment.

4. Deploying DeepSeek Locally

- Step-by-step guide to installing DeepSeek-R1 using Ollama.
- Configuring system resources to optimize performance.
- Troubleshooting common installation and runtime issues.

5. Enhancing Models with Retrieval-Augmented Generation (RAG)

- Understanding RAG and its role in improving LLM outputs.
- Integrating external knowledge bases for enriched responses.
- Implementing RAG in Python applications.

6. Applying Transfer Learning

- Concepts and benefits of transfer learning in LLMs.
- Techniques for fine-tuning models to cater to specific tasks.
- Practical exercises in customizing models using transfer learning.

7. Agent-Based AI Automation

- Exploring agentic RAG and its applications in dynamic decision-making.
- Designing intelligent agents for automated processes.
- Implementing agent-based systems in Python.

8. Optimizing Performance and Ensuring Security

- Best practices for maximizing model efficiency on local hardware.
- Security considerations in local AI deployments.
- Regular maintenance and updates for sustained performance.

This course promises a high-tech, engaging experience, empowering participants to harness the full potential of AI models within their own infrastructure.

© 2026 Faysal Aziz · Kuala Lumpur