Course Overview

Modern AI systems are no longer just notebooks and models—they are **applications that must be served, scaled, and integrated into real products**. This course teaches AI engineers how to build production-ready AI APIs using **FastAPI**, one of the most powerful and widely adopted Python frameworks for high-performance backend development.

Throughout the course, you will learn how to transform machine learning and LLM models into **robust, scalable services**. Starting with FastAPI fundamentals, the course gradually moves toward advanced topics such as **async programming, structured validation, AI model serving, retrieval systems, and background processing**.

By the end of the course, you will build a **complete AI backend system from scratch**, gaining the practical skills needed to deploy AI capabilities as real-world applications rather than isolated scripts.

Course Objectives

By completing this course, learners will be able to:

- Understand why FastAPI is particularly suited for AI and machine learning applications
- Build structured and maintainable API backends for AI systems
- Implement robust request validation using modern Python data models
- Serve machine learning and LLM models through production-style APIs
- Handle long-running AI tasks using asynchronous and background processing
- Integrate retrieval pipelines into AI APIs
- Design and implement a complete AI backend architecture

What You Will Learn

In this course, you will explore practical techniques for building AI-powered APIs, including:

- Designing high-performance AI APIs using FastAPI
- Structuring real-world AI projects for maintainability and scalability
- Implementing data validation and schema enforcement using Pydantic
- Using asynchronous programming to handle concurrent AI workloads
- Serving machine learning and language models through API endpoints
- Building retrieval-powered AI systems with simple RAG pipelines
- Implementing middleware for logging, monitoring, and observability
- Managing long-running AI computations using background tasks

---

Prerequisites

To benefit fully from this course, learners should have:

- Basic knowledge of **Python programming**
- Familiarity with **REST APIs**
- Basic understanding of **machine learning or AI workflows**
- Some experience with **Python libraries such as NumPy, Pandas, or ML frameworks**

Prior experience with FastAPI is **not required**.

Who This Course Is For

This course is ideal for:

- **AI engineers** who want to deploy models as scalable APIs
- **Machine learning engineers** transitioning from research to production systems
- **Backend developers** interested in building AI-powered services

---

- **Data scientists** who want to operationalize their models
- **LLM application developers** building real-world AI tools
- **Software engineers** exploring modern Python frameworks for AI systems

---

Course Syllabus

Chapter 1 — Why FastAPI for AI Engineers?
Understand the limitations of traditional AI workflows and why modern AI systems require API-driven architectures.

Chapter 2 — FastAPI Fundamentals (With Real Interaction)
Learn the essential concepts needed to build and run APIs using FastAPI, including routing, request handling, and interactive documentation.

Chapter 3 — Pydantic & Validation (Critical for AI)
Explore how structured data validation ensures reliability in AI systems and prevents common input-related failures.

Chapter 4 — Project Structure for Real AI Apps
Learn how to organize FastAPI applications using modular architecture patterns suitable for large AI systems.

Chapter 5 — Async Programming for AI APIs
Understand asynchronous programming and why it is crucial for handling concurrent AI workloads efficiently.

Chapter 6 — Serving ML Models
Learn how to load, manage, and serve machine learning models through scalable API endpoints.

Chapter 7 — Building an LLM API
Create an API interface for large language models and implement structured request-response pipelines.

Chapter 8 — Building a Mini RAG API
Develop a simplified Retrieval-Augmented Generation system that combines document retrieval with language model responses.

Chapter 9 — Middleware, Logging & Monitoring
Implement observability features that track system behavior, monitor requests, and improve reliability.

Chapter 10 — Background Tasks & Long Inference
Handle computationally heavy AI workloads using background task processing and queue-based architectures.

Chapter 11 — Final Capstone Project
Build a complete AI backend system that integrates model serving, retrieval pipelines, and asynchronous processing into a single production-style API.

Skills You Will Gain

After completing this course, you will develop practical skills such as:

- Designing scalable **AI backend architectures**
- Building **production-grade FastAPI applications**
- Implementing **robust request validation for AI APIs**
- Deploying **machine learning and LLM models as services**
- Managing **asynchronous and long-running AI workloads**
- Integrating **retrieval pipelines with AI models**
- Creating **maintainable and observable AI systems**