

Generative AI & LLM Engineering

Course Duration ~120 Hours · 16 Weeks · Live Online · Project-Based · Capstone Reviewed 1:1

RAG

Agentic AI

MCP

Evals & Observability

Fine-Tuning

LLMOps

In short: This is a live, project-based program that takes you from calling an LLM API to shipping evaluated, observable, production-grade AI systems. Across 12 modules you build RAG pipelines with citations, multi-agent systems on LangGraph, custom MCP servers, automated eval harnesses, LoRA fine-tunes, guardrail layers and autoscaled LLMOps deployments — closing with a capstone you design, build, evaluate, deploy and defend in a 1:1 review. Built for fresh IT graduates targeting an AI-first first role and for working professionals converting existing engineering skills into AI engineering.

Duration	16 weeks · ~120 hours · live + hands-on	Format	Live online · project-based · capstone reviewed 1:1
Prerequisite	Working Python. No ML/math background required.	Curriculum	12 modules, from transformer foundations to LLMOps and capstone
Portfolio	6 production-style projects plus a defended capstone	Ideal for	Fresh IT graduates and working professionals moving into AI engineering

Course Highlights

- Every module leads with what you build, not what you "learn about"
- Production RAG: hybrid search, re-ranking, GraphRAG, chunk-level citations
- Agentic AI with LangGraph, CrewAI and the Model Context Protocol (MCP)
- Evaluation-first engineering: eval sets, LLM-as-judge, regression suites in CI
- Fine-tuning with LoRA/QLoRA and DPO — and knowing when not to fine-tune
- AI safety and security aligned to the OWASP Top 10 for LLM applications
- LLMOps: vLLM serving, caching, autoscaling, CI/CD for prompts and models

Hands-on Projects & Portfolio

- **Enterprise RAG Assistant** — multi-source RAG over enterprise PDFs with hybrid search, re-ranking and inline citations
- **Custom MCP Server** — exposing internal tools to a customer-facing agent through one standard interface
- **Multi-Agent Research System** — research-and-report system built on LangGraph with planning, tool use and human-in-the-loop
- **Domain Fine-Tune with LoRA** — domain-specific fine-tuned model benchmarked against the base model
- **Automated Eval Harness** — nightly regression runs with a hallucination-rate dashboard wired into CI
- **Cost-Optimized Doc Pipeline** — document processing at scale using model routing and prompt caching

Tools & Stack

Python · PyTorch · Hugging Face · OpenAI · Anthropic Claude · Gemini · Llama · LangChain · LangGraph · LlamaIndex · CrewAI · MCP · pgvector · Qdrant · Pinecone · Weaviate · Neo4j · RAGAS · Langfuse · vLLM · FastAPI · Docker · AWS Bedrock

Career Outcomes

Demand for applied AI engineers who can ship RAG plus eval pipelines is growing far faster than the supply of engineers who can build them. This course targets the roles at the centre of that gap:

- AI Engineer · LLM Engineer · Applied AI Engineer · RAG Developer
- Agentic AI Engineer · AI Platform Architect · MLOps / LLMOps Engineer

Who This Course Is For

- Final-year students and fresh IT graduates targeting an AI-first first role
- Software / backend / full-stack engineers converting existing skills into AI engineering
- Data analysts, data scientists and ML engineers moving into LLM and agent systems
- Tech leads and architects who need to make build-vs-buy calls on GenAI systems

Module 1: Foundations of Generative AI & Transformers

8 Hours

YOU BUILD

- A model-selection playbook for a real business task — token-level cost analysis, context-window budgeting, and a scored comparison across model families

TOPICS

- How LLMs are actually trained; tokenization, context windows, attention intuition
- Model families: GPT, Claude, Gemini, Llama, DeepSeek, Mistral
- Open vs closed trade-offs; choosing a model for a job

Assessment: Quiz + Lab

Module 2: Production Prompt Engineering

8 Hours

YOU BUILD

- A versioned prompt library with structured JSON-schema outputs and an automated prompt regression test suite

TOPICS

- System prompts as versioned software contracts; zero/few-shot, chain-of-thought
- Structured outputs and JSON schemas; XML-tag structuring
- Prompt caching; prompt regression testing

Assessment: Quiz + Lab

Module 3: LLM API Engineering

10 Hours

YOU BUILD

- A streaming, tool-calling LLM service with retries and backoff, token budgeting, and cheap-model/strong-model routing

TOPICS

- Streaming; function/tool calling; multimodal and vision inputs
- Retries and backoff; token budgeting; latency optimization
- Model routing: cheap model to extract, strong model to reason

Assessment: Quiz + Lab

Module 4: Embeddings & Vector Databases

8 Hours

YOU BUILD

- A semantic search service over your own corpus with tuned chunking, metadata filtering, and index benchmarking on pgvector and Qdrant

TOPICS

- Embedding models; chunking strategies; similarity search
- pgvector, Qdrant, Pinecone, Weaviate
- Metadata filtering; index tuning

Assessment: Quiz + Lab

Module 5: RAG Systems That Actually Work

12 Hours

YOU BUILD

- A multi-source RAG assistant over enterprise PDFs with hybrid search (BM25 + dense), re-ranking, query rewriting, and chunk-level citations

TOPICS

- Naive vs advanced RAG; contextual retrieval
- Hybrid search (BM25 + dense); re-ranking; query rewriting
- GraphRAG with Neo4j; handling tables and PDFs; chunk-level citations

Assessment: Quiz + Lab**Module 6: Agentic AI & Orchestration**

12 Hours

YOU BUILD

- A multi-agent research-and-report system on LangGraph with planning, tool use, memory, and human-in-the-loop checkpoints

TOPICS

- ReAct loops; planning; tool use; memory
- LangGraph, CrewAI, LlamaIndex
- Multi-agent patterns and their failure modes; long-running/durable agents

Assessment: Quiz + Lab**Module 7: Model Context Protocol (MCP) & Tool Integration**

8 Hours

YOU BUILD

- A custom MCP server that exposes internal tools to a customer-facing agent, with sandboxed execution via E2B/Modal

TOPICS

- MCP servers and clients; A2A
- Exposing internal tools to agents; sandboxed execution (E2B/Modal)
- Replacing bespoke integration layers with a standard interface

Assessment: Quiz + Lab**Module 8: Evaluation & LLM Observability**

12 Hours

YOU BUILD

- An automated eval harness with nightly regression runs, a calibrated LLM-as-judge, and a hallucination-rate dashboard traced end-to-end

TOPICS

- Building eval sets from day one; RAGAS; retrieval precision/recall
- LLM-as-judge and judge calibration; regression suites in CI
- Tracing and monitoring with Langfuse / Phoenix / Helicone; cost and latency dashboards

Assessment: Quiz + Lab**Module 9: Fine-Tuning & Model Customization**

10 Hours

YOU BUILD

- A domain-specific model fine-tuned with LoRA and quantized for serving, benchmarked head-to-head against the base model

TOPICS

- When *not* to fine-tune; SFT; LoRA/QLoRA; PEFT
- DPO and preference data; Hugging Face pipelines
- Quantization; evaluating a fine-tune against the base model

Assessment: Quiz + Lab

Module 10: AI Safety, Guardrails & Security

8 Hours

YOU BUILD

- A guardrail layer for an existing agent — prompt injection defense, PII redaction, output validation, and an audit trail — tested against the OWASP LLM Top 10

TOPICS

- OWASP Top 10 for LLM applications
- Prompt injection and indirect injection defense; jailbreak resistance
- PII redaction; content filtering; output validation; responsible AI and audit trails

Assessment: Quiz + Lab**Module 11: LLMops & Deployment**

12 Hours

YOU BUILD

- A containerized, autoscaled AI service — FastAPI + vLLM with caching layers, CI/CD for prompts and models, and a cost-control dashboard

TOPICS

- Containerizing AI services; FastAPI serving; vLLM and inference optimization
- Caching layers; autoscaling; cost control at scale
- CI/CD for prompts and models; on-call for non-deterministic systems

Assessment: Quiz + Lab**Module 12: Capstone**

12 Hours

YOU BUILD

- One production-grade agentic AI application, end-to-end: designed, built, evaluated, deployed — and defended in a 1:1 review

TOPICS

- Architecture design review and scoping
- Evaluation plan, deployment, and observability
- 1:1 capstone defense

Assessment: Capstone Project + 1:1 Review