AI Engineering & MLOps

Build Production AI Systems That Hold Up Under Review

This course is for learners who already have data and ML foundations and now need production AI engineering discipline. Instead of stopping at notebooks or simple API calls, the track moves into structured outputs, retrieval quality, agent workflows, evaluation, local inference awareness, and MLOps habits that make AI systems more dependable.

The focus is not on collecting every buzzword in the ecosystem. The focus is on building systems you can explain, test, monitor, and improve when something goes wrong.

Why This Course?

The Market Reality

Global Context: Teams are moving from AI demos to production systems that need clearer retrieval, evaluation, guardrails, and monitoring. The engineering challenge is no longer just calling a model. It is making the system dependable.

Nepal Context: Nepal's AI ecosystem continues to grow through product companies, service teams, research organizations, and remote engineering work. The gap is not only model access. It is in engineers who can turn AI ideas into maintainable services with testing, monitoring, and better cost awareness.

Your Opportunity: This course positions you for **AI engineer, applied ML engineer, and MLOps engineer roles** where system reliability, retrieval quality, and production judgment matter more than notebook-only accuracy.

Nepal-Relevant Reality	Opportunity
Teams are experimenting with AI but struggle to ship it cleanly	Production-focused AI engineers stand out quickly
Many practitioners stop at notebooks and one-off demos	You learn service boundaries, evaluation, and deployment habits
RAG and agent workflows are becoming common product patterns	Retrieval and workflow-control experience becomes practical proof
Cost and reliability matter as much as model quality	Better judgment makes your work more valuable

Course Snapshot

Parameter	Details
Course Code	TR-10
Title	AI Engineering & MLOps
Duration	3 Months (12 Weeks)
Schedule	Monday to Friday (Mon–Fri, 5 Days/Week), 2 Hours/Day
Total Hours	120 Hours of Live Training
Batch Size	Maximum 10 Students
Course Fee	NPR 40,000
Prerequisites	Completion of Data Science & Machine Learning (TR-02) or equivalent is required. You should already be comfortable with Python, pandas, basic model training and evaluation, notebook workflow, and reading technical documentation. Before Day 1, review train/test split, leakage, prompt basics, and transformer fundamentals, then complete the AI readiness assessment. Saarathi Gate Assessment and advanced readiness review before Day 1.
Self-Study	Minimum 2 hours/day outside class (mandatory)
Outcome	AI Engineer / MLOps Engineer

Your Learning Week

Day	Activity
Mon–Fri	2-hour live class session (hands-on, project-based)
Mon–Fri	Minimum 2 hours self-study & practice (mandatory)
Saturday	No classes - flexible self-study, peer collaboration, project work
Sunday	Whole day self-learn time. Classrooms remain fully open for you to come in, study, collaborate with peers, and build projects.

> *This is an advanced track. Daily practice outside class is what turns good ideas into reliable systems you can defend during review.*

Week-by-Week Curriculum

Phase 1: Production AI Foundations & RAG (Weeks 1–3, 3 Weeks, 30 Hours)

Week	Focus Area	What You'll Master
Week 1	Production AI Foundations & Structured Outputs	L1 recap, prompt discipline, structured outputs, provider APIs, and the difference between a demo and a dependable service
Week 2	Vector Databases & Retrieval Quality	Embeddings, chunking, metadata, vector stores, and retrieval-quality trade-offs
Week 3	RAG Pipeline Development & Evaluation	RAG architecture, retrieval tuning, citation-aware responses, and evaluation basics

Phase 2: Agent Workflows & Advanced Retrieval (Weeks 4–6, 3 Weeks, 30 Hours)

Week	Focus Area	What You'll Master
Week 4	Tool-Using Agents & Workflow Control	Agent building blocks, tools, memory boundaries, and failure-aware workflow design
Week 5	LangGraph Multi-Step Systems	State graphs, branching workflows, checkpointing, and human-in-the-loop patterns
Week 6	CrewAI & Advanced Retrieval Patterns	Role-based orchestration, routing, reranking, hybrid retrieval, and framework comparison

Phase 3: Model Adaptation, Local Inference & Evaluation (Weeks 7–9, 3 Weeks, 30 Hours)

Week	Focus Area	What You'll Master
Week 7	Fine-Tuning Decisions & PEFT Workflows	When fine-tuning helps, dataset prep, LoRA/QLoRA workflows, and adaptation trade-offs
Week 8	Local Inference & Optimization	Ollama, llama.cpp awareness, quantization, local serving, and privacy/cost trade-offs
Week 9	Evaluation, Safety & Reliability	Benchmarks, rubric design, prompt-injection review, bias awareness, and guardrails

Phase 4: MLOps, Capstone & Career (Weeks 10–12, 3 Weeks, 30 Hours)

Week	Focus Area	What You'll Master
Week 10	MLOps Fundamentals	Experiment tracking, versioning, service boundaries, monitoring, and deployment workflow
Week 11	Production AI Capstone	Cost and latency awareness, failure handling, service integration, and capstone build
Week 12	Capstone Finalization & Career Launch	Portfolio packaging, architecture explanation, system-design discussion, and interviews

Skills You'll Gain

Languages & Frameworks

Technology	Proficiency Level
Python	Production AI implementation
LangChain & LangGraph	Retrieval and agent workflows
RAG	Retrieval-backed application design
MLflow	Experiment tracking and model workflow discipline
FastAPI	Service delivery and API boundaries
Ollama	Local inference awareness
Evaluation & Guardrails	Quality and safety review

Development Tools

Tool	Application
Jupyter Notebooks	Prototyping and evaluation
Git & GitHub	Version control and reviewable delivery
Docker	Packaging and deployment workflows

Topic Depth and Awareness

Section	Guidance
Purpose	This course intentionally separates what you need to master in depth from what you only need to understand with working awareness.
Depth	Structured outputs, retrieval quality, RAG workflows, agent orchestration, evaluation habits, and production delivery patterns practiced repeatedly in class
Awareness	Heavier fine-tuning, deeper local inference optimization, and broader vendor-specific platform choices introduced as comparison context
How to use this syllabus	Spend most of your self-study time strengthening the depth topics first. Use awareness topics to broaden judgment, not to split your focus too early.

Project Pool

> *All options below are **intermediate-level final projects**. Each student chooses **one** final project from this pool. Trainers may run smaller guided exercises during the course, but public phase-wise project sections are intentionally removed so the completion standard stays clear and consistent.*

#	Final Project Choice	What You Will Build	Core Stack / Tools
1	RAG Knowledge Base API	Build a retrieval-backed question-answering workflow with ingestion, evaluation, and service delivery basics.	FastAPI, LangChain / LangGraph, vector database, RAG evaluation
2	Document Intelligence Assistant	Build a document-processing assistant with chunking, retrieval, traceability, and human-readable outputs.	RAG, document parsing, embeddings, API integration
3	Evaluation & Monitoring Pack	Build an evaluation and monitoring layer for prompts, retrieval quality, latency, and failure modes.	RAGAS, evaluation metrics, monitoring, prompt testing
4	Agentic Research Workflow	Build a guided multi-step research assistant with tools, memory boundaries, and answer validation.	LangGraph / agent tooling, tool orchestration, memory design, workflow control

#	Final Project Choice	What You Will Build	Core Stack / Tools
5	Prompt & Retrieval Optimization Pack	Run structured experiments on prompt design, chunking, retrieval, and output reliability.	Embeddings, retrieval tuning, prompt engineering, quality analysis

Career Paths & Trajectory

Role Path	Focus and Proof	Stage and Timeline	What Actually Matters
Junior AI Engineer / Applied ML Engineer	Build retrieval-backed or model-driven application features with better evaluation and deployment discipline. Proof you leave with: RAG workflows, serving basics, and observability habits	Entry role - first 0–2 years	Reliable implementation, clean APIs, and understanding where AI systems fail in production.
AI Engineer / MLOps Engineer	Own inference services, evaluation loops, monitoring, and deployment workflows for production AI systems. Proof you leave with: RAG, agents, MLflow, and deployment-ready project proof	Growth role - 2–4 years	Strong system thinking, better latency and cost trade-offs, and fewer notebook-only solutions.
Senior AI Engineer / ML Platform Engineer	Improve shared AI infrastructure, evaluation standards, and team-wide deployment patterns. Proof you leave with: Monitoring, model versioning, and stronger production architecture judgment	Specialist path - 4–6 years	Platform reliability, guardrails, and helping teams ship AI systems with more discipline.
AI Solutions Architect / Applied AI Lead	Guide AI system choices around business value, safety, cost, and long-term maintainability. Proof you leave with: System design artifacts, vendor trade-off judgment, and stronger architecture reasoning	Senior design path - 6+ years	Good judgment on when to use AI, how to evaluate it, and how to align engineering choices with business reality.

Saarathi Gate & Completion Review

Before You Start: Saarathi Gate Assessment

All students complete the **Saarathi Gate Assessment** before Day 1. It is a short diagnostic review of aptitude, learning behaviour, and thinking style. It has **no pass/fail** and is used only to tailor support from the start.

After Course Completion: Saarathi Completion Review

The **Saarathi Academy Certificate** is issued after the selected final project is completed, documented, and reviewed by the trainer. There is **no separate certification exam** for this course.

Completion Requirements:

1. **Attendance:** Minimum 80% attendance
2. **Weekly Work:** Core deliverables, revision work, and practice tasks completed
3. **Final Project:** One intermediate-level project selected from the project pool and completed to trainer-approved quality
4. **Portfolio Proof:** Screenshots, documentation, case-study notes, or equivalent proof assets updated
5. **Trainer Review:** Practical execution, consistency, communication, and overall growth signed off by the trainer

Enrollment & Next Steps

Next Batch: Starting soon (contact for exact dates) **Offline Location:** Old Baneshwor Chowk, Kathmandu, Nepal **Mode:** Online + Offline **Contact (Call/WhatsApp):** 9761095364, 9744442469

» **[ENROLL NOW]** - Limited to 10 seats per batch

Reliable AI systems come from better retrieval, clearer evaluation, and stronger engineering discipline - not from hype.

Last Updated: Mar 30, 2026