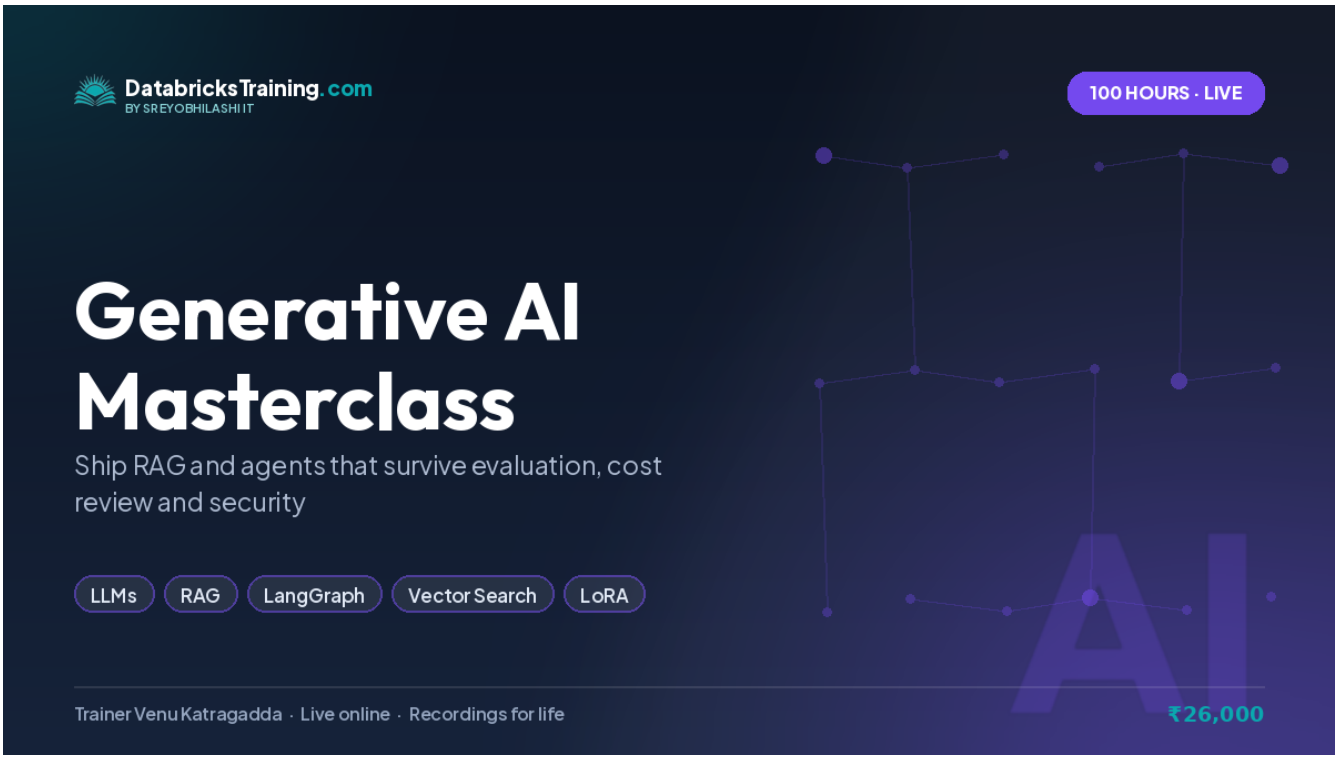

COURSE SYLLABUS

Generative AI Masterclass

Engineering-grade GenAI: you will ship a RAG system and an agent that survive evaluation, cost review and a security audit.

100 hours LIVE INSTRUCTION	10 modules STRUCTURED PATH	3 projects PORTFOLIO READY	**₹26,000** EMI AVAILABLE

Level	Intermediate to Advanced
Mode	Live, instructor-led online sessions. Every session recorded; lifetime access.
Trainer	Venu Katragadda — 14+ years in industry (L&T Infotech, Capgemini, Brillio, Manthan Systems). 1,200+ professionals trained.
Batch timings	Morning 6:00–7:30 AM IST · Morning 7:30–9:00 AM IST · Evening 7:30–9:00 PM IST · Weekend (Sat & Sun)
Certification	Databricks Generative AI Engineer Associate · Azure AI Engineer (AI-102) concepts
Enquiries	+91-8500002025 / +91-9247159150 (WhatsApp) · info@databrickstraining.com

What you will be able to do

Every outcome below is demonstrated in a lab or a graded project — not just discussed in a slide.

- Explain transformers, tokenisation, context windows and why models fail the way they do
- Build production RAG: chunking, embeddings, hybrid search, reranking, citations, evaluation
- Fine-tune open models with LoRA/QLoRA and know when fine-tuning is the wrong answer
- Build agents and tool-use systems, including MCP servers, with real guardrails
- Evaluate systematically (RAGAS, LLM-as-judge, golden sets) instead of vibes
- Control cost and latency: model routing, caching, batching, quantisation — with a real unit-economics model

Who this masterclass is for

- Data and software engineers building AI features
- Data scientists moving from classic ML to LLMs
- Architects evaluating GenAI platforms
- Trainers and tech leads who need depth, not demos

Prerequisites

- Solid Python
- Basic ML concepts helpful but not required
- API keys for at least one provider; open-source models run on free Colab/Kaggle GPUs

Tools and technologies

OpenAI GPT-5.x · Anthropic Claude 4.x · Google Gemini 2.x · Llama 4 · Mistral · Qwen · LangChain · LangGraph · LlamaIndex · Hugging Face · vLLM · Ollama · pgvector · Pinecone · Chroma · Weaviate · FAISS · Databricks Vector Search · MLflow · LangSmith · RAGAS · MCP · Amazon Bedrock · Azure AI Foundry · Vertex AI

Curriculum — 100 hours across 10 modules

Sessions run live in small batches. Roughly 60% of each session is hands-on lab work.

MODULE 1 Foundations: how LLMs actually work — 10 hrs

- Tokenisation, embeddings, attention, transformer blocks — built up from first principles
- Pretraining vs instruction tuning vs RLHF/RLAIF; base vs chat models
- Context windows, KV cache, temperature/top-p, sampling, structured output and JSON mode
- Reasoning models and extended thinking: when they help and what they cost
- The 2026 model landscape: GPT-5.x, Claude 4.x, Gemini 2.x, Llama 4, Mistral, Qwen — capability and price comparison
- Lab: token accounting and a cost model for a real workload

MODULE 2 Prompt engineering that survives production — 8 hrs

- System vs user vs tool messages; role design and instruction hierarchy
- Few-shot, chain-of-thought, self-consistency, decomposition, ReAct
- Structured outputs: JSON schema, function calling, constrained decoding, Pydantic validation
- Prompt versioning, templating, regression testing, prompt injection basics
- Context engineering: what to put in the window and what to leave out
- Lab: take a flaky prompt from 62% to 94% task accuracy on a golden set

MODULE 3 Embeddings, vector search and retrieval — 12 hrs

- Embedding models compared; dimensionality, normalisation, multilingual, domain adaptation
- Vector indexes: HNSW, IVF, ScaNN; recall vs latency vs memory trade-offs
- Vector stores: pgvector, Pinecone, Chroma, Weaviate, Qdrant, Databricks Vector Search, OpenSearch
- Hybrid search: BM25 + dense, reciprocal rank fusion, metadata filtering
- Reranking with cross-encoders; query rewriting and HyDE
- Lab: measured retrieval quality across five chunking strategies

MODULE 4 Building production RAG — 14 hrs

- Document ingestion: PDFs, tables, images, HTML; layout-aware parsing; OCR
- Chunking strategies: fixed, recursive, semantic, hierarchical, contextual retrieval
- Grounding, citations, refusal behaviour, handling 'I don't know'
- Advanced patterns: parent-document, multi-vector, GraphRAG, agentic RAG
- Freshness: incremental indexing, deletion, permissions-aware retrieval (row-level security)
- Serving architecture: API layer, caching, streaming responses, rate limits
- Lab: Project 1 end to end

MODULE 5 Evaluation, observability and guardrails — 10 hrs

- Building golden datasets; task-specific metrics; inter-annotator agreement
- RAGAS: faithfulness, answer relevance, context precision/recall
- LLM-as-judge: rubrics, position bias, calibration, when it lies to you
- Tracing with LangSmith / MLflow / OpenTelemetry; token and latency dashboards
- Guardrails: input/output filters, PII redaction, jailbreak and prompt-injection defence
- Regression gates in CI: fail the build when quality drops

- Lab: build an eval harness and wire it into GitHub Actions

MODULE 6 Agents and tool use — 12 hrs

- Agent loops: ReAct, plan-and-execute, reflection; when an agent is the wrong abstraction
- Function/tool calling in depth; parallel tools; error and timeout handling
- LangGraph: state machines, checkpointing, human-in-the-loop, subgraphs
- Multi-agent patterns: supervisor, hand-off, debate — and their failure modes
- Model Context Protocol (MCP): building an MCP server, resources, tools, prompts
- Computer-use and browser agents; sandboxing and blast-radius control
- Lab: Project 2 agent with red-team suite

MODULE 7 Fine-tuning and model adaptation — 12 hrs

- Decision framework: prompt → RAG → fine-tune → pretrain (and the cost of each)
- SFT dataset construction, formatting, deduplication, contamination checks
- PEFT: LoRA, QLoRA, adapters; rank/alpha selection; catastrophic forgetting
- Preference tuning: DPO, ORPO; RLHF at a conceptual level
- Quantisation: GGUF, AWQ, GPTQ, bitsandbytes; accuracy vs memory trade-offs
- Distillation to a smaller model for cost reduction
- Lab: Project 3 fine-tune with a proper eval comparison

MODULE 8 Serving, LLMOps and cost engineering — 10 hrs

- Inference servers: vLLM, TGI, Ollama, TensorRT-LLM; continuous batching, paged attention
- GPU sizing and economics: A10G vs L4 vs A100 vs H100; self-host vs API break-even maths
- Model routing and cascades; semantic caching; prompt caching; batch APIs
- Deployment on Databricks Model Serving, Amazon Bedrock, Azure AI Foundry, Vertex AI
- Versioning, canary rollouts, shadow traffic, rollback
- Lab: cut cost per 1,000 requests by 70% without losing quality

MODULE 9 Multimodal, security and responsible AI — 6 hrs

- Vision, audio and document understanding; multimodal RAG
- Image and video generation overview; when it matters for enterprise work
- OWASP Top 10 for LLM applications; supply-chain risk in model weights
- Data privacy, residency, DPAs, retention; what you can and cannot send to an API
- EU AI Act and NIST AI RMF at a practitioner level; model cards and documentation

MODULE 10 Architecture, capstone and career — 6 hrs

- Reference architectures for enterprise GenAI on AWS, Azure, GCP and Databricks
- Build vs buy; platform selection scorecard
- Capstone presentation and architecture defence
- GenAI engineer interview questions, portfolio framing, resume rewrite
- Staying current: how to read papers and release notes without drowning

Hands-on projects

You leave with work you can demo in an interview. Each project is code-reviewed and returned with written feedback.

PROJECT 1 Enterprise RAG over your own documents

Ingestion → chunking strategy comparison → hybrid retrieval → reranking → grounded answers with citations, plus a RAGAS evaluation report and a cost-per-query model.

PROJECT 2 Multi-tool agent

A LangGraph agent with SQL, search and internal-API tools, human-in-the-loop approval, retry/fallback logic, tracing and a red-team test suite.

PROJECT 3 Fine-tune and serve

LoRA fine-tune a small open model for a domain task, evaluate against the base model and a frontier API, then serve it with vLLM behind a cost/latency dashboard.

Frequently asked questions

Which models and versions do you teach?

Frontier APIs (OpenAI GPT-5.x, Anthropic Claude 4.x, Google Gemini 2.x) plus open-weight models (Llama 4, Mistral, Qwen) run locally with Ollama/vLLM. Model versions move fast — the course tracks current releases each batch and the principles are deliberately model-agnostic.

Do I need a GPU?

No. API work needs nothing special, and fine-tuning labs run on free Colab/Kaggle T4 GPUs using QLoRA. If you have a 16 GB+ GPU you can run everything locally.

What will the API calls cost me?

Budget roughly $20–$40 in API credits across the course. We use small models for iteration and teach cost control from Module 1 — several students finish under $15.

Is this course for developers or for managers?

Developers and architects. There is real code in every module. If you need a leadership-level overview, the 2-day executive workshop under Corporate Training is the better fit.

How to enrol

- Book a free demo. Sit in on a live session with the current batch — no card, no commitment. WhatsApp +91-9247159150 or call +91-8500002025.
- Pick your batch. Weekday morning, weekday evening or weekend. All times IST.
- Pay the fee (₹26,000, instalments and EMI available) and get your workspace, material and calendar invites.
- Start building. Labs from session one; recordings posted the same day.

What is included

- 100 hours of live instruction by Venu Katragadda
- Lifetime access to all session recordings
- 3 code-reviewed portfolio projects
- Preparation for: Databricks Generative AI Engineer Associate · Azure AI Engineer (AI-102) concepts
- Resume review, LinkedIn optimisation and one mock interview with written feedback
- Free re-attend of any future batch of the same masterclass
- Doubt-clearing sessions and a private student group

Refund policy in brief

When you cancel	Refund
Before the batch starts	Up to 90%
Within 5 days of start, max 2 classes attended	Up to 70%
After 7 days, or more than 5 classes attended	No refund — free transfer to a later batch

Full policy: databrickstraining.com/refund-policy/

Talk to us

Call	+91-8500002025 · +91-9247159150
WhatsApp	+91-9247159150 — usually answered within the hour
Email	info@databrickstraining.com
Website	databrickstraining.com/courses/generative-ai-masterclass/
Foundation tier	databrickstraining.in — shorter 70–80 hour courses if this one is too advanced
Institute	Sreyobhilashi IT · Udyam UDYAM-TS-02-0334643 · 112, Samshraya Homes, opp. PF Office, Balanagar, Kukatpally, Hyderabad, Telangana 500037