Generative AI Accelerator

29th September 2025

Where we are?

Intro to Generative AI

- Intro to AI
- History and evolution of AI
- Discriminative vs Generative AI
- AI Landscape
- Open Source vs Proprietary LMs
- Responsible AI - Ethics, Bias etc

LLM Foundations

- Transformers Intuition
- Next Token Prediction
- LLM Training Phases
- Pre-training/ Post-training
- Model behaviour parameters
- Different types of LLMs
- Limitations of LLMs

Prompt Engineering

- Prompting Techniques
- CoT/Tool Calling
- Prompt Evaluation
- Prompt Optimization
- Context Engineering
- Prompt Hacking/ Jailbreaks

RAG Systems

- RAG basics
- Different types of RAG
- **RAG Evaluation**
- Improving RAG performance

Designing LLM Systems

- Choosing the right stack
- Evaluating LLM systems
- Design Tradeoffs: Latency/Cost
- Performance Optimization
- Security and Privacy
- Case Study - Product Search

AI Agents

- Intro to Agents
- Tool Use and Memory
- Workflow vs Agents
- Agent orchestration patterns
- Agent Evaluation
- Model Context Protocol (MCP)

Capstone Project

- Project to apply your learnings
- Demo Day

Evaluating RAG Systems

How do we measure if Retrieval Augmented Generation really works?

Why evaluate RAG?

LLMs ≠ Truth Engines:
Large Language Models generate fluent text, not verified facts.

RAG ≠ Perfect Grounding:
Retrieval adds context, but irrelevant or missing documents can still mislead the model.

Real-World Impact:
Inaccurate responses can cause misinformation in customer support, wrong recommendations, or compliance risks in enterprises.

Evaluation = Feedback Loop:
It helps identify whether the problem lies in:
1️⃣ Retrieval → wrong or missing documents
2️⃣ Generation → hallucinated or incomplete answers
3️⃣ Grounding → weak connection between retrieved context and final output

👉 **Without systematic evaluation, you can't improve.**