Advanced Generative AI

Course Content

Lesson 1 - Scaling Generative AI Applications

- Distributed serving
- Model compression
- Load balancing

Lesson 2 - Advanced Fine-tuning Techniques

- Parameter-efficient fine-tuning (PEFT)
- LoRA adapters
- Quantization

Lesson 3 - Retrieval Augmented Generation (RAG)

- Hybrid search
- Knowledge retrieval
- Integrating RAG pipelines

Lesson 4 - Evaluation Metrics for Generative Systems

- BLEU, ROUGE, METEOR
- Human-in-the-loop evaluations
- Safety benchmarks

Lesson 5 - Responsible AI and Governance

- Fairness audits
- Model cards
- Data documentation

Lesson 6 - Human-in-the-Loop Systems

- Reinforcement learning with feedback
- Task routing
- Active learning

Lesson 7 - Cost Optimization for GenAI Deployments

- Efficient inference

- Hardware accelerators
- Cloud vs on-prem trade-offs

Lesson 8 - Domain-Specific GenAI (Legal, Healthcare, etc.)

- Domain constraints
- Regulatory compliance
- Ethical deployment

Lesson 9 - Scalability and Monitoring

- Infrastructure monitoring
- Scaling best practices
- Retraining strategies

Lesson 10 - Advanced GenAI Applications Showcase

- Complex conversational systems
- Vision-language models
- Multimodal architectures

Capstone Project

- Architect and deploy an advanced generative AI pipeline with evaluation metrics.