Comprehensive Course Recap

CS 435 Elective: Generative AI Security, Ethics, and Governance

Week 1 / Lecture 1

What Was Learned

- Introduction to Generative AI & LLMs
- Basic Transformer concepts
- Hyperscalers overview

Term: Generative AI

AI that creates new text, images, or data.

Term: LLM

Large Language Model trained on massive corpora.

Term: Transformer

Architecture using multi-head self-attention.

Term: Foundation Models

Base models adaptable to many tasks.

Term: Hyperscalers

Big cloud providers: AWS, GCP, Azure.

Term: AGI vs. Narrow AI

AGI: broad human-like intelligence; Narrow AI: specialized tasks.

Assignment: AI/Generative AI Timeline

Purpose

Purpose

Trace evolution from perceptrons to GPT-4.

What Was Learned

Historic milestones & milestone research papers.

Exercise: Recap Quiz 1

Alignment problem, fairness, trust, accountability.

What Was Learned

AGI vs. narrow AI
Paperclip thought experiment
Bias & logs for accountability

Week 2 / Lecture 2

What Was Learned

- Transformer deep dive
- Tokenization approaches
- AI libraries overview

Term: Transformer Architecture

Self-attention, parallel processing, scaling laws.

Term: Tokenization

Splitting text into subwords for model input.

Assignment: LeetCode Problem & LLM

Purpose

Purpose

Practice code generation & debugging with an LLM.

What Was Learned

Prompt engineering, next-token prediction, test-driven fixes.

Exercise: Recap Quiz 2

ML "tribes," RNN vs. Transformer, self-attention steps.

What Was Learned

Domingos' five tribes, ethical challenges in text generation.

Week 3 / Lecture 3

What Was Learned

- Retrieval-Augmented Generation (RAG)
- Vector DB fundamentals

Term: RAG

Combines LLM with external knowledge for factual answers.

Guest Lecture: Alla Abdella

Vector DB + RAG pipeline demonstration.

Week 4 / Lecture 4

What Was Learned

- Responsible Innovation
- Fairness, bias, accountability

Term: Fairness vs. Equity vs. Equality

Different ways AI outcomes treat demographics.

Assignment: Coding Debiaser

Purpose

Purpose

Mitigate bias with reweighing or other strategies.

What Was Learned

AIF360 usage, measuring disparate impact, Theil index.

Exercise: Recap Quiz

AI ethics, real-world bias examples, case studies.

Week 5 / Lecture 5

What Was Learned

- FAccT (Fairness, Accountability, Transparency)
- OWASP Top 10 for LLMs
- Red-teaming AI

Term: AI Security

Attacks: prompt injection, data poisoning, model extraction.

Assignment: AI Explainability 360

Purpose

Purpose

Explore LIME, SHAP, integrated gradients.

What Was Learned

Interpreting local vs. global model decisions.

Exercise: MCQ Quiz on Fairness Regulations

Key laws: Belmont, GDPR. Tools for bias mitigation.

Week 6 / Lecture 6

What Was Learned

- AI Governance Regulations
- MITRE ATLAS for ML threats

Term: EU AI Act

Risk-based approach for AI classification.

Term: MITRE ATLAS

Framework for adversarial ML tactics and mitigations.

Assignment: Regulations Assessment

Summarize global AI laws and code with AI assistance.

What Was Learned

US vs. EU patchwork, NYC Local Law 144, compliance best practices.

Exercise: Generative AI Security & Regulations MCQ

Safety vs. security, AWS vs. GCP vs. Azure compliance.

Week 7 / Lecture 7

What Was Learned

- AWS AI security
- Amazon Titan, Bedrock
- SageMaker Clarify, Model Monitor

Term: Amazon Bedrock

Managed generative AI service on AWS.

Term: SageMaker Clarify

Detects bias pre- and post-training.

Assignment: Hyperscalers AI Summary

Compare AWS, GCP, Azure features.

What Was Learned

Cost, security, bias detection in each platform.

Week 8 / Lecture 8

What Was Learned

- Federated Learning, DP
- Azure AI Content Safety
- Prompt Shields, Groundedness

Term: Federated Learning

Training models across distributed clients, protecting data locally.

Term: Differential Privacy

Noise added to preserve individual anonymity.

Term: Prompt Shields

Prevent malicious injection in LLM prompts.

Assignment: Privacy & Content Safety Quiz

Gradient inversion, brand risk detection, real-time constraints.

Week 9 / Lecture 9

What Was Learned

- RAG pipeline design
- LucidChart system diagrams
- GCP usage, Hugging Face integration

Term: FAISS / GPU-based vector DB

For semantic search and retrieval.

Assignment: Semantic Search & RAG

PDF ingestion, embeddings, GPT-based Q&A.

What Was Learned

Top-k retrieval, pipeline orchestration, better factual grounding.

Week 10 / Lecture 10

What Was Learned

- The Great Ethical AI Debate
- Structured debate on AI in classrooms, jobs, military

Activity: Debates

Team For vs. Team Against, final poll for persuasion shift.

Assignment: Debate Reflection

Summarize stance using real regulations & ethical frameworks.

Week 11 / Lecture 11

What Was Learned

- NeMo Guardrails, CoLang
- Brand-damaging content rules
- Hugging Face LLaMA usage

Term: NeMo Guardrails

Policy-based approach to block harmful queries.

Term: CoLang

Language to define LLM guardrail rules.

Term: LLaMA

Open-source large model from Meta (Hugging Face).

Assignment: AI Judge API

Classify brand-risk questions with moderation endpoints.

What Was Learned

REST calls, minimal JSON usage, prompt filtering logic.

Week 12 / Lecture 12

What Was Learned

- Ethical AI Startup Evaluation
- 10-question VC framework
- Market viability & compliance

Term: VC & Startup Terms

Seed, Series A/B, TAM, runway, exit strategy.

Assignment: Startup Evaluation

Analyze AI startup's ethics posture, growth potential.

What Was Learned

Governance, risk management, investing rationale.

Key Concept: Security & Adversarial ML

Prompt injection, data poisoning, safe LLM APIs.

Key Concept: Ethics & Governance

Fairness, transparency, accountability, bias mitigation.

Key Concept: Hyperscalers Ecosystem

AWS, GCP, Azure distinct AI toolchains & compliance.

Key Concept: Technical Foundations

Transformers, embeddings, RAG, fine-tuning, MLOps.

Key Concept: Responsible AI

Continuous monitoring, interpretability, regulated deployment.

Final Project

Custom validator for Guardrails AI or secure LLM.

Major Readings

Stochastic Parrots, OECD AI, UNESCO Ethics, NIST RMF.