Course Summary & Future Trends

CS 203: Software Tools and Techniques for AI

Prof. Nipun Batra, IIT Gandhinagar

The Journey We've Taken

Part 1: The Data Foundation

- Week 1-3: Collection (Scraping), Validation (Pydantic), Labeling (Label Studio).
- Week 4-5: Active Learning & Augmentation.
- *Key Takeaway*: Data is the most critical part of the pipeline.

Part 2: Model Engineering

- Week 6: LLM APIs.
- Week 7: Training & Fine-tuning (PEFT).
- Week 8: Reproducibility (Docker/MLflow).
- *Key Takeaway*: Models must be reproducible and trackable.

The Journey (Continued)

Part 3: Deployment & MLOps

- Week 9-10: Demos (Streamlit) & APIs (FastAPI).
- Week 11: CI/CD (GitHub Actions).
- Week 12-13: Optimization (ONNX/Quantization) & Profiling.
- Week 14: Monitoring (Drift).
- *Key Takeaway*: A model is not useful until it is served, reliable, and monitored.

The Full Stack AI Engineer

You now have the toolkit to build end-to-end systems:

1. **Scrape** data from the web.
2. **Validate** it rigorously.
3. **Train** a model (or fine-tune an LLM).
4. **Package** it in Docker.
5. **Serve** it via FastAPI.
6. **Deploy** it to the cloud.
7. **Monitor** it for drift.

Future Trends in AI Engineering

1. LLM Ops (LLOps)

- Managing prompts as code.
- Eval-driven development (RAGAS, TruLens).
- Vector Database management (Chroma/Pinecone) - *Check backup slides!*

2. AI Agents

- Systems that take action, not just chat.
- Tool use and planning (LangGraph) - *Check backup slides!*

3. Edge AI

- Running SLMs (Small Language Models) on phones/laptops.
- ExecuTorch, MLX (Apple Silicon).

Final Project Ideas

1. **Personal Health Assistant**: RAG app over your medical reports.
2. **Legal Document Analyzer**: Fine-tuned BERT for contract review.
3. **Smart Scraper**: Agent that navigates websites to find pricing.
4. **Edge Cam**: Object detection running on Raspberry Pi.

Staying Updated

- **Newsletters**: The Batch (DeepLearning.AI), Import AI.
- **Conferences**: NeurIPS (Datasets track), SysML.
- **GitHub**: Star the repos we used (FastAPI, Pydantic, Streamlit).

Thank You!

"The best way to predict the future is to invent it."

Keep building.