Building Scalable Multi Agent AI Systems

Mentors: Shubham Ingale, Atharv Kurde

Final Project: Build a RAG Agent

Tentative Plan:

Phase 1: Deep Learning (Weeks 1–4)

For the vibes (3b1b, Obviously!): Neural networks - YouTube

Resources till week3: Practical Deep Learning using PyTorch - YouTube

Also checkout this playlist, one of the best resources I have come across for LLM building from scratch (I have seen and coded all the content in the playlist personally : Neural Networks: Zero to Hero - YouTube (If you are wondering who Andrej Karpathy is, Google him! You are going to be amazed.)

There are 14 videos in this playlist. Try to complete all of them by week3. Make notes! Also code along with him. Only watching the videos won't do any good.

Week 1: Neural Networks & PyTorch Basics
✅ Train a simple MLP (Video 1-7)

Week 2: CNNs for Image Classification
✅ Build a CNN (Video 8-11)

Week 3: RNNs/LSTMs for Sequential Data
✅ Train a LSTM (Video 12-14)

Also read this blog - Understanding LSTM Networks -- colah's blog

Week 4: Transformers & Self-Attention (Andrej Karpathy's video :
Let's build GPT: from scratch, in code, spelled out.

You need to get done with the above mentioned video. For more exploration you can watch the videos from his playlist mentioned at the start of the document.
✅ Implement a basic self-attention layer
✅ Understand encoder-decoder architecture (GPT)

Phase 2: Hugging Face & LLM Deployment (Weeks 5–9)

Week 5: Hugging Face Basics
✅ Use `transformers` pipelines for text generation, summarization, classification

Getting Started With Hugging Face in 15 Minutes | Transformers, Pipeline, Tokenizer, Mod…

Introduction - Hugging Face LLM Course

Week 6: Running Models Locally with Ollama and RAG
✅ Set up Ollama and explore Groq (Both can we used to use open source models)
✅ Run LLaMA, Mistral, Phi, etc. locally and experiment with prompts

Running LLM Locally: A Beginner's Guide to Using Ollama | by Arun Patidar | Medium

ChatGroq | 🦜🔗 LangChain

Learn RAG From Scratch – Python AI Tutorial from a LangChain Engineer

Week 7: LangGraph - Building AI Agent (Finally!)
✅ Build ReAct (reasoning and acting) Agents (You can either use transformers pipelines or Ollama)

✅ Build RAG Agent

a-practical-guide-to-building-agents.pdf

LangGraph Complete Course for Beginners – Complex AI Agents with Python

Week 8: Model Optimization (Optional)
✅ Try quantized models (e.g., GGUF with Ollama)
✅ Experiment with ONNX or `optimum` for model speed-up

Week 9: Wrap-up & Documentation
✅ Write documentation on your workflow (GitHub)
✅ Share learnings and examples of local + Hugging Face usage