Building Your Own LLM from Scratch

Duration: 40 Hours

Module 1: Foundations of Large Language Models

- Evolution of NLP → From N-grams to Transformers
- What is an LLM? Key characteristics
- Understanding Tokens, Embeddings, Context Windows
- Transformer Architecture Deep Dive
- Self-Attention Mechanism
- Encoder vs Decoder vs Decoder-only models
- Overview of popular models (LLaMA, GPT, Mistral)

Module 2: Data Engineering for LLM Training

- Importance of data in LLMs
- Types of datasets (web text, code, domain-specific)
- Data collection strategies
- Data cleaning & deduplication
- Text preprocessing pipelines
- Tokenization strategies (BPE, SentencePiece)
- Dataset formatting for training

Module 3: Model Architecture & Design

- Designing a small LLM
- Choosing hyperparameters:
    - Layers, heads, hidden size
- Positional encoding techniques
- Activation functions (ReLU, GELU)
- Initialization strategies
- Memory optimization techniques

Module 4: Distributed Training & Infrastructure

- GPU architecture basics (A100, multi-GPU setups)
- Distributed training concepts:
    - Data Parallelism
    - Model Parallelism
    - Pipeline Parallelism
- Frameworks:
    - PyTorch Distributed
    - DeepSpeed
    - FSDP (Fully Sharded Data Parallel)
- Mixed precision training (FP16/BF16)
- Checkpointing & fault tolerance

Module 5: Training the LLM

- Training objectives (causal language modeling)
- Loss functions & optimization
- Learning rate scheduling
- Gradient clipping
- Monitoring training (loss curves, perplexity)
- Debugging training issues
- Cost optimization strategies

Module 6: Evaluation & Fine-Tuning

Topics:

- Evaluation metrics
- Benchmark datasets
- Fine-tuning techniques:
    - Full fine-tuning
    - LoRA
- Prompt-based evaluation

Module 7: Deployment & Optimization

- Model compression
- Inference optimization
- Serving LLMs via APIs
- Integration with applications
- Safety, bias, and guardrails
