

Deploying Small Language Models (LFWS307)

Course Duration: 24 Hours

TOC

Course Introduction

Hugging Face Model Ecosystem

Lab 2.1. Download Phi-3-mini and Qwen2.5-1.5B. Compare model cards, licenses, and file sizes. Convert safetensors to GGUF.

Llamafire: Zero-Dependency Deployment

Lab 3.1. Create llamafire from Phi-3-mini GGUF. Test CLI completion and HTTP API. Benchmark tokens/sec on CPU vs GPU

Quantization with llama.cpp

Lab 4.1. Quantize Qwen2.5-1.5B to Q4/Q5/Q8. Benchmark size, speed, and perplexity. Select optimal quantization for 8GB RAM target.

Llamafire HTTP Serving

Lab 5.1. Deploy llamafire server. Build Python/curl client. Test streaming completions. Load test with 10 concurrent users.

Production Serving with Batuta

Lab 6.1. Build Batuta serving pipeline. Compare latency vs llamafire. Achieve <100ms p99 with continuous batching.

RAG with Patcha + Hugging Face Embeddings

Lab 7.1. Index 1000 docs using all-MiniLM-L6-v2 embeddings. Build RAG pipeline with Phi-3. Compare RAG vs pure generation accuracy.

Edge Deployment

Lab 8.1. Deploy Q4 quantized model to ARM device (or emulator). Achieve interactive inference with 4GB RAM constraint.

Browser Deployment with Presentar

Lab 9.1. Deploy Phi-3 Q4 to browser via Presentar. Achieve <500ms first-token latency. Build chat interface with streaming.

Monitoring with Entrenar

Kubernetes Deployment

Capstone: Multi-Target Deployment

Course Summary