Topic 10: Serving and Deployment of AI Systems

1. Ollama: An open-source toolkit that makes running large language models locally easy: https://github.com/ollama/ollama
2. vLLM: A production-grade inference engine built for throughput and memory efficiency: https://docs.vllm.ai/en/stable/getting_started/quickstart.html
3. NVIDIA TensorRT: NVIDIA’s toolkit for accelerating large language model inference on GPUs with quantization and streaming: https://github.com/NVIDIA/TensorRT
4. Text Generation Inference: A robust, open-source server for deploying LLMs at scale: https://huggingface.co/docs/inference-endpoints/en/engines/tgi
5. Llama.cpp: Portable C++ implementation for running quantized LLMs efficiently on CPUs or edge devices: https://github.com/ggml-org/llama.cpp
6. Ray Serve: A Python-native serving framework from Ray for scalable model APIs. Supports dynamic batching, streaming, and distributed inference for LLMs: https://docs.ray.io/en/latest/serve/index.html
7. Deploying LLMs on Kubernetes: A practical guide to deploying, scaling, and optimizing LLMs on Kubernetes clusters with GPUs: https://innitor.ai/blog/best-practices-for-deploying-open-source-llms-on-kubernetes-with-nvidia-gpus