Scaling AI Deployment: Techniques and Tools

Overview

Deploying large-scale AI systems requires managing infrastructure, latency, and cost while maintaining model quality. Scaling involves parallelization and monitoring.

Deployment Strategies

• Containerization with Docker and Kubernetes • Model serving with TensorRT, ONNX Runtime, or FastAPI • Autoscaling with cloud orchestration tools

Performance Optimization

• Quantization and pruning • Caching responses for repeated queries • Using GPUs/TPUs efficiently

Monitoring and Maintenance

• Track inference latency and errors • Automate retraining using CI/CD pipelines • Regular validation with test datasets