

## Performance Optimization

Optimize model inference speed and resource efficiency

### Model Quantization

INT8 inference for 4x speedup

### Batching

Process multiple requests together

### Caching

Cache frequent predictions or features

### GPU Optimization

Use TensorRT, ONNX Runtime

### Profiling

Identify bottlenecks in pipeline

### Tools & Platforms

NVIDIA TensorRT

ONNX Runtime

PyTorch JIT

TF Lite