Efficient Large Language Model Inference

