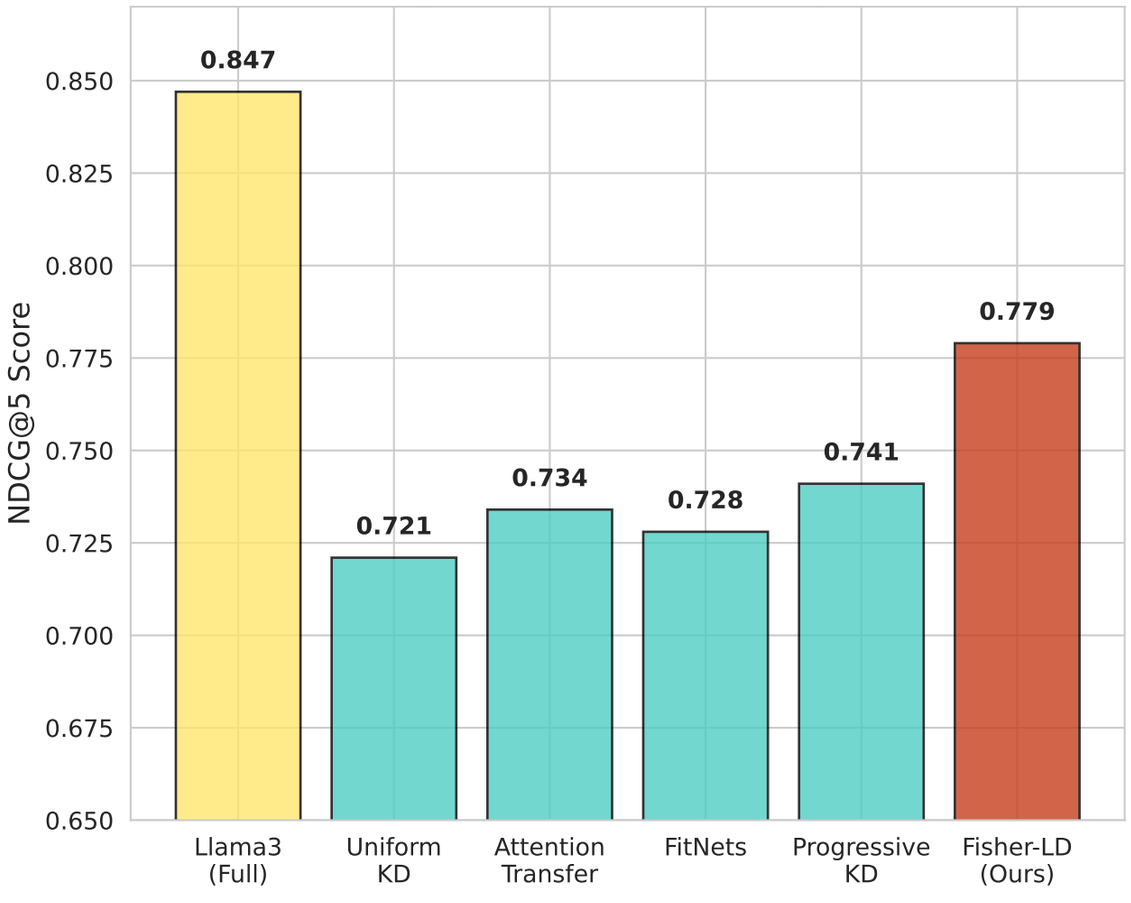
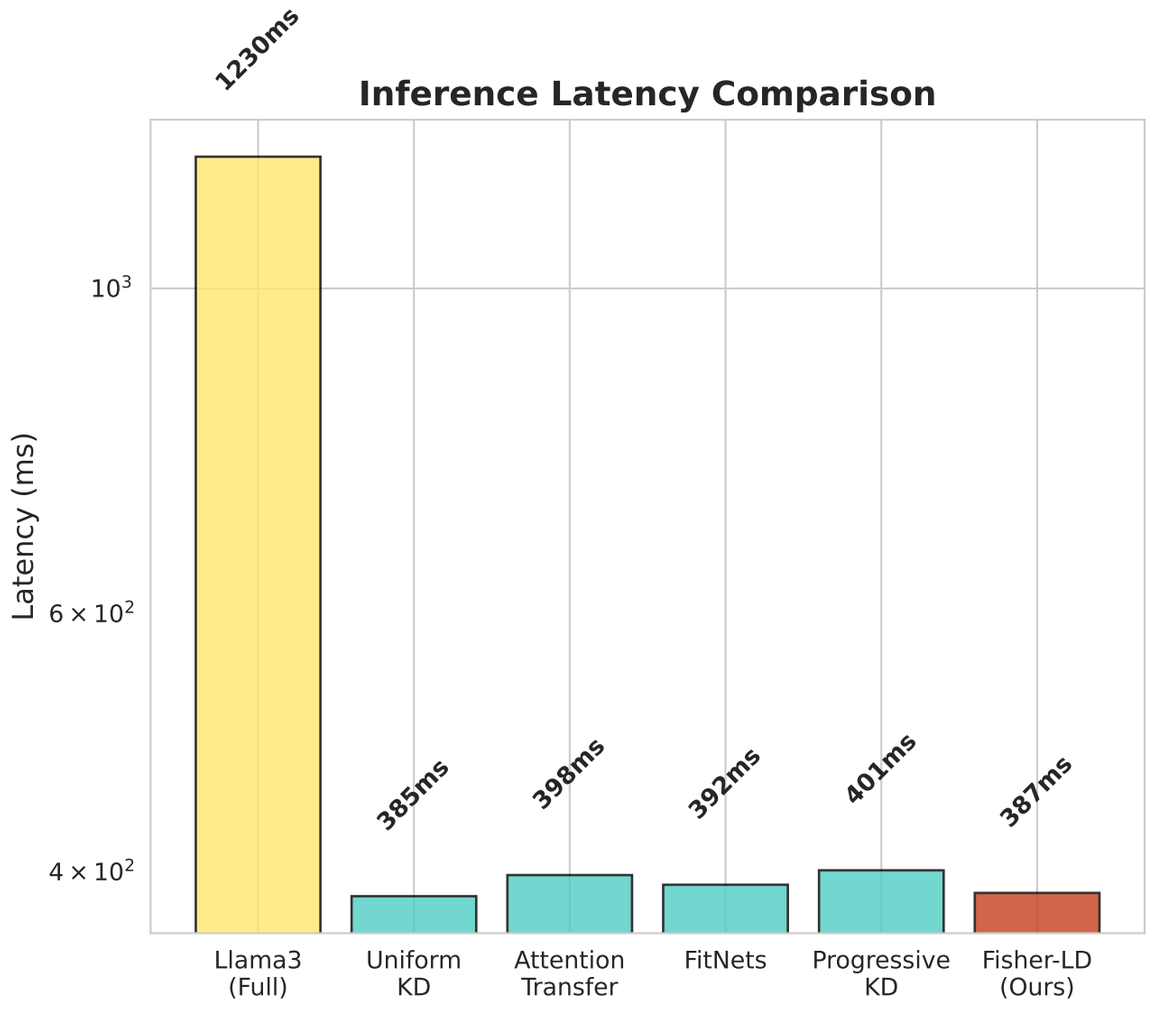
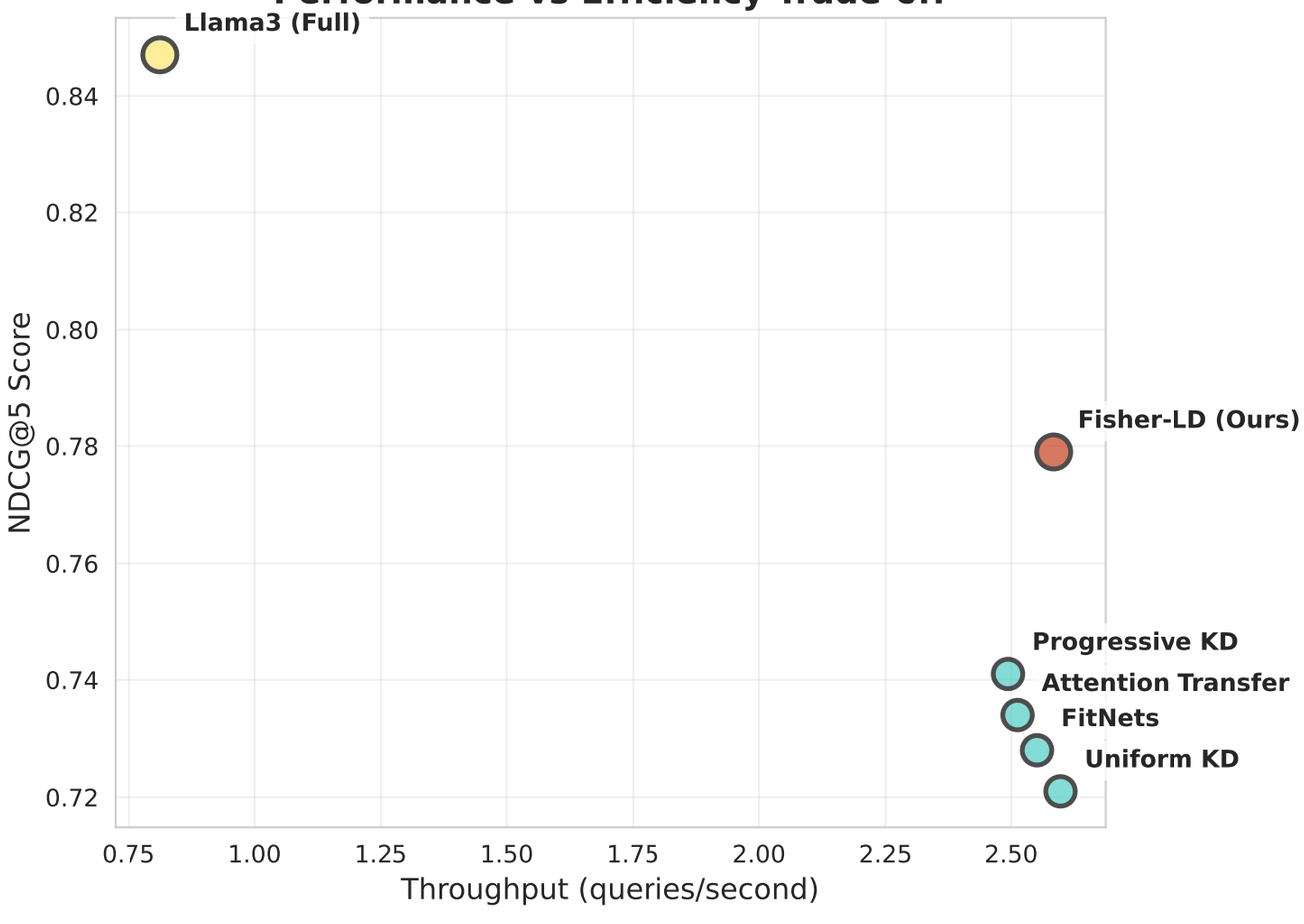
NDCG@5 Performance Comparison



Inference Latency Comparison



Performance vs Efficiency Trade-off



Model Compression vs Quality Retention