

Inference Latency vs. Network Bandwidth (DistillBERT-base)

