

Experimental Setup: Llama 3.2-1B Quantization Study



HuggingFace Hub
meta-llama/Llama-3.2-1B

Llama 3.2-1B
1B params • RoPE • SwiGLU

Quantization

FP16
Baseline

NF4
4-bit

FP4
4-bit

Im-evaluation-harness
Gao et al., 2023

CoQA Benchmark
50 samples • Zero-shot

Results: F1 Score, Memory, Latency, Throughput

Modal Cloud
Serverless GPUs

NVIDIA A100
40GB • CUDA 12.x

PyTorch 2.1+
Transformers 4.36+

BitsAndBytes
4-bit Quantization