LLaMA-70B Inference Throughput (Groq LPU)

Throughput (tokens/sec) vs Quantization Precision

- FP16 (CPU): 12 tok/s
- INT8: 1450 tok/s
- INT4 (AWQ): 3124 tok/s
- 3.5-bit (Ours): 4188 tok/s (+34.1%)