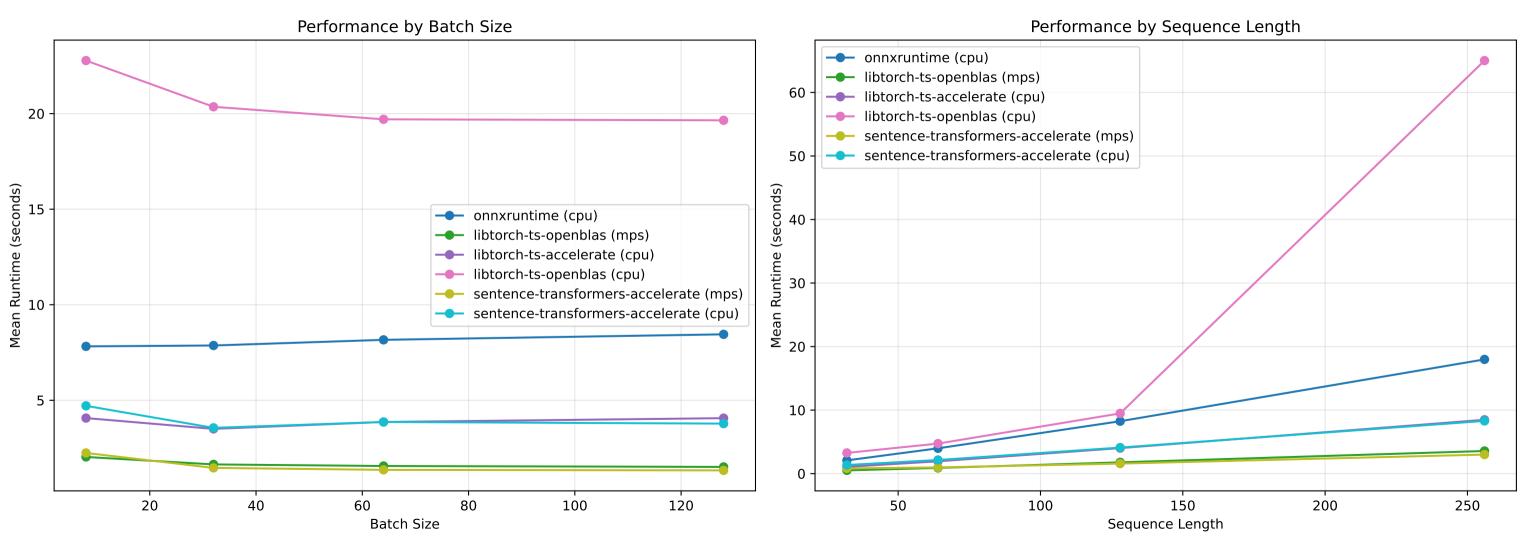
Benchmark Performance Overview



Batch Size Performance by Contestant (by Sequence Length) Seq Length = 32Seq Length = 64Seq Length = 128onnxruntime (cpu) onnxruntime (cpu) 6 libtorch-ts-openblas (mps) libtorch-ts-openblas (mps) 6 libtorch-ts-accelerate (cpu) libtorch-ts-accelerate (cpu) libtorch-ts-openblas (cpu) libtorch-ts-openblas (cpu) 10 sentence-transformers-accelerate (mps) sentence-transformers-accelerate (mps) 5 sentence-transformers-accelerate (cpu) sentence-transformers-accelerate (cpu) 5 Mean Runtime (seconds) Mean Runtime (seconds) Mean Runtime (seconds) onnxruntime (cpu) libtorch-ts-openblas (mps) libtorch-ts-accelerate (cpu) libtorch-ts-openblas (cpu) sentence-transformers-accelerate (mps) sentence-transformers-accelerate (cpu) 2 2 1 80 20 40 60 100 120 20 40 60 80 100 120 20 40 60 80 100 120 Batch Size **Batch Size** Batch Size Seq Length = 25670 60 50 Mean Runtime (seconds) onnxruntime (cpu) libtorch-ts-openblas (mps)

libtorch-ts-accelerate (cpu) libtorch-ts-openblas (cpu)

80

20

10

20

40

60

Batch Size

sentence-transformers-accelerate (mps) sentence-transformers-accelerate (cpu)

100

120