

Sparse vs Full Attention Performance Comparison

Sequence Length: 64,000, Heads: 4, Sparsity: 90%

