Memory and Computation Scaling with Context Length

Crossover Point
(~1024 tokens)

16×

4×

2×

4×

1×

1×

Attention (O(n²))

Feed-Forward (O(n))

Relative Computation

16×

12×

8×

4×

2×

0×

0   512   1024   2048   3072   4096

Feed-forward
dominates

Attention dominates
(Memory bottleneck)

Context Length (tokens)

Quadratic scaling makes context length the primary cost driver for long sequences