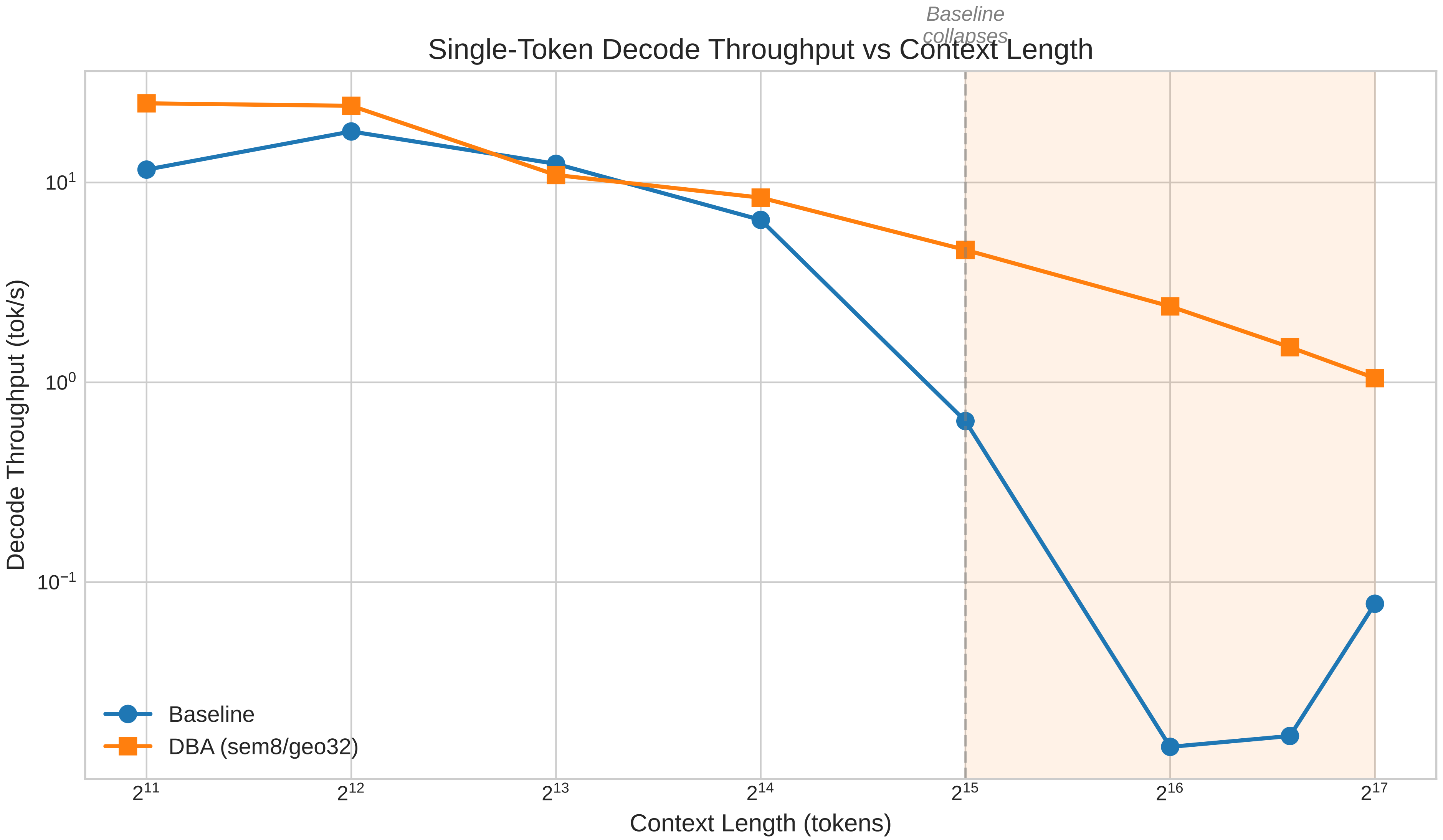
Baseline collapses
Single-Token Decode Throughput vs Context Length
Decode Throughput (tok/s)
10^1
10^0
10^{-1}
2^{11}
2^{12}
2^{13}
2^{14}
2^{15}
2^{16}
2^{17}
Context Length (tokens)
Baseline
DBA (sem8/geo32)