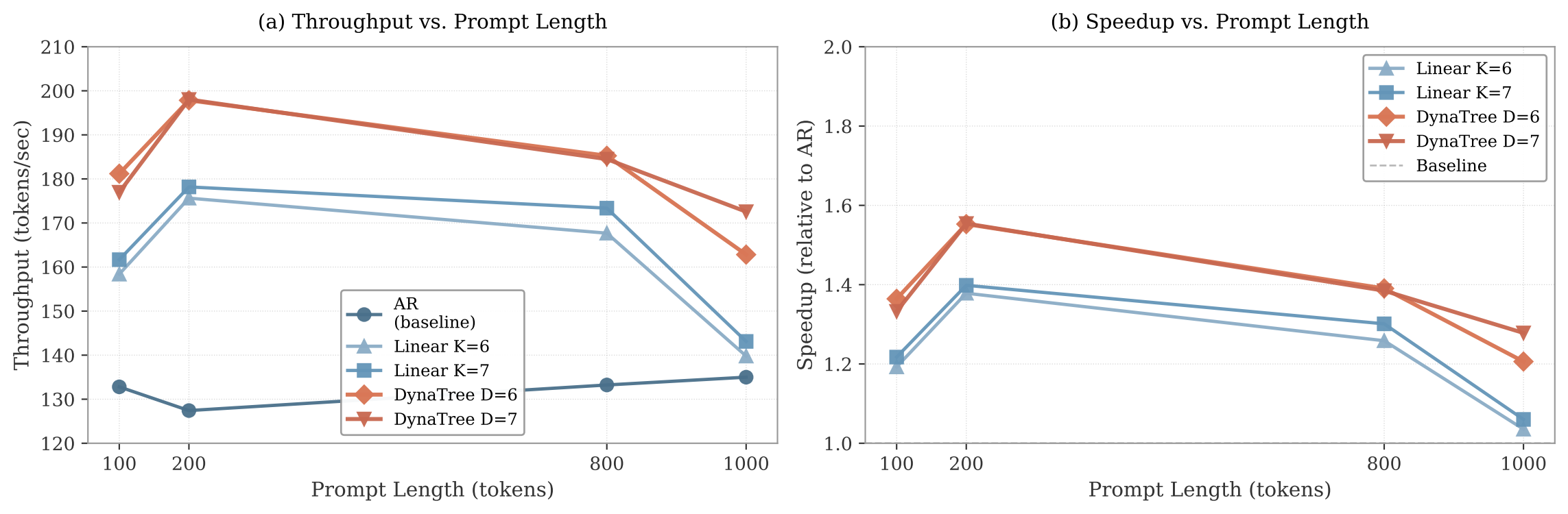
(a) Throughput vs. Prompt Length
Throughput (tokens/sec)
Prompt Length (tokens)
AR (baseline)
Linear K=6
Linear K=7
DynaTree D=6
DynaTree D=7
(b) Speedup vs. Prompt Length
Speedup (relative to AR)
Prompt Length (tokens)
Linear K=6
Linear K=7
DynaTree D=6
DynaTree D=7
Baseline