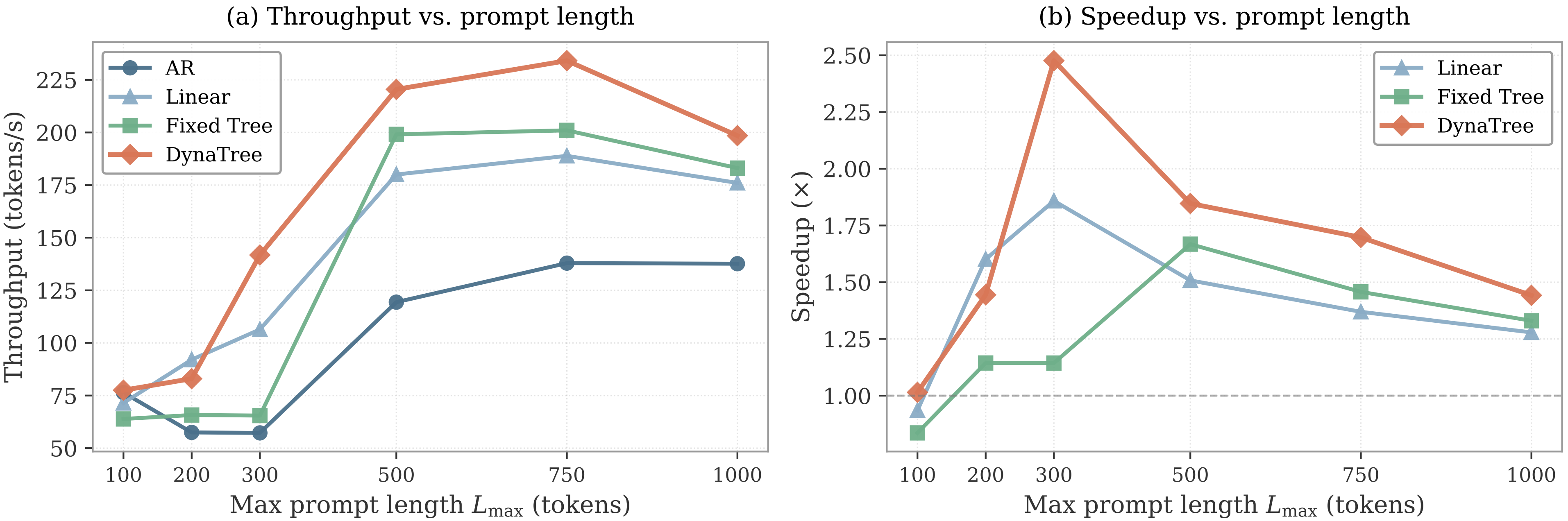
(a) Throughput vs. prompt length
Throughput (tokens/s)
225
200
175
150
125
100
75
50
100
200
300
500
750
1000
Max prompt length $L_{\max}$ (tokens)
AR
Linear
Fixed Tree
DynaTree
(b) Speedup vs. prompt length
Speedup (×)
2.50
2.25
2.00
1.75
1.50
1.25
1.00
100
200
300
500
750
1000
Max prompt length $L_{\max}$ (tokens)
Linear
Fixed Tree
DynaTree