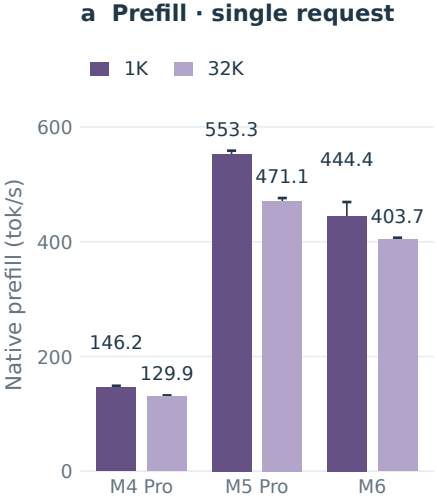
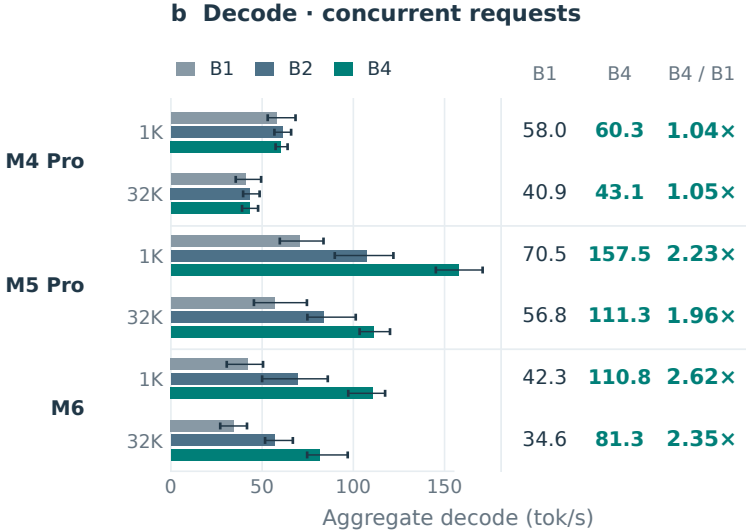

a Prefill · single request
1K
32K
600
400
200
0
Native prefill (tok/s)
553.3
471.1
444.4
403.7
146.2
129.9
M4 Pro
M5 Pro
M6


b Decode · concurrent requests
B1
B2
B4
B1
B4
B4 / B1
M4 Pro
1K
58.0
60.3
1.04×
32K
40.9
43.1
1.05×
M5 Pro
1K
70.5
157.5
2.23×
32K
56.8
111.3
1.96×
M6
1K
42.3
110.8
2.62×
32K
34.6
81.3
2.35×
0
50
100
150
Aggregate decode (tok/s)