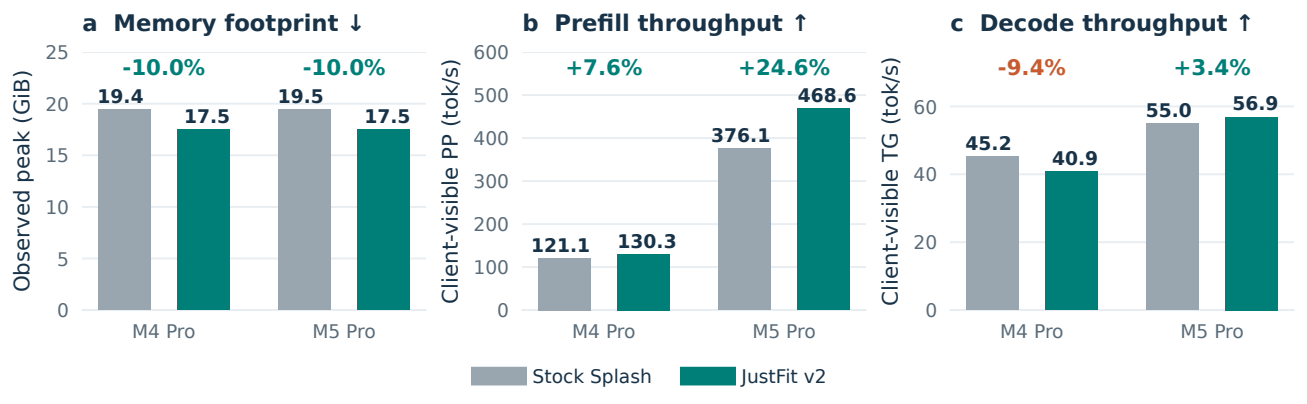

a Memory footprint ↓
Observed peak (GiB)
25
20
15
10
5
0
-10.0%
-10.0%
19.4
17.5
19.5
17.5
M4 Pro
M5 Pro
b Prefill throughput ↑
Client-visible PP (tok/s)
600
500
400
300
200
100
0
+7.6%
+24.6%
121.1
130.3
376.1
468.6
M4 Pro
M5 Pro
c Decode throughput ↑
Client-visible TG (tok/s)
60
40
20
0
-9.4%
+3.4%
45.2
40.9
55.0
56.9
M4 Pro
M5 Pro
Stock Splash
JustFit v2