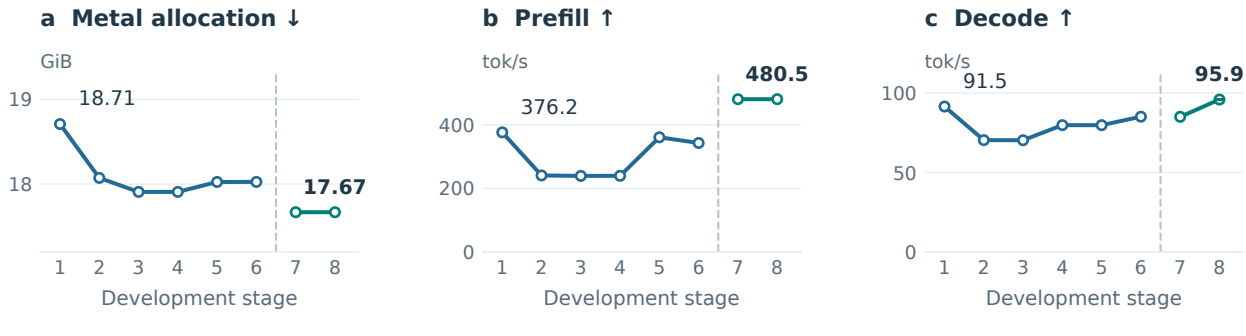

a Metal allocation ↓
GiB
19
18
18.71
17.67
1 2 3 4 5 6 7 8
Development stage
b Prefill ↑
tok/s
400
200
0
376.2
480.5
1 2 3 4 5 6 7 8
Development stage
c Decode ↑
tok/s
100
50
0
91.5
95.9
1 2 3 4 5 6 7 8
Development stage