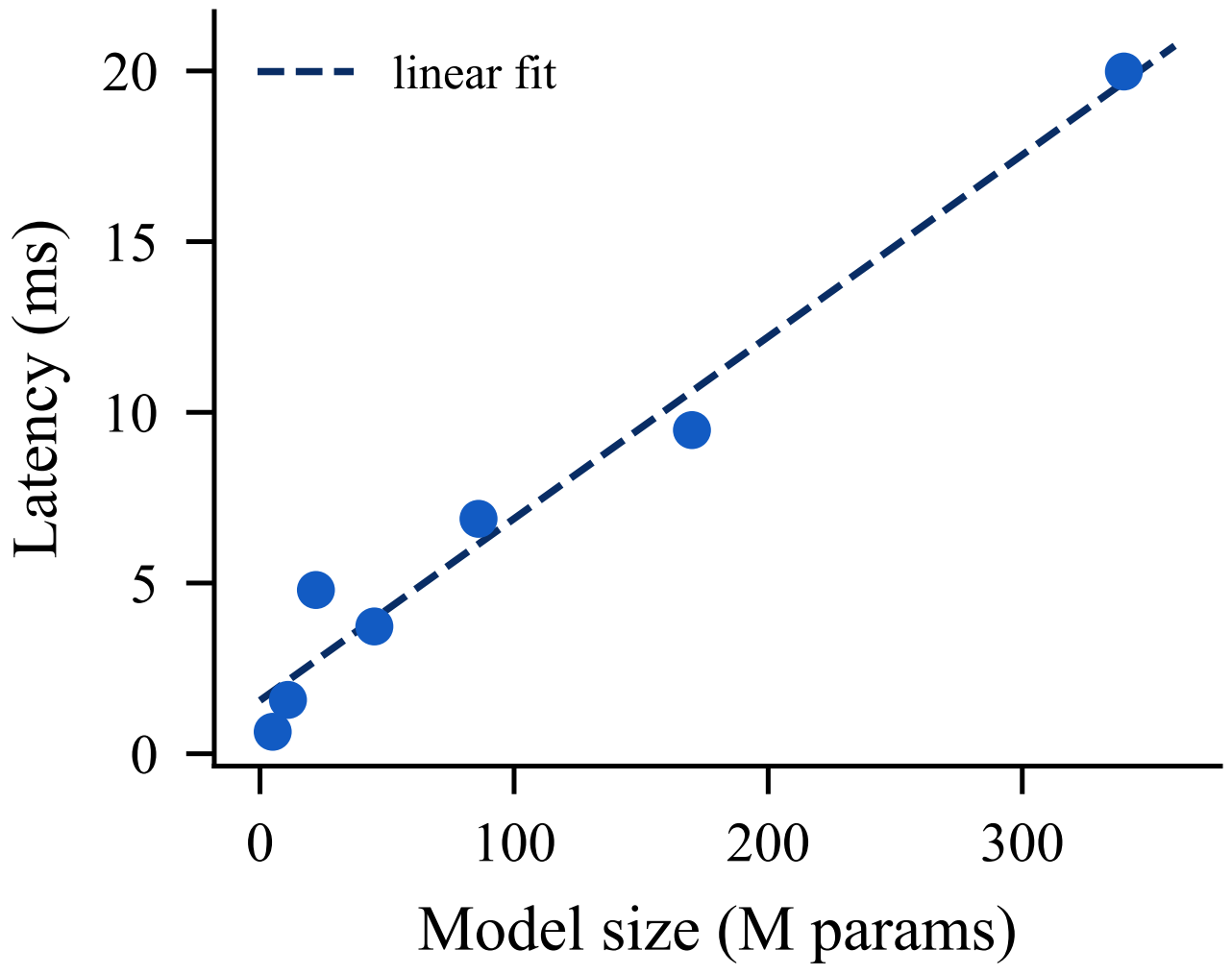

linear fit
20
15
10
5
0
Latency (ms)
0
100
200
300
Model size (M params)