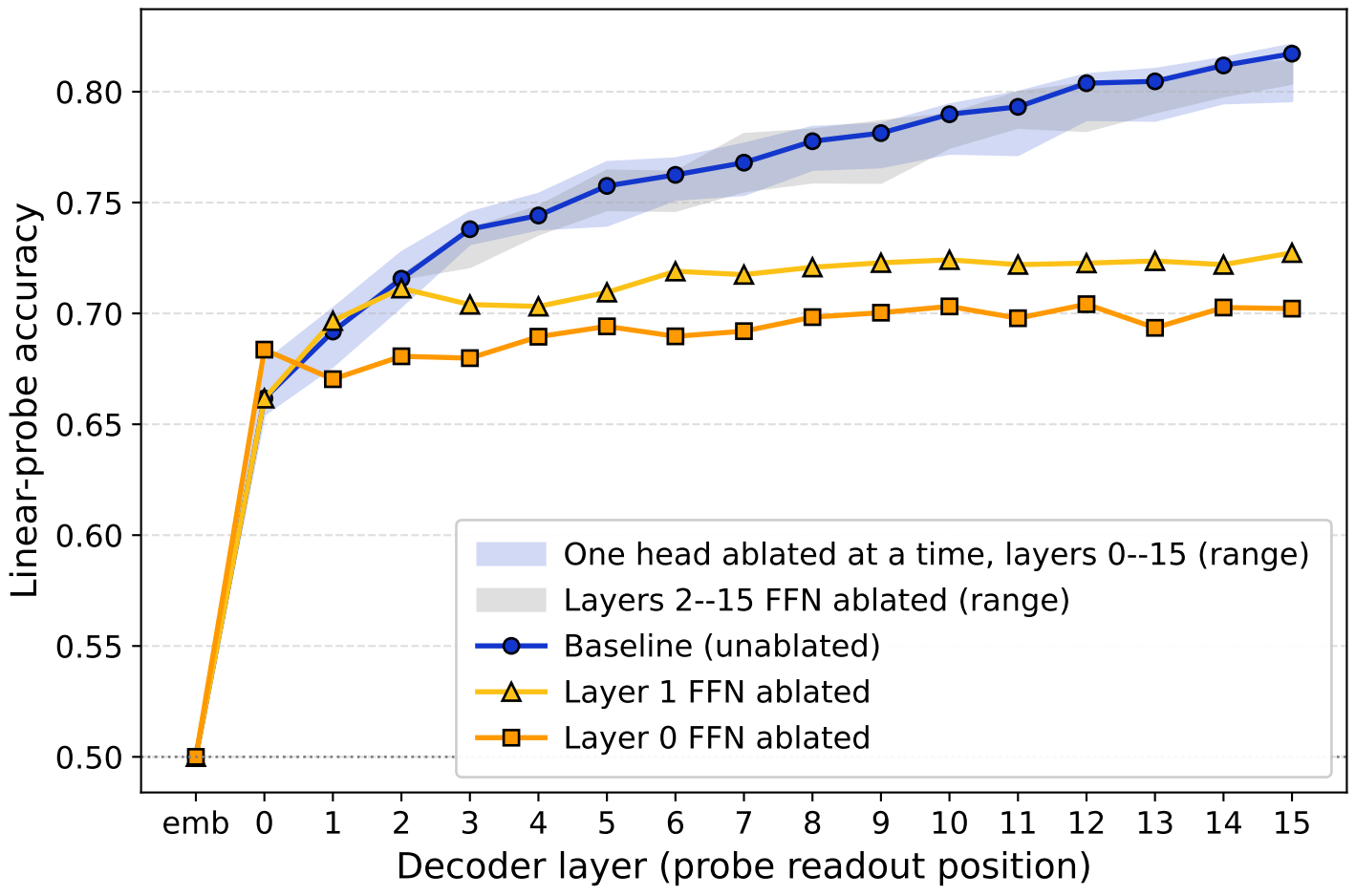

Linear-probe accuracy
0.80
0.75
0.70
0.65
0.60
0.55
0.50
emb
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Decoder layer (probe readout position)
One head ablated at a time, layers 0--15 (range)
Layers 2--15 FFN ablated (range)
Baseline (unablated)
Layer 1 FFN ablated
Layer 0 FFN ablated