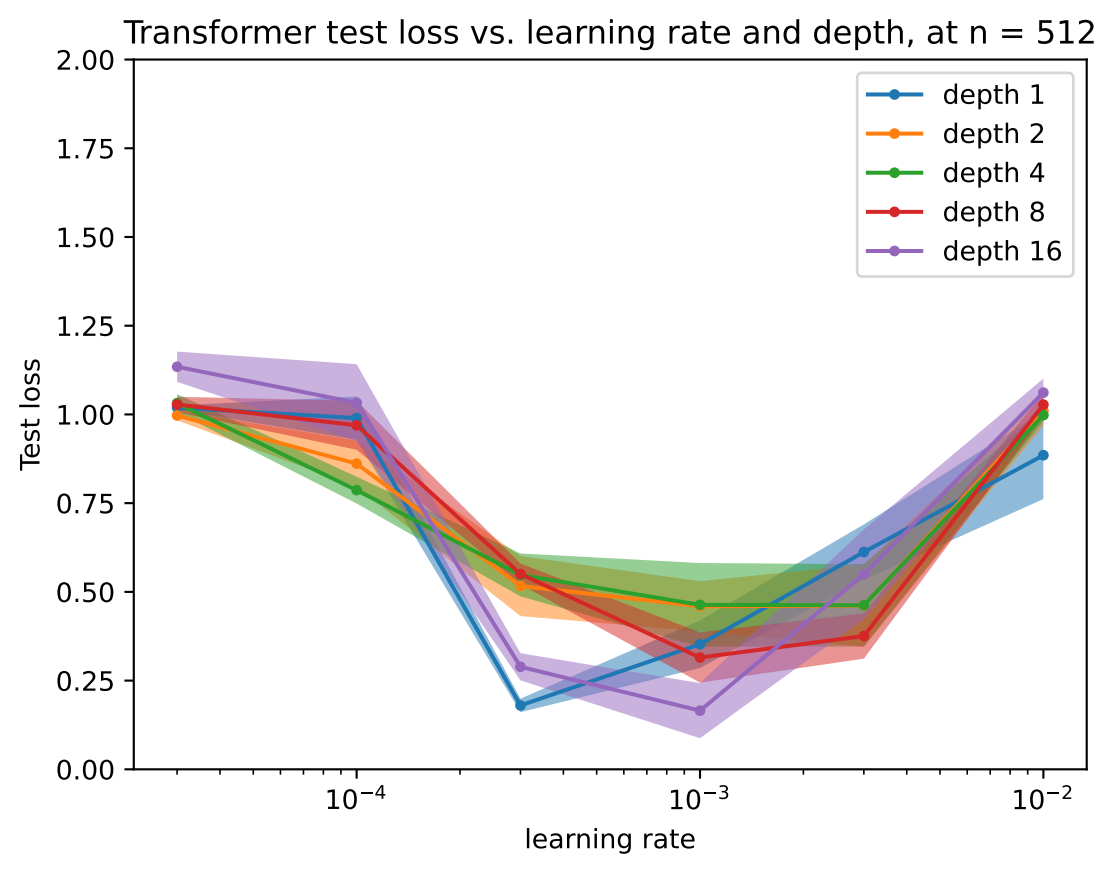
Transformer test loss vs. learning rate and depth, at n = 512
depth 1
depth 2
depth 4
depth 8
depth 16
Test loss
learning rate
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
10^{-4}
10^{-3}
10^{-2}