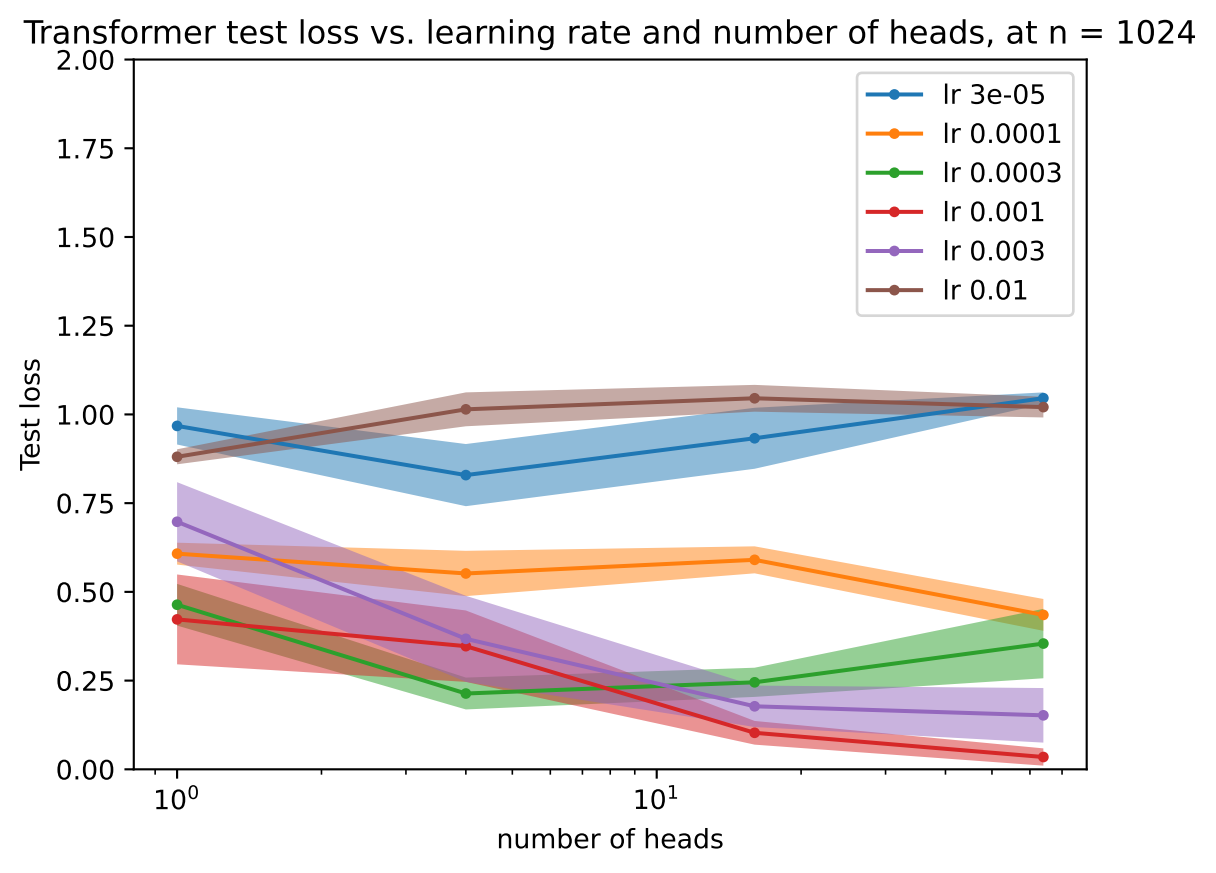
Transformer test loss vs. learning rate and number of heads, at n = 1024
Test loss
number of heads
lr 3e-05
lr 0.0001
lr 0.0003
lr 0.001
lr 0.003
lr 0.01
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
10^0
10^1