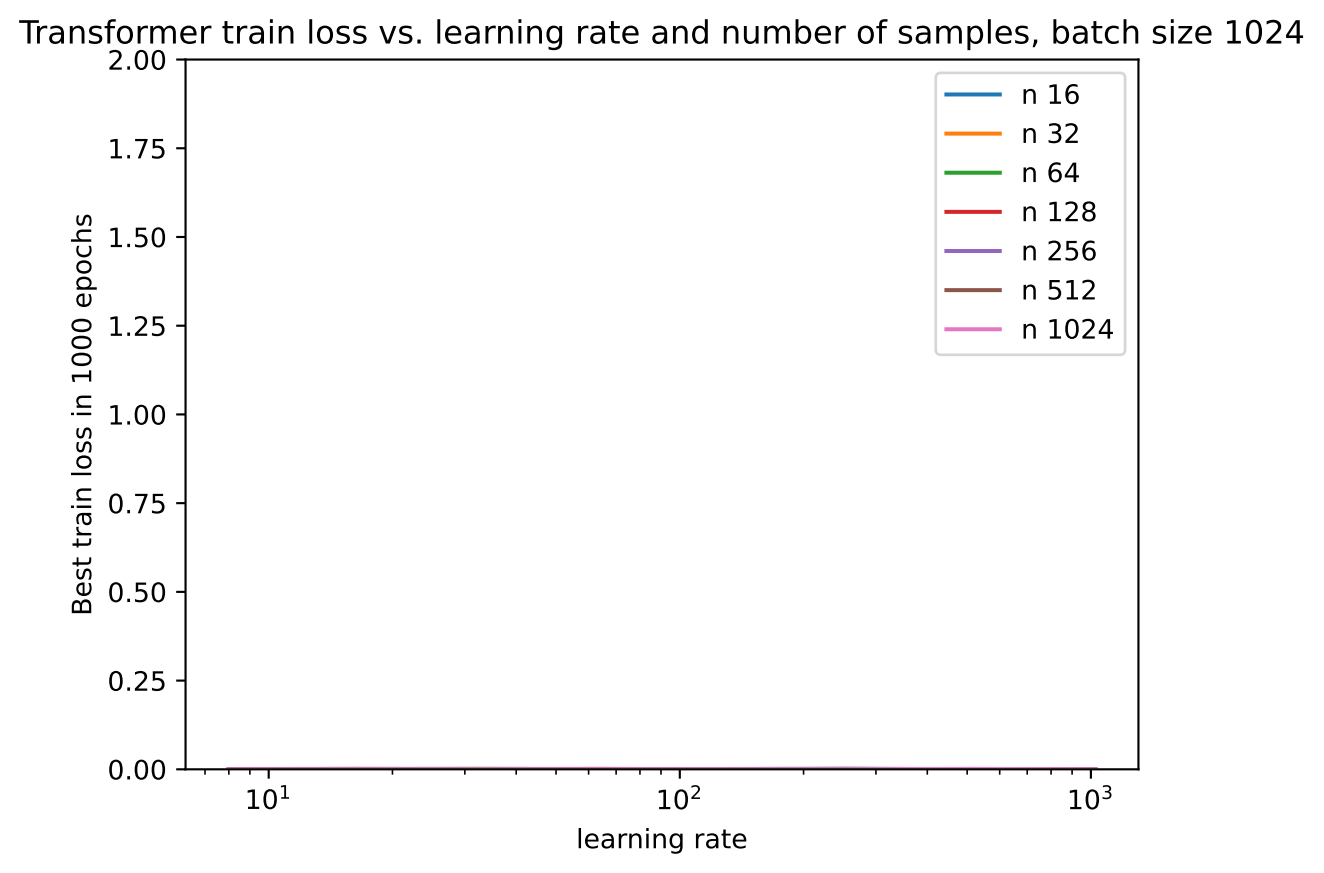
Transformer train loss vs. learning rate and number of samples, batch size 1024
Best train loss in 1000 epochs
learning rate
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
10^1
10^2
10^3
n 16
n 32
n 64
n 128
n 256
n 512
n 1024