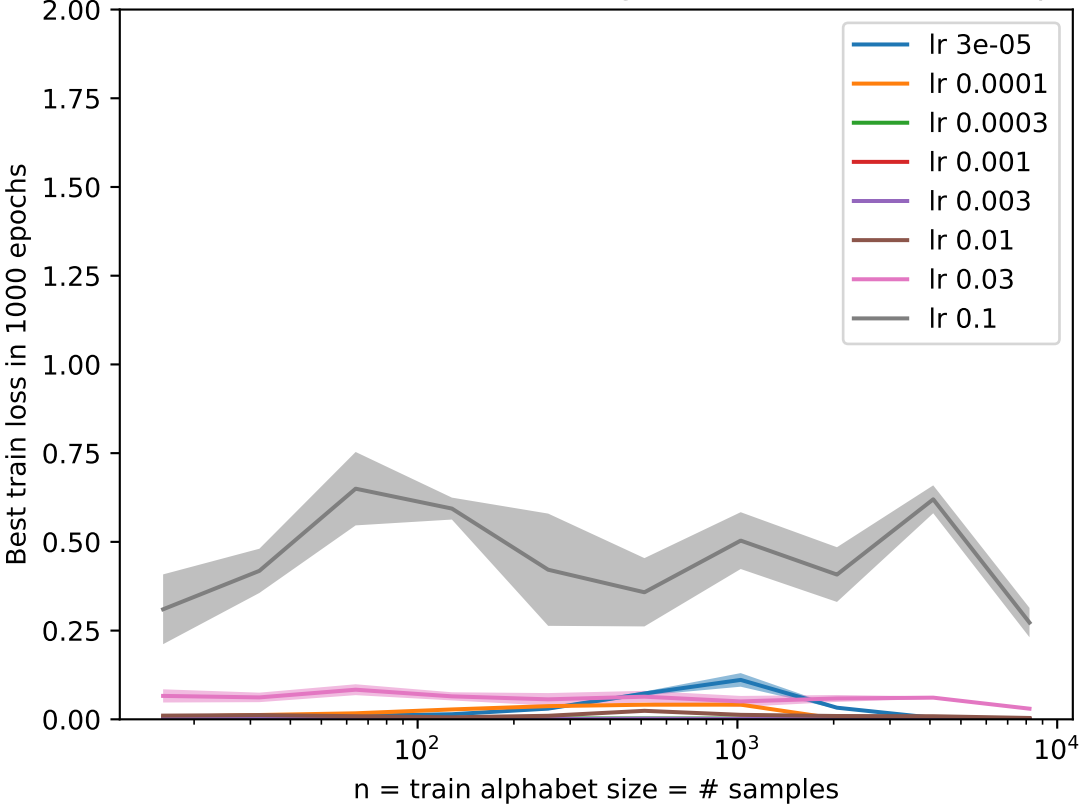
Transformer train loss vs. learning rate and number of samples
Best train loss in 1000 epochs
n = train alphabet size = # samples
lr 3e-05
lr 0.0001
lr 0.0003
lr 0.001
lr 0.003
lr 0.01
lr 0.03
lr 0.1
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
10^2
10^3
10^4