

Transformer train loss vs. learning rate and depth, at n = 1024

