

Transformer train loss vs. learning rate and number of heads, at $n = 1024$

