

Transformer test loss vs. learning rate and number of heads, at $n = 512$

