High score: 0.005

Impact of restoring MLP after corrupted input

single restored MLP layer within gpt2

p(worsened)

Low score: 0.004