High score: 0.042

Impact of restoring MLP after corrupted input

p(evacuated)

Low score: 0.021

single restored MLP layer within gpt2