Impact of restoring MLP after corrupted input

High score: 0.498

Low score: 0.451

p(on)

single restored MLP layer within gpt2