MLP Layers Effect (Noising)
gpt2 - Letter-String Analogy Task ('+1' vs No Rule) 0.0 Average Effect on Logit Difference -0.1 -0.2 -0.3-0.4

Layer

10

11