MLP Layers Effect (Noising) gptj6b - Letter-String Analogy Task ('+1' vs No Rule) 0.5 0.0 Average Effect on Logit Difference -0.5 -1.0-1.5-2.0-2.510 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27

Layer