Llama-3.3-70B-Instruct layer 0 layer 1 layer 2 layer 3 0.6 0.4 0.04 0.2 0.4 0.02 0.2 0.1 0.2 0 0 -0.02-0.1-0.05 0 0.3 0.2 0.10 -0.4 -0.20.05 0.1 -0.4-0.20 0.2 0 0.2 layer 5 layer 4 layer 6 layer 7 0.5 0.5 0.2 0 -0.2-0.5-0.5-0.4-0.5~0.2° 0.2° 0.4 0.4 0.20 -0.5-0.50 0.5 0.5 0 layer 8 layer 9 layer 10 layer 11 0.5 0.5 0.5 0.5 0 0 -0.5 -0.5-0.5-1-0.5-0.50.5 -0.50 0.5 -0.50 0.5 -1 -0.5 0 0.50 layer 14 layer 13 layer 15 layer 12 1 0.5 0.5 -0.5-0.5 -1 $^{-1}$ -1-1-10 1 -10 1 2 -10 1 layer 16 layer 17 layer 18 layer 19 1 layer 22 layer 20 layer 21 layer 23 1 -1-2 -3 -2 -10 $-2 \ -1 \ 0$ 2 layer 27 layer 25 layer 26 layer 24 2 0 -2 0 2 0 2 2 -2 0 -2-2 layer 30 layer 28 layer 29 layer 31 2 -4 -2 0 -2 0 2 -4 -2 02 -4 -2 0layer 35 layer 34 layer 32 layer 33 -4 -5 -4 -2 0 2 4 0 5 layer 38 layer 36 layer 37 layer 39 -5 -5 layer 40 layer 41 layer 42 layer 43 2 -2**-**5 0 -5 0 5 -5 -5 5 layer 46 layer 44 layer 45 layer 47 -5 -5 5 -5 5 -5 0 5 -5 0 5 0 0 layer 48 layer 50 layer 49 layer 51 -5 -5 -5 5 10 **-**5 0 5 10 **-**5 0 5 -5 5 0 10 0 10 layer 52 layer 53 layer 54 layer 55 5 0 -5 -5 -5 0 5 -5 0 5 10 -5 0 5 10 10 -5 10 layer 56 layer 57 layer 58 layer 59 -10 -10 -10 -5 0 5 10 -10 -10 -5 0 5 10-10 -5 0 5 10-10 -5 0 5 10 layer 60 layer 61 layer 62 layer 63 0 -10 -10 -10 -10 10 -10 -10-10-1010 10 10 layer 65 layer 67 layer 66 layer 64 0 -10 -10 -10 -10 -10 -10 10 0 10 -10 0 -100 10 0 10 layer 68 layer 69 layer 70 layer 71 -10 -10 -10 -10-15 -10 0 10 -15 -10 0 10 -10 0 10 -10 0 10 layer 74 layer 73 layer 75 layer 72 10 -10 -10 -10 -10 -10 0 10 20 -10 0 -10 0 1010 20 -20 -10 0 10 20 layer 78 layer 79 layer 76 layer 77 20 20 10 10 10 -10-10-20 -20 -10 0 10 20 10 20 -200 20 0 20

honest_no

honest_yes

lying_no

lying_yes

Loading [MathJax]/extensions/MathMenu.js