llama2-13b-hf Average Between-Subject Alignment 0.0050 norm plato randtok 0.0045 0.0040 0.0035 0.0030 0.0025 0.0020 5 10 20 0 15 25 30 35 40 Layer Index