GPT-J Steered Accuracy for Layer 15 uncentred mean-centred 8.0 0.7 Average Accuracy
0.0 0.0 0.0
0.0 5.0 0.0 0.2 0.1 0.0 antonym capitalize present-past singular-plural country-capital english-french Task