

Social IQa, qwen-2.5-7b to Llama-3.1-70B

Accuracy

0.81
0.80
0.79
0.78
0.77
0.76

0.00

0.25

0.50

0.75

1.00

Routing Ratio

- average-token-prob
- verbalization-1s
- verbalization-2s
- p(true)
- trained-probe
- perplexity
- jaccard-degree
- ood-probe

