

average-token-prob

verbalization-1s

verbalization-2s

p(true)

trained-probe

perplexity

jaccard-degree

ood-probe

Routing Ratio

