

MultiArith, phi-3.5-mini to GPT-4.1 mini



average-token-prob

verbalization-1s

verbalization-2s

$p(\text{true})$

trained-probe

perplexity

jaccard-degree

ood-probe