

SVAMP, mistral-7b to GPT-4.1 mini



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