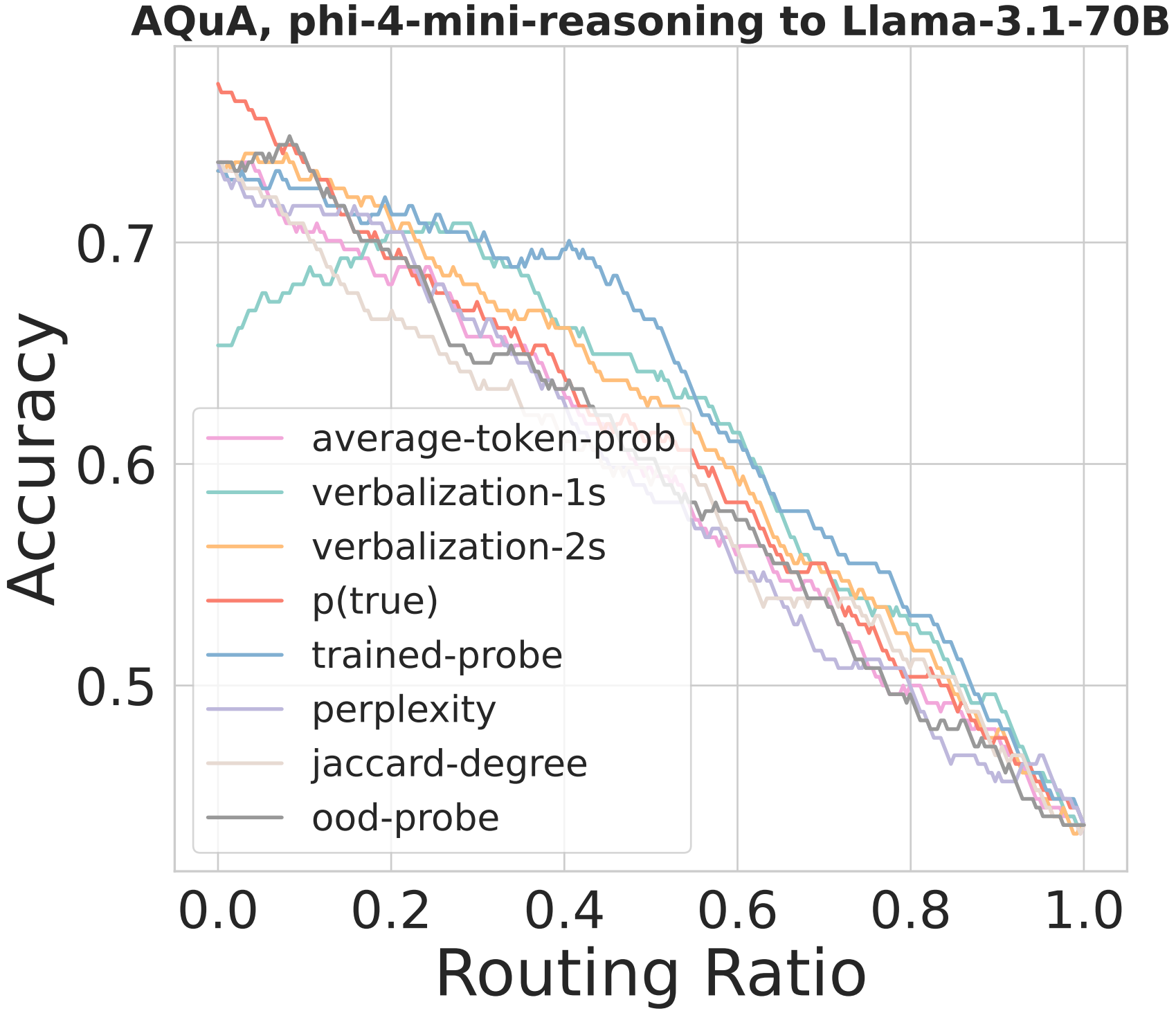
AQuA, phi-4-mini-reasoning to Llama-3.1-70B
Accuracy
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Routing Ratio
average-token-prob
verbalization-1s
verbalization-2s
p(true)
trained-probe
perplexity
jaccard-degree
ood-probe