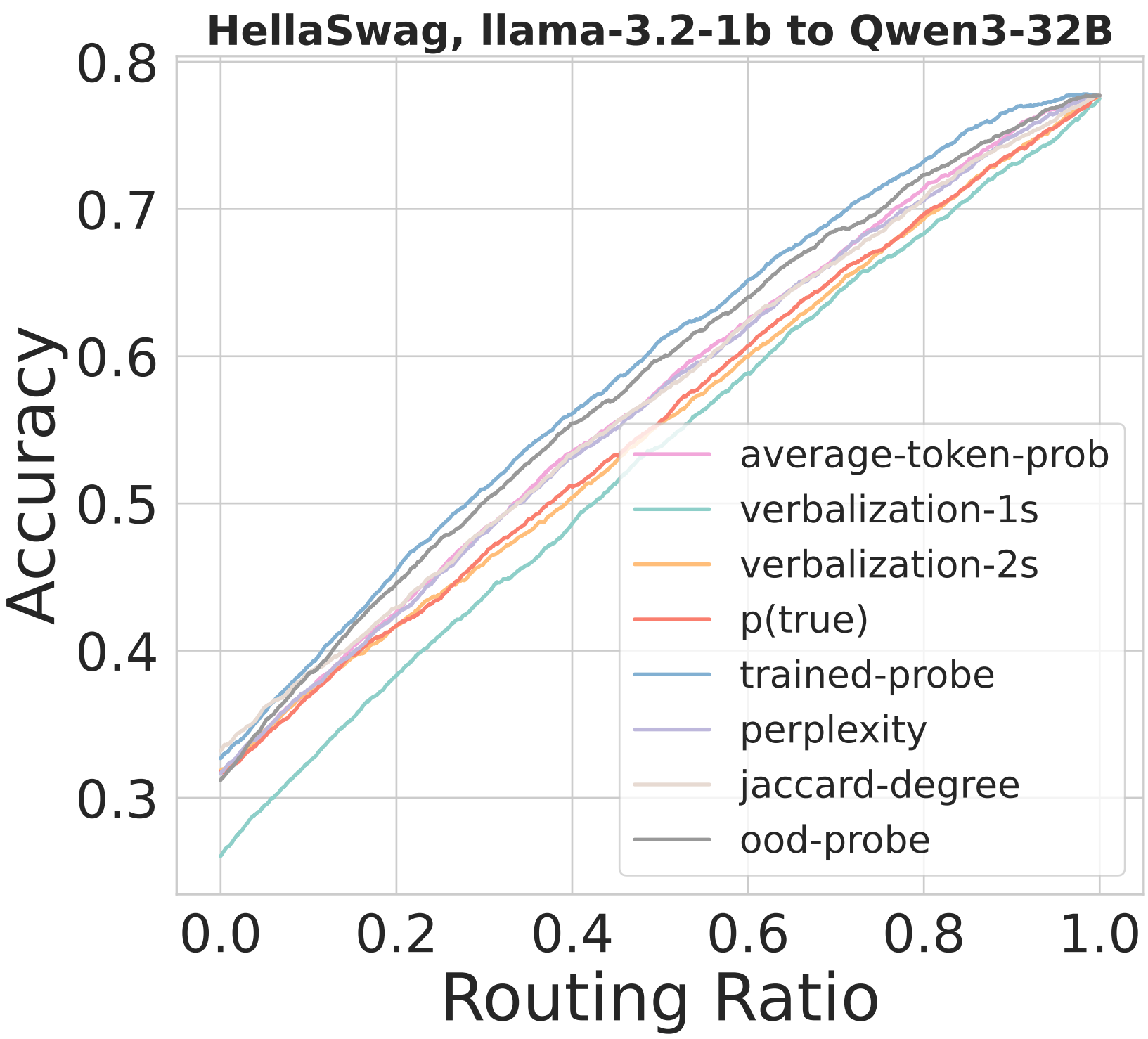
HellaSwag, llama-3.2-1b to Qwen3-32B
Accuracy
Routing Ratio
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0
average-token-prob
verbalization-1s
verbalization-2s
p(true)
trained-probe
perplexity
jaccard-degree
ood-probe