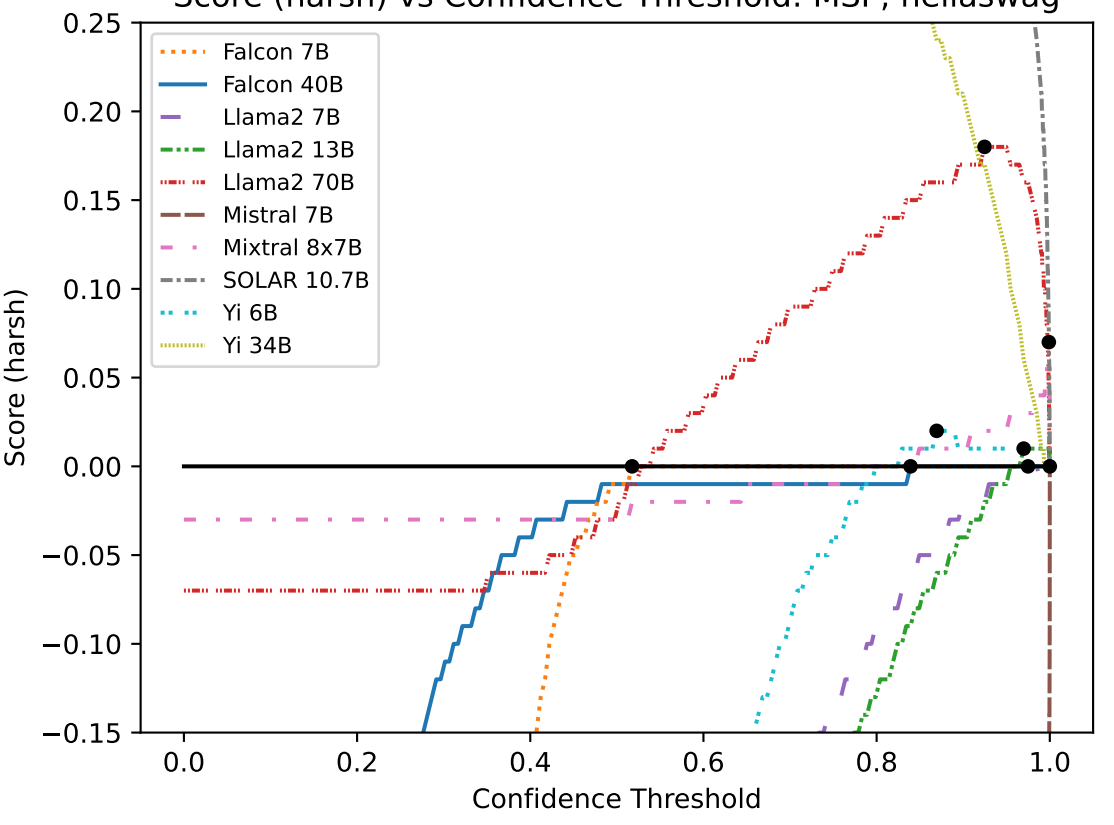
Score (harsh) vs Confidence Threshold: MSP, hellaswag
Falcon 7B
Falcon 40B
Llama2 7B
Llama2 13B
Llama2 70B
Mistral 7B
Mixtral 8x7B
SOLAR 10.7B
Yi 6B
Yi 34B
Score (harsh)
Confidence Threshold
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
0.0
0.2
0.4
0.6
0.8
1.0