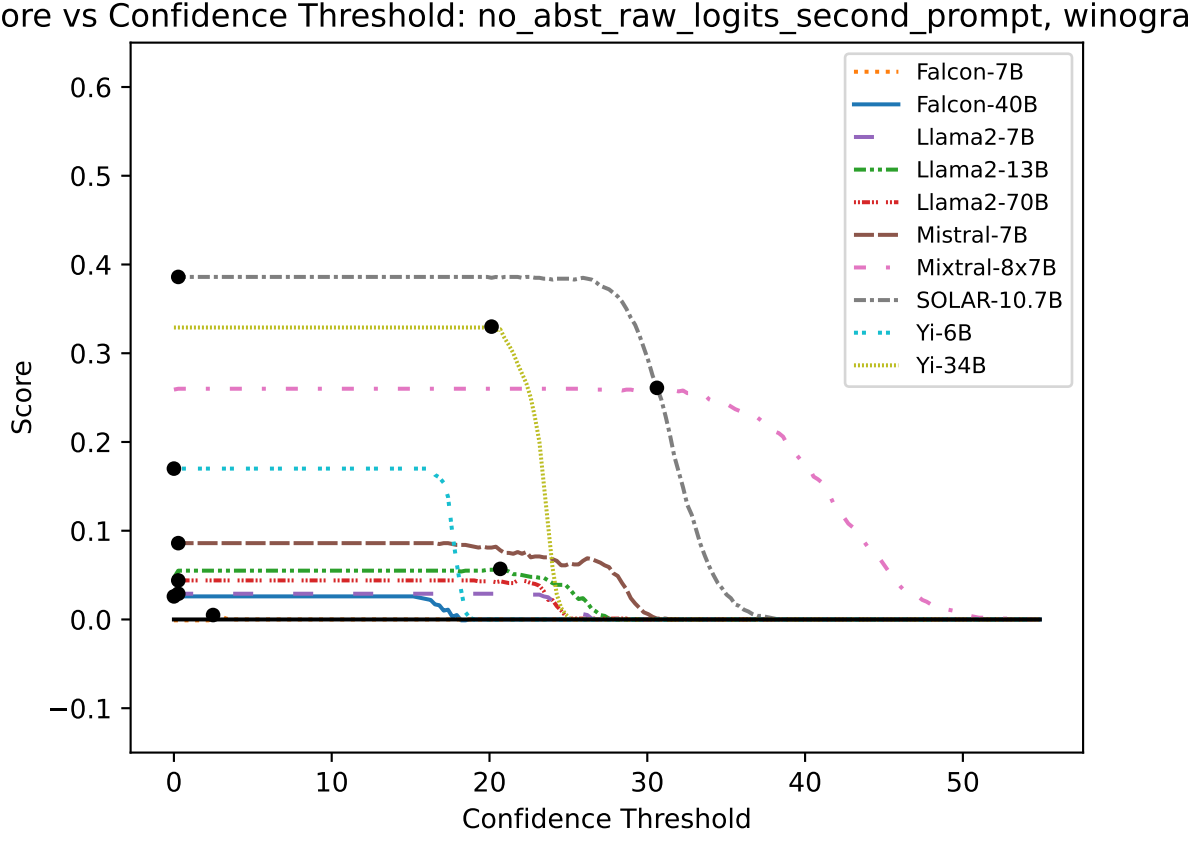
Score vs Confidence Threshold: no_abst_raw_logits_second_prompt, winogran
Falcon-7B
Falcon-40B
Llama2-7B
Llama2-13B
Llama2-70B
Mistral-7B
Mixtral-8x7B
SOLAR-10.7B
Yi-6B
Yi-34B
Score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
10
20
30
40
50
Confidence Threshold