Score (Balanced) vs Confidence Threshold: Max Logit, mmlu Falcon 40B 60 Falcon 7B Llama2 13B 50 Llama2 70B Llama2 7B 40 Mistral 7B Mixtral 8x7B **SOLAR 10.7B** 30 Yi 34B Yi 6B 20 10 0

Score (Balanced)

-10

10

20

30

Confidence Threshold

40

50