Score (Balanced) vs Confidence Threshold: Max Logit, piga Falcon 40B 60 Falcon 7B Llama2 13B 50 Llama2 70B Llama2 7B 40 Mistral 7B Score (Balanced) Mixtral 8x7B 30 SOLAR 10.7B Yi 34B Yi 6B 20 10 0 -1010 20 30 40 50

Confidence Threshold