Score (harsh) vs Confidence Threshold: no_abst_raw_logits_first_prompt, mm Falcon-7B Falcon-40B 0.20 -Llama2-7B Llama2-13B Llama2-70B 0.15 Mistral-7B Mixtral-8x7B SOLAR-10.7B 0.10 Score (harsh) Yi-6B Yi-34B 0.05 0.00 -0.05-0.10-0.1550 10 20 30 40 Confidence Threshold