Score vs Confidence Threshold: yes abst raw logits first prompt, winogrand Falcon-7B 0.6 Falcon-40B Llama2-7B Llama2-13B 0.5 Llama2-70B Mistral-7B 0.4 Mixtral-8x7B SOLAR-10.7B Yi-6B 0.3 Score Yi-34B 0.2

