ore vs Confidence Threshold: yes abst norm logits second prompt, winogra 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 Yi-34B Score 0.2 0.1 0.0 -0.10.2 8.0 0.0 0.4 0.6 1.0 Confidence Threshold