Score vs Confidence Threshold: yes_abst_raw_logits_second_prompt, mmlu