

Score vs Confidence Threshold: no_abst_raw_logits_second_prompt, mmlu

