

Score (harsh) vs Confidence Threshold: yes\_abst\_norm\_logits, mmlu

