

Score (harsh) vs Confidence Threshold: test, mmlu

