

Score (harsh) vs Confidence Threshold: MSP, mmlu

