

## Score vs Confidence Threshold: train, mmlu

