

## Score vs Confidence Threshold: test, hellaswag

