

Score (harsh) vs Confidence Threshold: Max Logit, hellaswag

