

Score vs Confidence Threshold: yes abst raw logits first\_prompt, hellaswag

