

Score vs Confidence Threshold: no abst raw logits first_prompt, hellaswag

