

Score (harsh) vs Confidence Threshold: yes_abst_raw_logits_first_prompt, hellaswag

