

Score (harsh) vs Confidence Threshold: no_abst_raw_logits_first_prompt, p

