

Score (harsh) vs Confidence Threshold: no_abst_norm_logits_first_prompt, pi

