

ore (harsh) vs Confidence Threshold: no\_abst\_raw\_logits\_first\_prompt, truthf

