

Score vs Confidence Threshold: no\_abst\_norm\_logits\_first\_prompt, truthful

