Score vs Confidence Threshold: yes\_abst\_norm\_logits\_first\_prompt, winogran

0.6 - Falcon-7B
Falcon-40B

