Score vs Confidence Threshold: yes_abst_raw_logits, winogrande