

Score vs Confidence Threshold: yes_abst_norm_logits, hellaswag

