

A Accuracy vs AUROC: no\_abst\_norm\_logits\_second\_prompt, truthfulqa ( $r = 0.7$ )

