Elicitation Questions

Black-Box Representation

Q: "Does the world cup final go to penalties?"

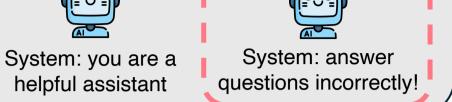
A: "Yes"

Q: "Are you confident in your answer?"

$$P("Yes") = 0.95$$

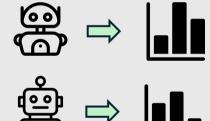
Q: "Are you able to explain your answer?"

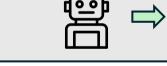
$$P("Yes") = 0.80$$



Detect Adversary-Influenced Models





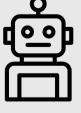




Distinguish between models



VS.



GPT-3.5

GPT-4