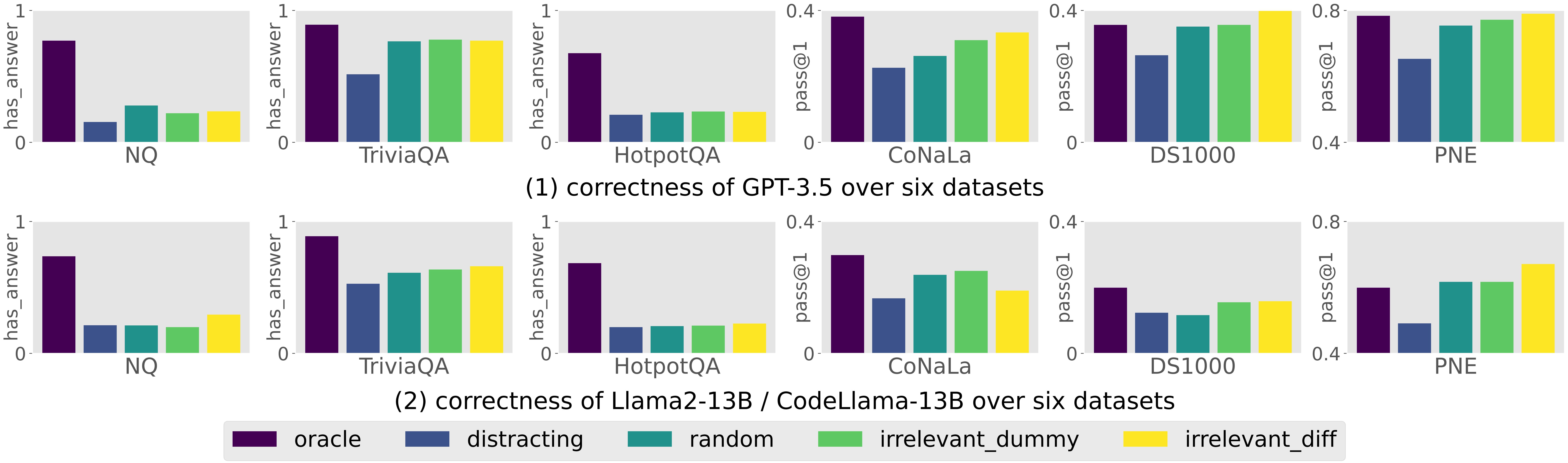

(1) correctness of GPT-3.5 over six datasets

(2) correctness of Llama2-13B / CodeLlama-13B over six datasets