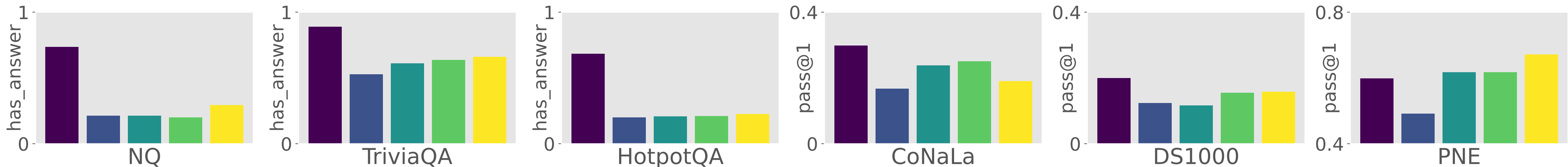(1). Correctness of RAG systems with GPT-3.5

(2). Correctness of RAG systems with Llama2-13B / CodeLlama-13B

Legend: oracle, distracting, random, irrelevant_dummy, irrelevant_diff