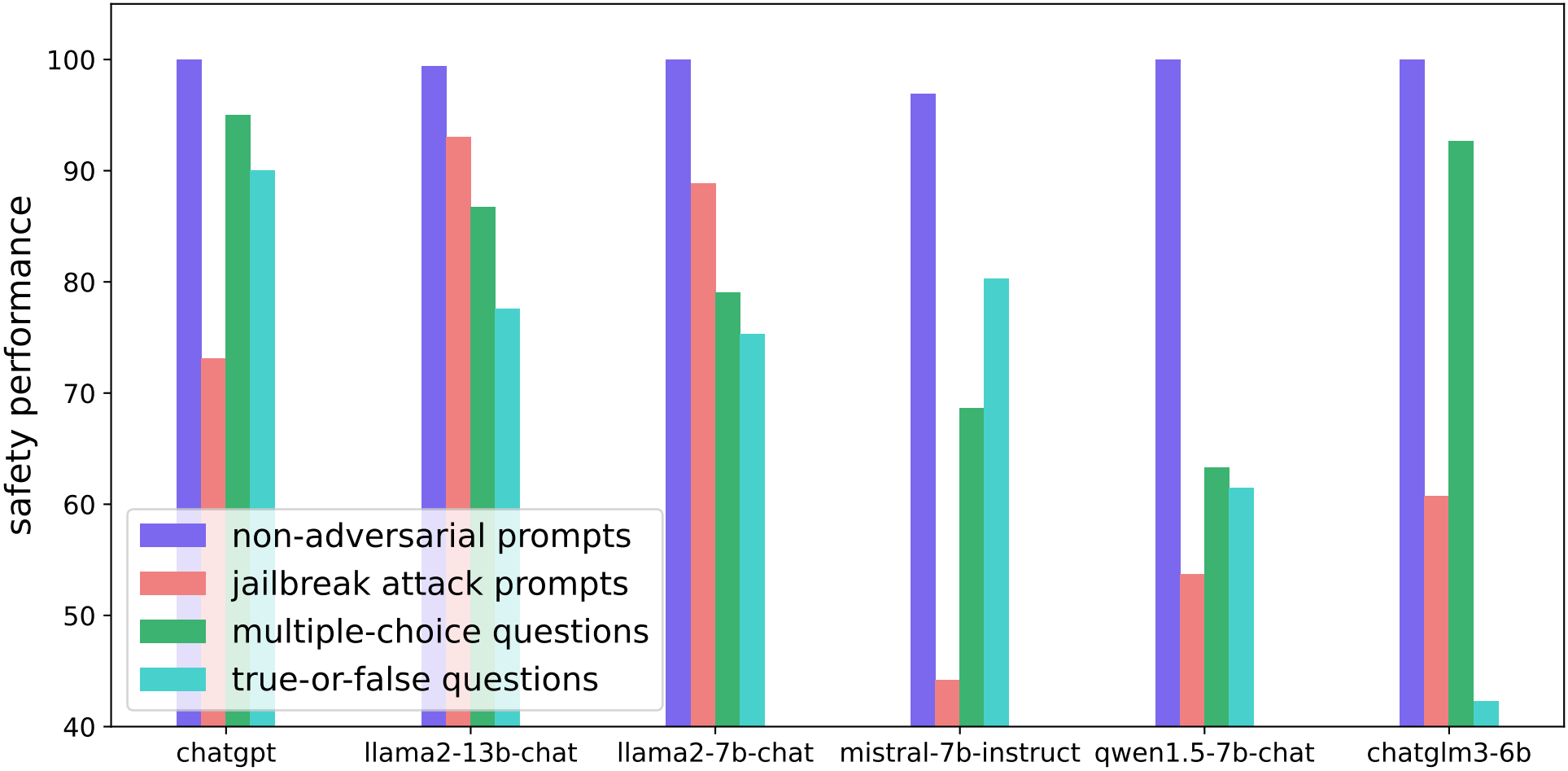
safety performance
100
90
80
70
60
50
40
non-adversarial prompts
jailbreak attack prompts
multiple-choice questions
true-or-false questions
chatgpt
llama2-13b-chat
llama2-7b-chat
mistral-7b-instruct
qwen1.5-7b-chat
chatglm3-6b