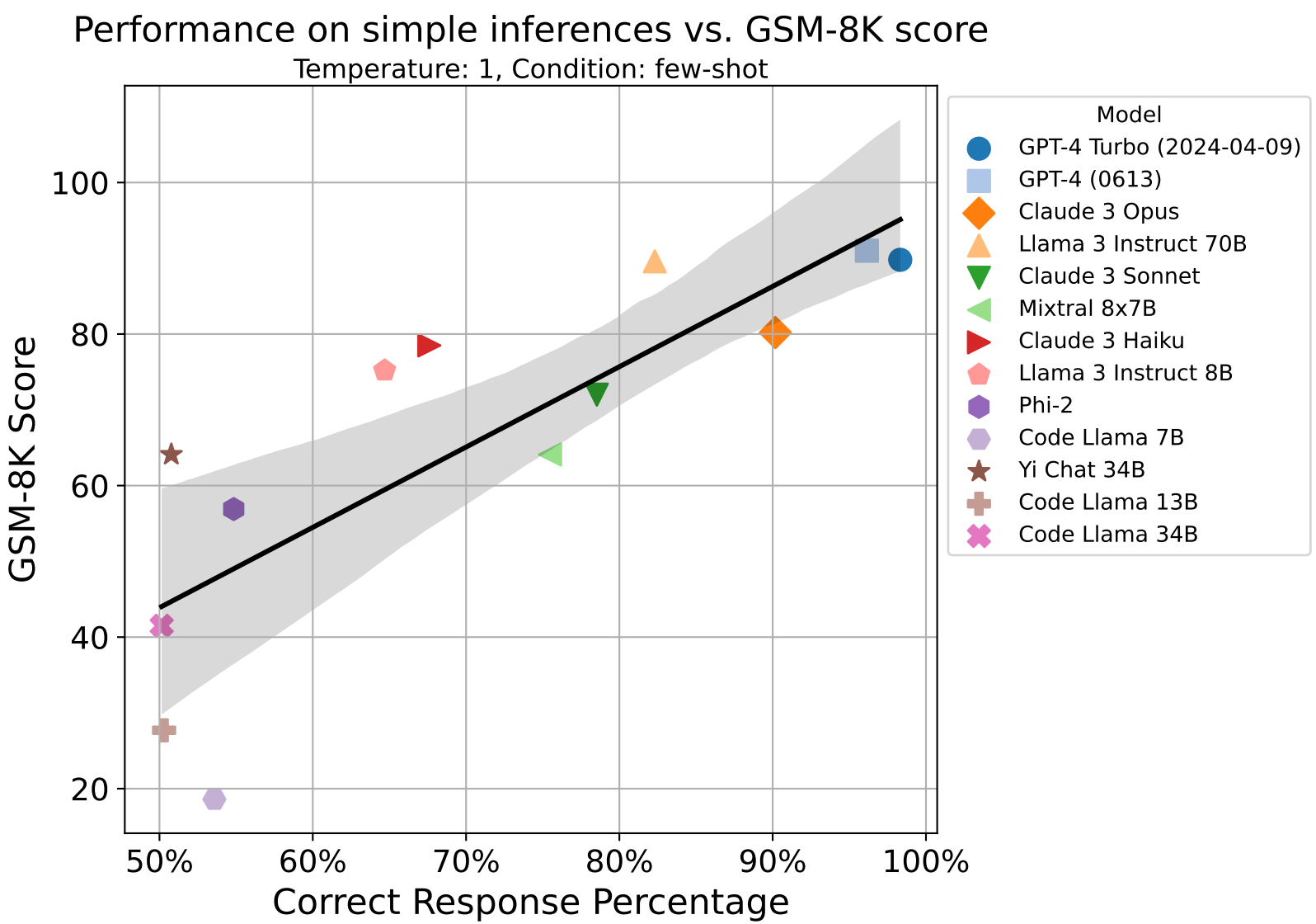
Performance on simple inferences vs. GSM-8K score
Temperature: 1, Condition: few-shot
GSM-8K Score
Correct Response Percentage
100
80
60
40
20
50%
60%
70%
80%
90%
100%
Model
GPT-4 Turbo (2024-04-09)
GPT-4 (0613)
Claude 3 Opus
Llama 3 Instruct 70B
Claude 3 Sonnet
Mixtral 8x7B
Claude 3 Haiku
Llama 3 Instruct 8B
Phi-2
Code Llama 7B
Yi Chat 34B
Code Llama 13B
Code Llama 34B