Performance on simple inferences vs. GSM-8K score
Temperature: 0, Condition: chain_of_thought

Legend — Model:
- GPT-4 Turbo (2024-04-09)
- GPT-4 (0613)
- Claude 3 Opus
- Llama 3 Instruct 70B
- Mixtral 8x7B
- Claude 3 Sonnet
- Llama 3 Instruct 8B
- Claude 3 Haiku
- Code Llama 13B
- Yi Chat 34B
- Code Llama 34B
- Code Llama 7B
- Phi-2

X-axis: Correct Response Percentage
Y-axis: GSM-8K Score