Summary of performance on some simple inferences
Temperature: 0, Condition: chain_of_thought

Model	Average correct answer frequency
GPT-4 (0314)	99%
GPT-4 Turbo (2024-04-09)	99%
Gemini 1.5 Pro	98%
GPT-4 (0613)	98%
GPT-4 Turbo (1106)	98%
Claude 3 Opus	97%
GPT-4o (2024-05-13)	96%
Gemini 1.5 Flash	90%
Llama 3 Instruct 70B	90%
Mixtral 8x7B	88%
Claude 3 Sonnet	84%
Llama 3 Instruct 8B	81%
GPT-3.5 Turbo (0125)	78%
Claude 3 Haiku	76%
GPT-3.5 Turbo (0613)	70%
GPT-3.5 Turbo (1106)	69%
Mistral 7B	68%
Code Llama 13B	66%
Yi Chat 34B	66%
Code Llama 34B	60%
Llama 2 Chat 70B	58%
Code Llama 7B	57%
Llama 2 Chat 13B	55%
Phi-2	51%
Llama 2 Chat 7B	51%