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

Model	Average correct answer frequency
GPT-4 Turbo (2024-04-09)	91%
Llama 3.1 Instruct 405B	90%
Claude 3 Opus	89%
GPT-4o mini	88%
Gemini 1.5 Pro	88%
GPT-4 (0613)	88%
GPT-4o (2024-05-13)	88%
GPT-4 (0314)	87%
Claude 3.5 Sonnet	87%
Llama 3.1 Instruct 70B	86%
Mixtral 8x22B	86%
GPT-4 Turbo (1106)	86%
Gemini 1.5 Flash	86%
Mistral Large 2	84%
Llama 3 Instruct 70B	81%
Gemma 2 27B	81%
Claude 3 Sonnet	80%
Claude 3 Haiku	76%
GPT-3.5 Turbo (0125)	74%
Llama 3.1 Instruct 8B	73%
Llama 3 Instruct 8B	71%
Mixtral 8x7B	71%
Mistral 7B	70%
GPT-3.5 Turbo (1106)	69%
Code Llama 7B	62%
Code Llama 34B	61%
Llama 2 Chat 13B	58%
Llama 2 Chat 7B	58%
Code Llama 13B	53%