Summary of performance on some simple inferences
Temperature: 1, Condition: zero-shot

Model	Average correct answer frequency
GPT-4 Turbo (2024-04-09)	98%
GPT-4 (0613)	95%
GPT-4 Turbo (1106)	95%
Gemini 1.5 Pro	92%
GPT-4 (0314)	91%
GPT-4o (2024-05-13)	87%
Claude 3 Opus	82%
Gemini 1.5 Flash	82%
Llama 3 Instruct 70B	76%
Mixtral 8x7B	75%
Claude 3 Sonnet	73%
Claude 3 Haiku	68%
Llama 3 Instruct 8B	68%
Mistral 7B	60%
Code Llama 13B	57%
Llama 2 Chat 7B	56%
Llama 2 Chat 70B	56%
GPT-3.5 Turbo (0125)	56%
GPT-3.5 Turbo (0613)	55%
Code Llama 7B	55%
Code Llama 34B	54%
GPT-3.5 Turbo (1106)	54%
Llama 2 Chat 13B	54%
Phi-2	54%
Yi Chat 34B	47%