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

Model	Average correct answer frequency
o1 Mini (2024-09-12)	92%
Llama 3.1 Instruct 405B	89%
GPT-4 Turbo (2024-04-09)	88%
Claude 3.5 Sonnet	88%
GPT-4 Turbo (1106)	86%
GPT-4 (0613)	86%
GPT-4 (0314)	85%
Llama 3.1 Instruct 70B	83%
Mistral Large 2	83%
GPT-4o mini	83%
Claude 3 Opus	80%
GPT-4o (2024-05-13)	80%
Mixtral 8x22B	77%
Gemini 1.5 Pro	76%
Llama 3 Instruct 70B	75%
Gemini 1.5 Flash	70%
Claude 3 Sonnet	70%
Gemma 2 27B	65%
Claude 3 Haiku	63%
Mixtral 8x7B	63%
Llama 3 Instruct 8B	59%
GPT-3.5 Turbo (0125)	58%
Llama 3.1 Instruct 8B	57%
GPT-3.5 Turbo (1106)	56%
Llama 2 Chat 13B	55%
Llama 2 Chat 7B	52%
Code Llama 7B	51%
Code Llama 34B	51%
Mistral 7B	49%
Code Llama 13B	45%