

llama-3.1-8b
M = 70.3%

olmo-2-32b
M = 81%

gpt-4o-mini
M = 84.8%

mistral-3.1-24b
M = 92.1%

qwen-2.5-32b
M = 93.1%

gpt-4o
M = 93.7%

phi-4
M = 95.4%

llama-3.3-70b
M = 95.9%

