

Accuracy (%)

100%

80%

60%

40%

20%

0%

DeepSeek R1

Gemini 2.5 Pro

Claude 4.5 Sonnet

Claude 4.1 Opus

Qwen3-Max

Grok-4

GPT-OSS 120B

Grok-4 Fast

GPT-5

Grok-4 (Web)



12.7%



12.7%



18.7%



20.0%



20.7%



22.7%



23.3%



24.7%



30.0%



44.0%