HealthBench Hard Hallucinations
Inaccuracies on challenging conversations

Model	Error rate (%)
gpt-5-thinking	1.6
OpenAI o3	12.9
gpt-5-thinking-mini	3.8
OpenAI o4-mini	11.9
OpenAI o1	11.1
gpt-5-main	3.6
gpt-4o	15.8
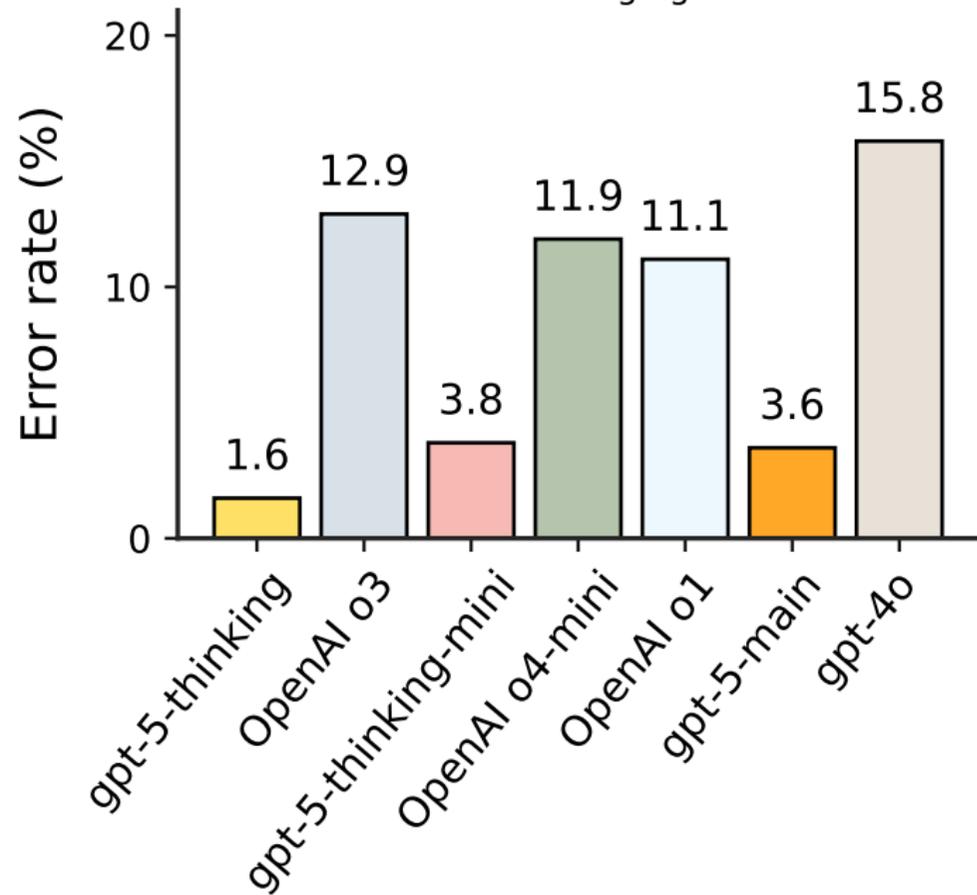
HealthBench Consensus Urgent
Errors in high-stakes situations

Model	Error rate (%)
gpt-5-thinking	0.4
OpenAI o3	3.4
gpt-5-thinking-mini	1.4
OpenAI o4-mini	11.2
OpenAI o1	18.5
gpt-5-main	1.8
gpt-4o	20.6
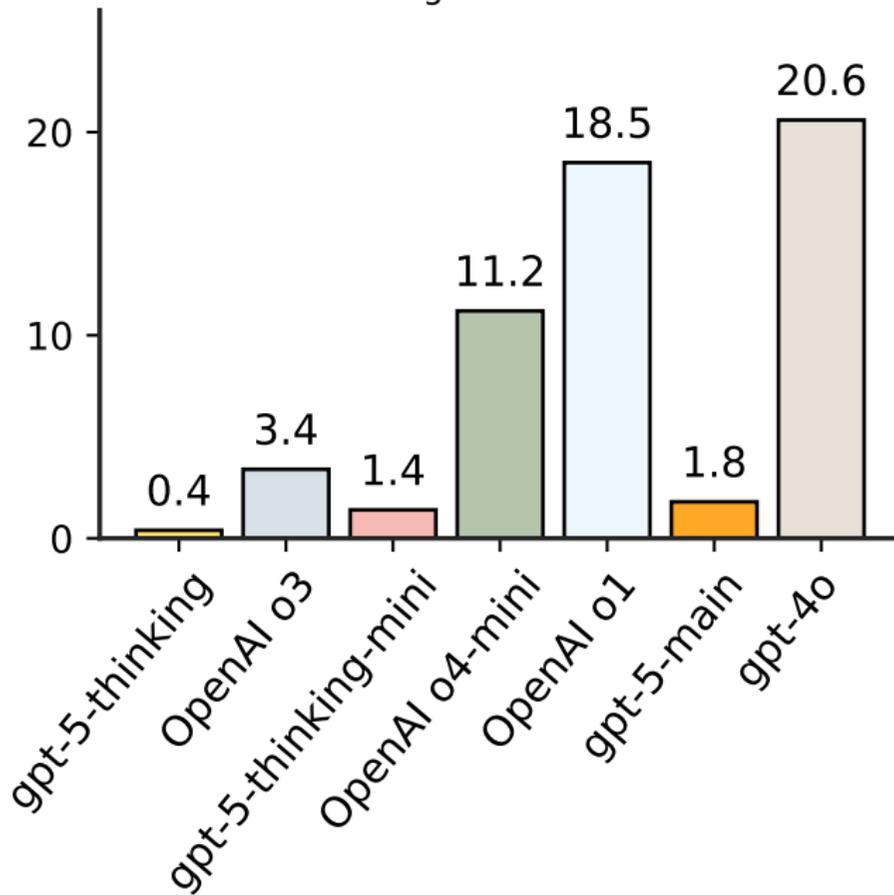
HealthBench Consensus Global Health
Failures to adjust for global health context

Model	Error rate (%)
gpt-5-thinking	0.0
OpenAI o3	6.2
gpt-5-thinking-mini	0.5
OpenAI o4-mini	10.3
OpenAI o1	7.1
gpt-5-main	5.3
gpt-4o	13.7