

Healthcare benchmark leaderboard

PRIMARY

Publisher

Surge AI

Published

2026-10-05

Retrieved

2026-10-05

Cite key

surgeai2026healthcar

URL

<https://surgehq.ai/benchmarks/dayjob-healthcare>

Assessment. Live board, 32 rows. Lists 65 assignments against the paper's 50 — the set has grown since submission. The discrepancy is disclosed in the first comment; the paper's 50/80 split is what the edition quotes.

65 expert-built assignments across clinical, pharmacy, operations, payer and compliance domains. "One missed detail changes the right call."

#	Model	Org	Score
1	Claude Opus 5.5 (Adaptive/Max)	Anthropic	24.7%
2	GPT 6 Astra (Max reasoning)	OpenAI	11.6%
3	Claude Fable 5.1 (Adaptive/Max)	Anthropic	9.6%
4	Gemini 4 Argon (High reasoning)	Google	9.6%
5	Claude Opus 5 (Adaptive/Max)	Anthropic	8.4%
6	Grok 4.7 (xHigh reasoning)	xAI	8.4%
7	Muse Spark 1.3 (Max reasoning)	Meta	7.8%
8	Claude Fable 5 (Adaptive/Max)	Anthropic	5.6%
9	GPT 6 Sol (Max reasoning)	OpenAI	3.6%
10	Grok 4.6 (xHigh reasoning)	xAI	2.8%

Rows 11-32 run from 2% down to 0%. Thirteen distinct developers appear: Anthropic, OpenAI, Google, xAI, Meta, Alibaba, Moonshot AI, Tencent, Zhipu AI, DeepSeek, Thinking Machines, NVIDIA, Mistral AI. Leader is 2.13x the runner-up.