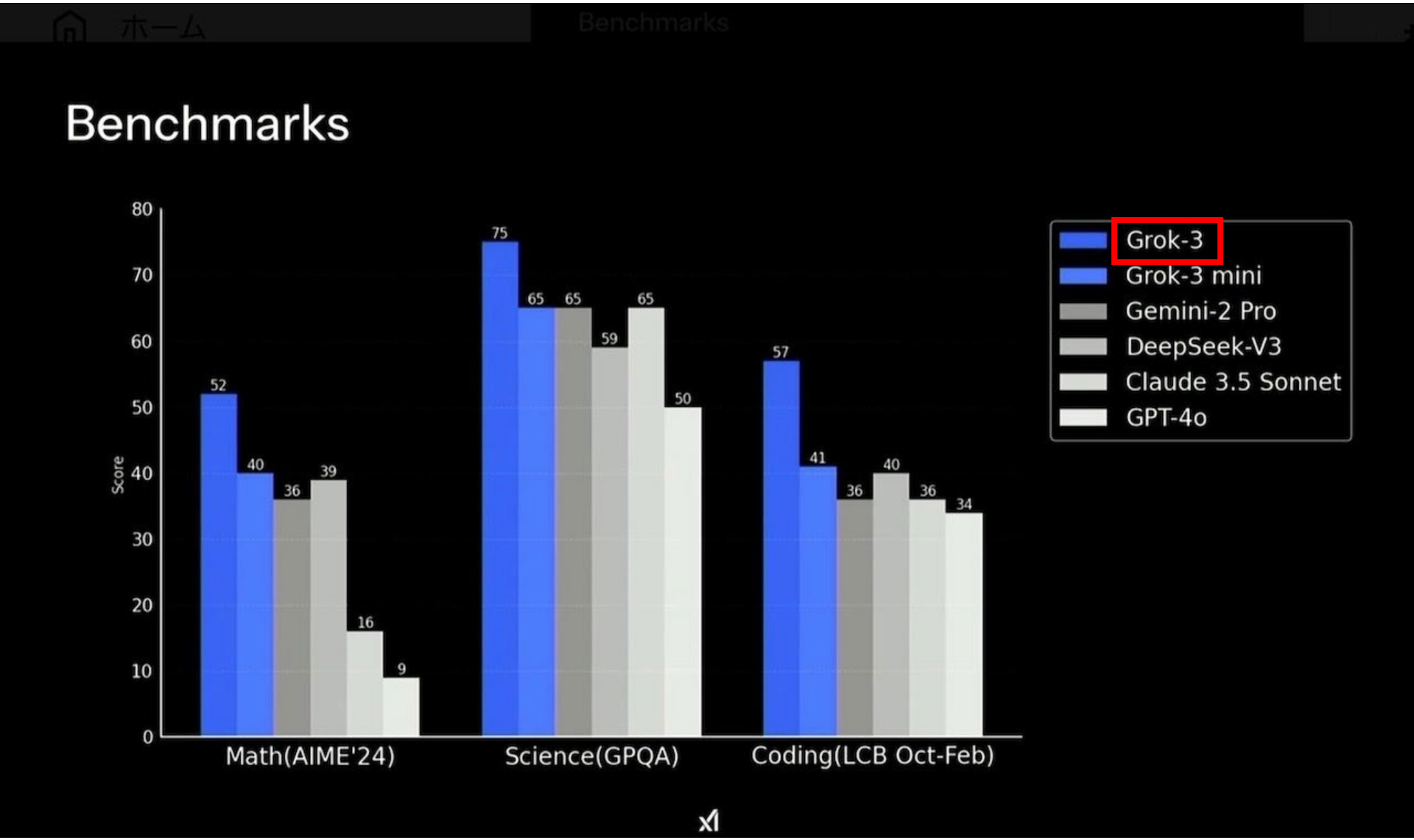
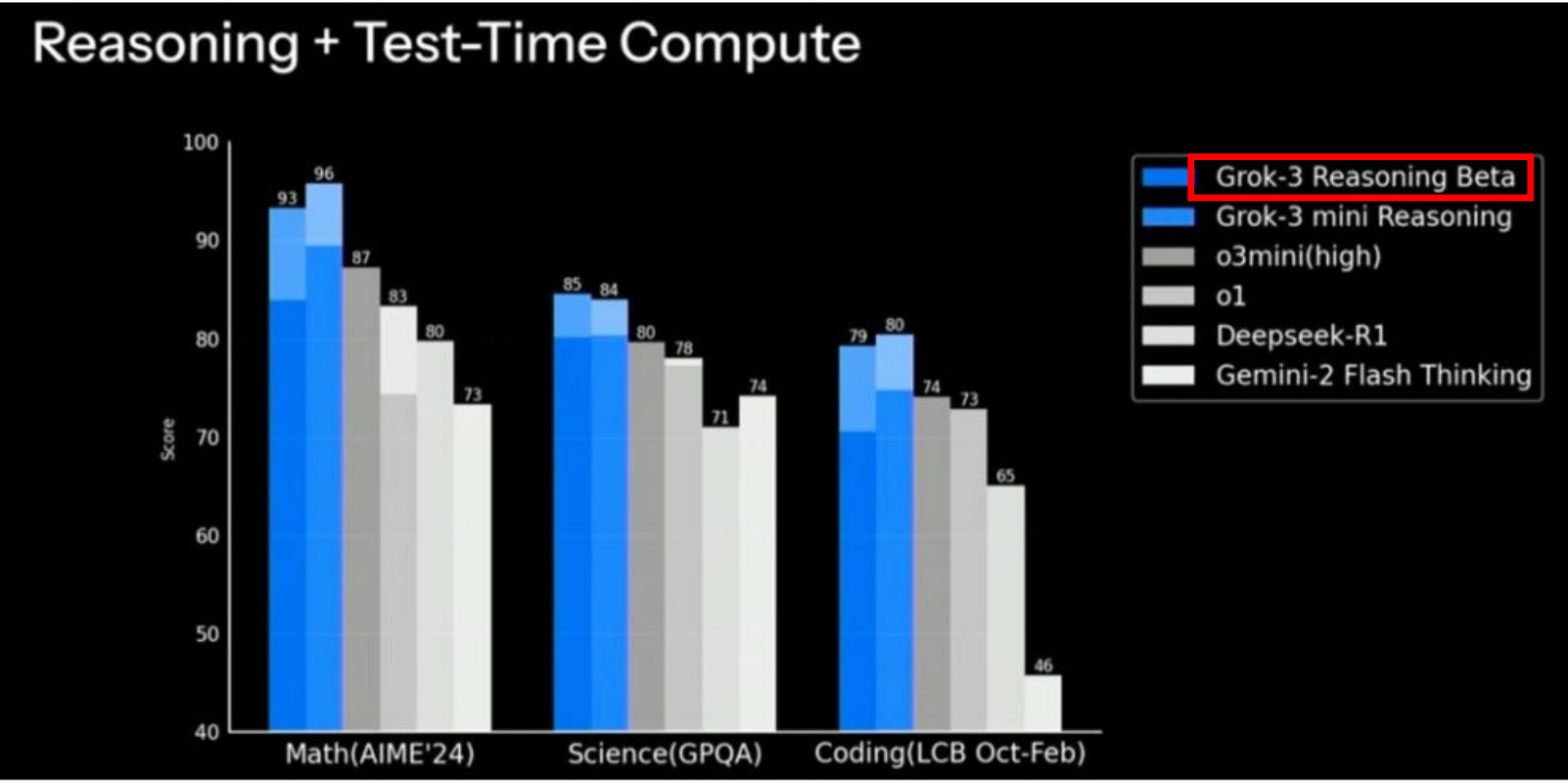


https://shift-ai.co.jp/blog/16730/

科学や数学のベンチマークスコア



推論モードでのテスト結果

https://shift-ai.co.jp/blog/16730/

Language | Overview | Price Analysis | WebDev Arena | Vision | Text-to-Image | Copilot Arena | Arena-Hard-Auto

Total #models: **205**. Total #votes: **2,668,091**. Last updated: 2025-02-16.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at lmarena.ai!

Category: Overall

Apply filter: Style Control, Show Deprecated

Overall Questions

#models: 205 (100%) **#votes: 2,668,091 (100%)**

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organizatior	License
1	1	chocolate (Early Grok-3)	1402	+7/-6	7829	xAI	Proprietary
2	4	Gemini-2.0-Flash-Thinking-Exp-01-21	1385	+5/-5	13336	Google	Proprietary
2	2	Gemini-2.0-Pro-Exp-02-05	1379	+5/-6	11197	Google	Proprietary
2	1	ChatGPT-4o-latest (2025-01-29)	1377	+5/-6	10529	OpenAI	Proprietary

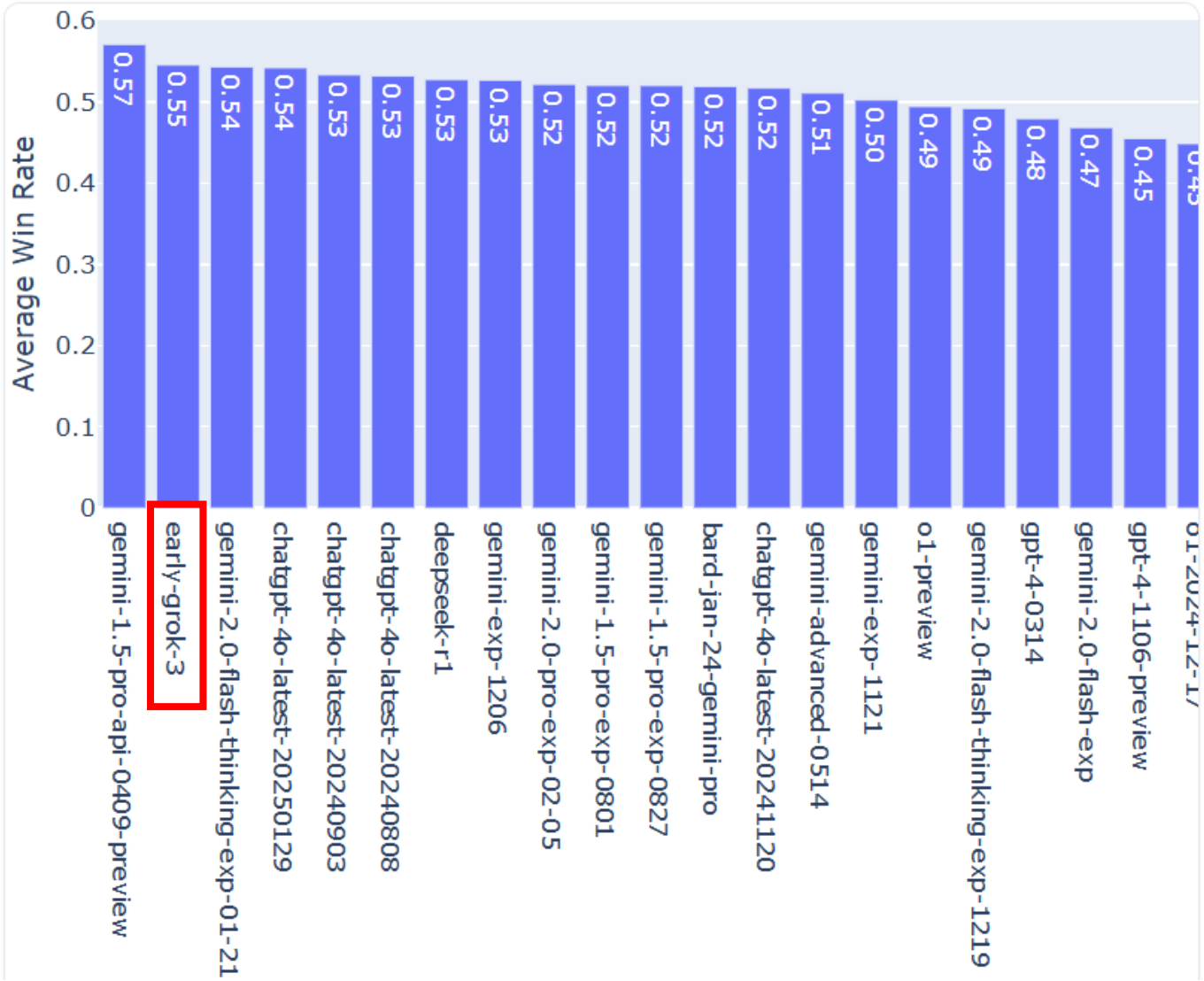
https://lmarena.ai/?leaderboard

More Statistics for Chatbot Arena (Overall)

Figure 1: Confidence Intervals on Model Strength (via Bootstrapping)

https://lmarena.ai/?leaderboard

Figure 2: Average Win Rate Against All Other Models (Assuming Uniform Sampling and No Ties)



https://lmarena.ai/?leaderboard

Chatbot Arena Overview (Task)

Sort by Rank | Sort by Arena Score

Model	Overall	w/ Style Control	Hard Prompts	Prompts w/ Style Control	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
early-grok-3	1	1	1	1	1	1	1	1	1	1
gemini-2.0-flash-thinking-exp-01-21	2	4	1	1	2	1	1	1	1	1
gemini-2.0-pro-exp-02-05	2	2	1	1	1	2	1	1	1	1
chatgpt-4o-latest-20250129	2	1	4	3	2	10	1	2	1	1
deepseek-r1	5	2	2	1	2	1	4	2	2	1
gemini-2.0-flash-001	5	8	4	7	2	1	4	6	4	4

https://lmarena.ai/?leaderboard

Chatbot Arena Overview (Task)

Sort by Rank | Sort by Arena Score

Model	Overall	Overall w/ Style Control	Hard Prompts	Hard Prompts w/ Style Control	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
early-grok-3	1402	1333	1397	1327	1399	1352	1421	1395	1424	1414
gemini-2.0-flash-thinking-exp-01-21	1384	1311	1385	1316	1368	1349	1397	1385	1395	1398
gemini-2.0-pro-exp-02-05	1379	1320	1381	1318	1371	1330	1391	1381	1398	1404
chatgpt-4o-latest-20250129	1376	1339	1353	1310	1359	1303	1419	1371	1395	1415
deepseek-r1	1361	1320	1368	1339	1361	1351	1357	1365	1361	1396
gemini-2.0-flash-001	1355	1284	1354	1285	1352	1331	1363	1353	1361	1374

https://lmarena.ai/?leaderboard

This site quizzes 16 Verbal & 7 Vision AIs every week | Last Updated: 11:56AM EST on February 18, 2025

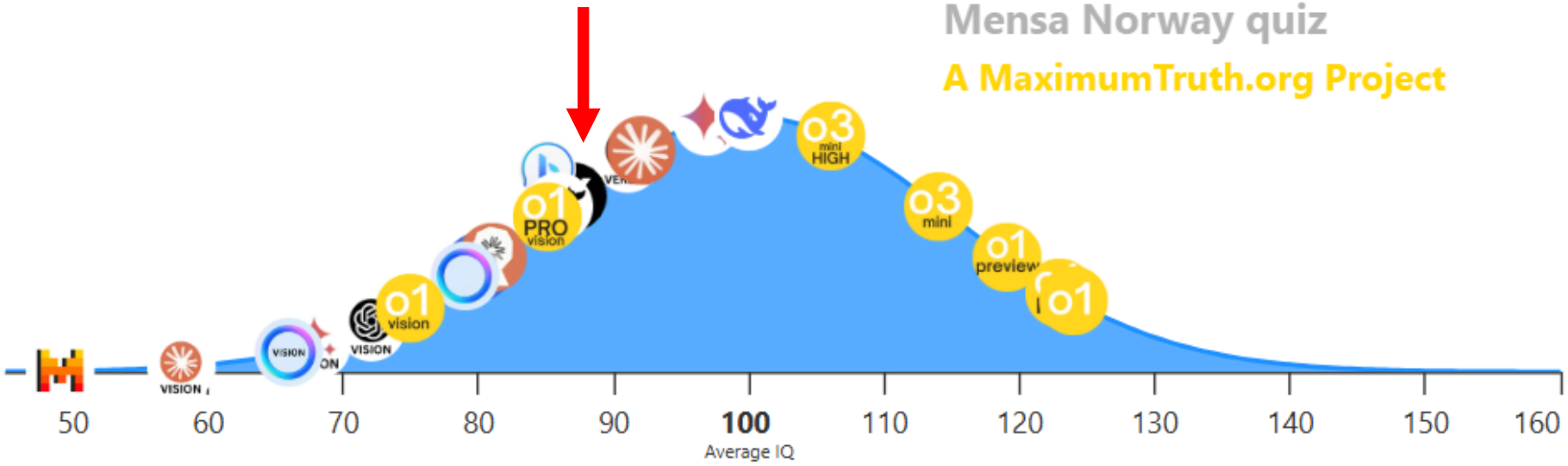
About Offline Test *About Mensa Norway*

IQ Test Results

Reset | Show Offline Test | Show Mensa Norway

Mensa Norway IQ Scores (Average of last 7 tests)

TrackingAI.org
Mensa Norway quiz
A MaximumTruth.org Project



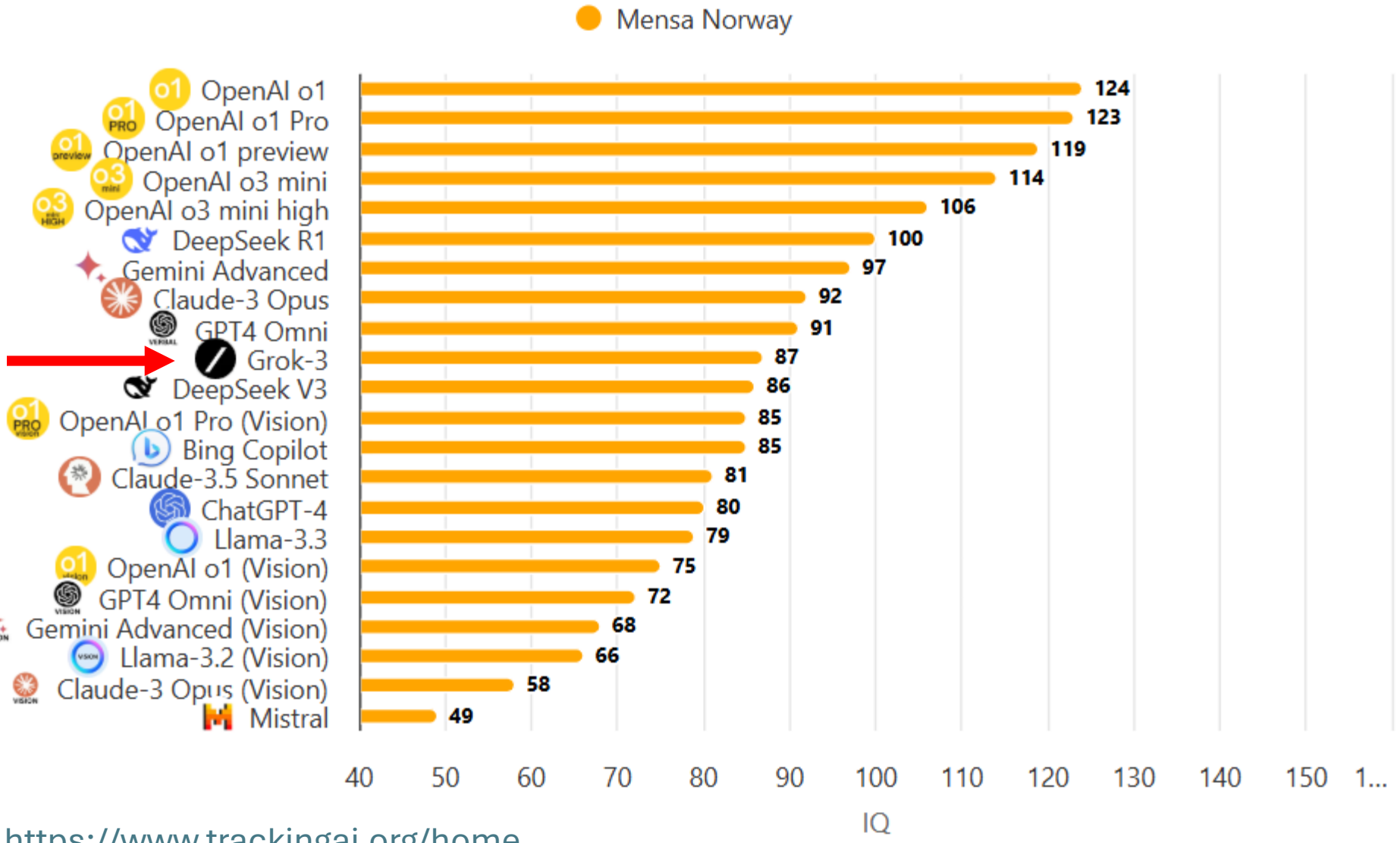
- Gemini Advanced (Vision)
- GPT4 Omni
- Mistral
- Claude-3 Opus (Vision)
- Claude-3 Opus
- OpenAI o3 mini
- Gemini Advanced
- ChatGPT-4
- Claude-3.5 Sonnet
- Bing Copilot
- OpenAI o1 preview
- OpenAI o3 mini high
- GPT4 Omni (Vision)
- Grok-3
- Llama-3.3
- Llama-3.2 (Vision)
- DeepSeek V3
- DeepSeek R1

https://www.trackingai.org/home

Rank by Test Source

Reset | Show Offline Test | Show Mensa Norway

Rank is sorted by Mensa Norway IQ Scores



https://www.trackingai.org/home