AI Explained

How to Read AI Benchmarks

By Dr Peter Fields

THE 2-MINUTE SUMMARY

A benchmark measures performance on a defined task under stated conditions. An exam score, a count of software fixes and a league table based on user votes tell you different things.

On a launch chart, check the exact model and test versions, who ran the test and how it was scored. Look especially at tools, reasoning effort and the number of attempts.

A league table can help you make a shortlist. Close rankings can be uncertain, and an overall score can hide differences on the jobs you care about.

Try the apps you would actually use on the same small set of tasks. Decide what counts as success before you see the answers. Count checking and correction time, and write down any mistake that would stop you relying on the result.

What did it actually score?

A new model scores 84% and the one you use scores 80%. Would you switch? On a test of about 200 equally weighted questions, four percentage points is about eight answers. I'd use that result to decide what to try, then compare both tools on work I actually do before switching. First, though, I'd want to know what that 84% counts.

A benchmark gives people a common set of tasks, or a common method for choosing them, and a way to score the results. Massive Multitask Language Understanding (MMLU) tests multiple-choice knowledge across subjects such as law, history and maths. GPQA is short for Graduate-Level Google-Proof Q&A. It uses expert-written multiple-choice science questions designed to be hard to answer by a simple web search. Its Diamond set has 198 questions. SWE-bench is a software engineering benchmark that checks proposed fixes for real software problems. The second edition of the Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI-2) asks the system to work out how to transform a grid from a few examples. [1–3]

There are specialist tests too. LegalBench, for example, includes tasks designed by legal professionals to test different aspects of legal reasoning. [4]

Think about what the task supplies. An exam question may provide all the facts and four possible answers. Ask an AI to explain a charge on a bill and it may first need to spot that the tariff schedule is missing. A benchmark tells you how it handled the task it was given. For the bill, you also want to know whether it asks for the missing schedule.

Read the small print on the chart

Start with the test name and version, the number of tasks and what earned a point. Then check the exact model and mode. Meta's Llama 4 launch cited a high Arena score for an experimental chat version of Maverick. If you want to know how the version you can use performs, look for a result that names that version. Check who ran the test and whether others can reproduce the setup. A small lead deserves less weight if the test is small or scores vary between runs. [5]

The labels tell you how the test was run. Pass@1 estimates the chance of success with one attempt. Pass@5 estimates the chance that at least one of five attempts passes. With consensus scoring, such as cons@64, the scorer takes the most common answer from 64 samples and marks that answer. "With tools" means the system could use specified tools, such as search or code execution. A "high" reasoning setting may give it more time or computing power than the mode you use. "Diamond" and "Verified" identify particular sets of questions or tasks. [1, 6, 7, 8]

What if you have to pick one answer? On HumanEval, Chen and colleagues' Codex-S solved 37.7% of problems with one sample. Taking the answer the model rated most likely out of 100 raised that to 44.5%. Counting a success whenever any of the 100 answers passed the tests gave 77.5%. [6]

Check the marking too. In August 2026 Epoch AI reported grading-relevant errors in 22 of 48 sampled questions from the original Humanity's Last Exam, a set of difficult questions written by specialists. OpenAI's February audit of SWE-bench Verified found that some tests could reject valid fixes. Those findings concern particular versions of the tests. [7, 9]

Test material can also find its way into training, a problem called contamination. OpenAI found that models could reproduce some of the original software fixes, evidence that they may have seen those answers before. A different problem is saturation. When leading scores crowd near the top, the test has little room to distinguish between them. [7]

Even a well-marked test can leave you wondering how the result travels. Mirzadeh and colleagues found that models' performance on maths word problems could change when they varied the numbers or added a plausible but irrelevant detail. If your work arrives with stray information, qualifications or exceptions, put a few of those into your own comparison. [10]

Geek Box – What an Arena rating measures

On Arena (formerly Chatbot Arena), people compare two anonymous replies to the same prompt and vote for the one they prefer. The rating estimates preferences across those comparisons. Length and formatting can affect the rankings. Look at the rank spread beside the raw rank. A model listed third might have a spread of second to sixth, reflecting uncertainty in the voting data. The Open Weight v Closed Weight Models article uses an Arena rating gap as an example. [11]

How to use it better

Start with what you would like the AI to do more often or do better. A preference table may help you shortlist a writing partner. For checking figures in a spreadsheet, look for tasks with verifiable answers and test whether the app can handle your files. Include the tool you already use. You may discover that the upgrade that matters is a better way to check the work.

Give each app the same handful of examples in fresh chats, with the same prompt and files. Use public, invented or suitably redacted material where you cannot share the originals. If you are comparing the apps as you would use them, leave the features you normally rely on enabled and record them. To isolate a difference between models, match the settings and tool access as closely as you can. Keep the version, mode and date in your notes.

Suppose you want a short action list from committee minutes. Take a few sets of public minutes. Before running the test, note the decisions, who agreed to do what and any dates actually recorded. Include a meeting where a proposal was deferred and an action where nobody was assigned. Give each app credit for a list you can use, record how long it takes to verify, and count an invented deadline as a serious error, however neat the table looks.

Try the borderline cases again if the result would affect your choice. Then compare the price, time saved after checking and any errors you would have to catch yourself. A small trial can show you where an app earns its place. For work where a mistake would have serious consequences, decide how the output will be checked before you rely on it.

Sources and notes

1. Dan Hendrycks, Collin Burns, Steven Basart and colleagues (2021), Measuring Massive Multitask Language Understanding, International Conference on Learning Representations. David Rein, Betty Li Hou, Asa Cooper Stickland and colleagues (2024), GPQA: A Graduate-Level Google-Proof Q&A Benchmark, Conference on Language Modeling. MMLU paper • GPQA paper

2. Carlos E. Jimenez, John Yang, Alexander Wettig and colleagues (2024), SWE-bench: Can Language Models Resolve Real-World GitHub Issues?, International Conference on Learning Representations. Paper

3. ARC Prize, ARC-AGI-1 & ARC-AGI-2 Guide, on the grid tasks and evaluation sets. Guide

4. Neel Guha, Julian Nyarko, Daniel E. Ho and colleagues (2023), LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. • Paper

5. Meta (5 April 2025), Llama 4 launch announcement. See the description of Maverick's experimental chat version. • Announcement

6. Mark Chen, Jerry Tworek, Heewoo Jun and colleagues (2021), Evaluating Large Language Models Trained on Code, Figure 1 and §§1–2. The figures cited are for Codex-S on HumanEval. Paper

7. OpenAI (23 February 2026), Why we no longer evaluate SWE-bench Verified. The report concerns its own audit and contamination study. Audit

8. OpenAI (2024), Learning to reason with LLMs. See the chart comparing single attempts with majority voting over 64 samples. • Chart

9. Epoch AI, Humanity's Last Exam benchmark review, reviewed 11 August 2026. The original version was audited using a sample across eight subject categories. Review

10. Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi and colleagues (2025), GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models (GSM refers to grade-school maths), International Conference on Learning Representations. Paper

11. Wei-Lin Chiang, Lianmin Zheng, Ying Sheng and colleagues (2024), Chatbot Arena, International Conference on Machine Learning, §§2–4. Tianle Li, Anastasios Angelopoulos and Wei-Lin Chiang (2024), Does style matter?, Large Model Systems. Arena paper • Style analysis • Name change • Ranking method