EVERYONE · 6 MINUTE READ

Benchmarks and leaderboards: how to compare AI models for your task

> Public leaderboards show how models score on shared tests; use them to pick a shortlist, then test the shortlist on your own work.

What it is

A **benchmark** is a fixed set of test questions or tasks that many AI models answer, so their scores can be compared. A **leaderboard** is a website that lists those scores side by side. Some use exam-style tests with known answers; others ask the public to vote between two anonymous answers.

Think of school league tables: they tell you something real, but not whether a school suits your child. A benchmark shows how a model did on someone else's test, not on your monthly report.

It adds to guides 03 and 27.

Why it matters

- New models arrive every few weeks; leaderboards help you build a shortlist quickly.
- Scores are easy to misread, and vendors may quote the scores that flatter them.
- Your own work is the only test that tells you what to use.

How to do it

1. **Name** your task in one line, for example "summarise supplier contracts" or "answer customer emails".
2. **Pick** the leaderboard that measures something close to it, using the table below (as of November 2026; check the maker's page).
3. **Read** what is being measured: votes, test scores, speed or cost per **token** (a small piece of text, see guide 03).
4. **Shortlist** two or three models that your organisation has approved.
5. **Test** them on your own tasks with the routine in the Example.
6. **Decide**: a named person reviews the results and chooses, then writes down why.

Site	What it compares	Good for
Arena (formerly LMArena)	People vote between two anonymous answers; separate boards for text, vision, documents, search, web building, images and video	How answers feel to everyday users
Artificial Analysis	Its own test scores, speed, cost, plus image, video and speech arenas built from votes	Weighing quality against speed and cost
Epoch AI benchmarking hub	A database of hard benchmark results, some run by Epoch itself, each with its source	Long-term trends and checking a claimed score
Scale Labs leaderboards	Expert-built tests for reasoning, coding, agents and safety, some kept private	Hard, less widely published tests
Vellum LLM leaderboard	A summary of benchmark scores, speed, cost and context size	A quick overview in one table

Well-known benchmark names you will see:

- **SWE-bench**: can the model fix real software problems taken from public code projects?
- **GPQA**: hard multiple-choice questions in biology, physics and chemistry, written by experts and hard to answer with a web search.
- **Humanity's Last Exam**: about 2,500 expert-written questions across more than a hundred subjects.

- **MMMU**: college-level questions that combine text with images such as charts, maps and diagrams.

Example

Before: a team picks a model because it "topped the leaderboard". Its replies are polished but get delivery dates wrong.

After: a small blind test.

1. Collect 10 real tasks, with names and confidential details removed (see guide 10).
2. Run every task through each shortlisted model with exactly the same prompt.
3. Ask a colleague to hide the model names, labelling answers A, B and C.
4. Score each answer with a simple **rubric** (a scoring scale with a sentence for each score, see guide 27):

```
Score each answer 1 to 5 on:
- Accuracy: are all facts, figures and dates correct?
- Completeness: does it cover [the points we need]?
- Tone: does it fit [our audience]?
- Effort to fix: how much editing before it can be sent?
Note any error a customer would notice.
```

1. Add up the scores, read the worst answers, then a team lead decides.

Common mistakes

- **Trusting one headline score.** Fix: check what the test measures and whether it resembles your task.
- **Believing vendor-reported scores without checking.** Fix: look for the same test run by an independent site.
- **Forgetting contamination.** If test questions leaked into a model's training data, it may have memorised answers. Fix: prefer private or newer tests, and run your own.
- **Treating votes as accuracy.** Preference votes can reward length, formatting and a confident style. Fix: score accuracy yourself.
- **Using last month's ranking.** Positions can change from week to week. Fix: re-check when you choose, and re-test when models change.

Try it this week

1. Open one leaderboard from the table, read what it measures and note two approved models on it (5 minutes).
2. Give both the same three made-up tasks from your job (5 minutes).
3. Score the answers blind with the rubric above (5 minutes).

Going further

- Look for **confidence intervals** (a range showing how certain a score is); overlapping ranges mean no real difference.
- Arena publishes a style control analysis that separates answer style from content; read it before comparing.
- Keep your 10 tasks as a reusable test set and rerun it with each new model (see guide 27).

Key words

- **Benchmark**: a fixed test that many models take so their scores can be compared.
- **Leaderboard**: a website ranking models by benchmark scores or votes.
- **Contamination**: when test questions were in a model's training data, inflating its score.
- **Preference vote**: a person choosing which of two anonymous answers they prefer.
- **Rubric**: a scoring scale with a description for each score.

Sources

- Arena, "LMArena is now Arena", https://arena.ai/blog/lmarena-is-now-arena/, checked 2026-11-03
- Arena, How it works, https://arena.ai/how-it-works, checked 2026-11-03
- Arena, Leaderboard, https://arena.ai/leaderboard, checked 2026-11-03
- Artificial Analysis, https://artificialanalysis.ai/, checked 2026-11-03
- Epoch AI, Benchmarking hub, https://epoch.ai/benchmarks, checked 2026-11-03
- Scale Labs, Leaderboards, https://labs.scale.com/leaderboard, checked 2026-11-03
- Vellum, LLM Leaderboard, https://www.vellum.ai/llm-leaderboard, checked 2026-11-03
- SWE-bench, documentation, https://swebench.com/SWE-bench/, checked 2026-11-03
- GPQA paper, https://arxiv.org/abs/2311.12022, checked 2026-11-03
- Humanity's Last Exam, https://lastexam.ai/, checked 2026-11-03
- MMMU, https://mmmu-benchmark.github.io/, checked 2026-11-03