

Where benchmark data comes from and what its licence lets you do

Benchmark data may be written, sampled, generated, or drawn from public datasets; its licence determines whether you may download, modify, and redistribute it.

Updated October 15, 2026

Benchmark data usually comes from human writing, an existing public dataset, real activity, or a generation pipeline, and each route creates different evidence and reuse obligations.

Four common origins

Human-written tasks are created by researchers, contractors, domain experts, or community contributors. They can target a precise capability and avoid artifacts in older datasets. Their quality depends on instructions, review, contributor consent, and disagreement handling. “Human-written” is not a licence; the publisher still needs permission to distribute the work.

Public-dataset tasks adapt records from an existing corpus. A science benchmark might select questions from an educational resource, while a coding benchmark might pair issue reports with patches. This route can provide realistic language and labels, but the benchmark inherits constraints from the source. Check whether selection, transformation, and redistribution are permitted and whether attribution must follow individual records.

Synthetic tasks are produced by a deterministic program, a model, or both. A committed seeded generator is especially useful: another team can rebuild the split, inspect assumptions, and create fresh variants. Synthetic does not mean automatically clean. A generator can repeat templates, leak the answer into the prompt, or encode one model's stylistic preferences.

Live or commissioned tasks come from new events or paid authors under an explicit agreement. They can resist memorization and better reflect current systems. They also require careful privacy review, contributor terms, and a repeatable process for refreshing the benchmark.

The ABI family illustrates several shapes without distributing third-party task text. [Benchmark Ranking Stability](#) varies seeds and prompts to study ranking movement, while [Benchmark Checkout Consent](#) uses simulated commerce state to test an agent decision.

Separate the assets

A benchmark is rarely one copyrighted object. Treat these as separate assets:

- the paper and documentation;
- runner and grader code;
- task prompts, images, audio, and labels;
- source-dataset records;
- model outputs and run logs;

- the compiled database or index.

A repository's MIT or Apache licence often covers its code, not an embedded dataset. A paper may be readable online under publisher terms while its supplementary data has a Creative Commons licence. A model card can describe a dataset without granting any right to copy it. Record a source and licence for each asset you plan to use.

Read the licence as an instruction

CC0 is a public-domain dedication intended to remove copyright and database-right restrictions to the extent possible. Creative Commons recommends it where broad data reuse is the goal. CC BY permits redistribution and adaptation with attribution. CC BY-SA adds a share-alike condition. Noncommercial and no-derivatives variants impose further limits and may not fit a commercial evaluation service or a transformed task set.

Software licences solve a different problem. MIT and Apache-2.0 are common for runners and graders. Apache-2.0 includes an express patent grant; MIT is shorter. Neither label should be copied from the code repository onto unrelated task data.

“Proprietary” can still mean usable if a contract or published terms grant the access you need. “Unknown” means the source did not state a usable licence, not that the data is free of rights. If a dataset card says only “research use,” read the actual terms before treating that phrase as permission.

This is practical reading guidance, not legal advice. When a high-value dataset has unclear ownership, ask counsel or the publisher rather than guessing.

Keep attribution and provenance

For every imported source, store its name, URL, licence identifier or exact terms link, retrieval date, and refresh cadence. If attribution is required, preserve the creator name and notice. Creative Commons' data guidance recommends machine-readable metadata including title, author, source link, licence link, attribution statement, and rights statement.

Provenance also makes debugging possible. If a task is challenged, you should be able to identify its source, transformation, and version without searching old chat logs. A stable source ID lets a refresh update the row instead of creating a duplicate.

What common licences usually allow

A quick triage can save time:

- **CC0:** broad reuse; attribution is courteous even when not legally required.
- **CC BY:** copy and adapt with credit and the required licence notice.
- **CC BY-SA:** copy and adapt with credit, plus share-alike duties for adaptations.
- **MIT or Apache-2.0:** generally suitable for software; verify whether data is included.
- **Proprietary or custom:** follow the stated contract and access limits.
- **No licence found:** do not redistribute the content; record metadata and seek permission.

Always check the exact version. SPDX identifiers such as CC-BY-4.0 and Apache-2.0 avoid vague labels like “Creative Commons.”

A safe import routine

Before a row lands, open the primary source. Confirm what the benchmark measures, where its tasks came from, and which licence covers the material you will store. Write the description in your own words. Import only short factual metadata unless the licence clearly permits the task content. Record unknowns as unknowns.

Then test the retrieval steps on a clean machine. A nominally open dataset that requires a dead script is not meaningfully reproducible. Keep a refresh note: rerun a seed script, fetch a tagged release, or request a new export. These small records turn “we found it online” into a chain another evaluator can audit.

Frequently asked questions

Is a benchmark open because its paper is public?

No. The paper, code, prompts, and dataset may each have different licences or no stated licence.

Can I copy data that has no licence?

Do not assume so. Ask the rights holder or use metadata only until permission is clear.

Sources

[Creative Commons licensing best practices for scientific data](https://creativecommons.org/2026/04/20/licensing-best-practices-for-the-sharing-of-scientific-data/) — <https://creativecommons.org/2026/04/20/licensing-best-practices-for-the-sharing-of-scientific-data/>

[Creative Commons public-domain guidance](https://creativecommons.org/public-domain/) — <https://creativecommons.org/public-domain/>