

math-vault: Curated, Traceable Snapshots of Public Mathematical Reasoning Datasets

Xinmu Ge

Shanghai Innovation Institute · Shanghai Jiao Tong University · g3ra1d@sjtu.edu.cn

Independent dataset documentation report · 31 August 2026 · Related dataset DOI:
10.5281/zenodo.21411214

Abstract. math-vault is a versioned snapshot of public mathematical-reasoning datasets assembled for traceable evaluation. The v0.1.0 dataset release contains 15 evaluation subsets and 32,226 parser-valid rows in a canonical JSONL interface with required `id`, `problem`, and `answer` fields. It maintains source, derived, and canonical layers rather than treating all copies as equivalent. Per-file parent pointers and recorded SHA256 hashes make materialization auditable. Nineteen Omni-MATH rows that the current parser cannot reliably process remain preserved as isolated audit issues and are excluded from the evaluation count. The canonical and DAPO dedup data-materialization scripts use only the Python standard library. Code and documentation are MIT-licensed; each dataset retains its upstream licensing terms.

1. Purpose and scope

Mathematical-reasoning evaluation often combines public datasets whose fields, provenance, and answer conventions differ. math-vault provides a frozen release that makes those distinctions explicit. It does not claim ownership of the upstream data or collapse their licenses into a single data license. Instead, it records snapshots and transformations, then supplies a common evaluation-facing representation for tools that need stable identifiers and gold answers.

The release is designed to feed math-eval [1] directly. This is an interface relationship: math-vault supplies evaluation rows and provenance, while math-eval provides generation and evaluation machinery. The OPD study [2] documents one downstream use case, and should not be read as an endorsement by any upstream dataset author.

15 evaluation subsets	32,226 parser-valid rows	19 isolated audit issues
--------------------------	-----------------------------	-----------------------------

2. Release structure

The frozen v0.1.0 dataset release distinguishes three layers. **source/** contains upstream artifact snapshots and permits only serialization changes that preserve rows, fields, ordering, and content. **derived/** holds transformations such as filtering, deduplication, prompt removal, or sampling, together with their parents and reconstruction rules. **canonical/** mechanically materializes data usable by math-eval.

Every canonical evaluation row has a non-empty unique string identifier, a problem, and an answer. Optional solution, source, and metadata fields are retained when needed for audit. `canonical/manifest.json` records the source name, parent path, recorded parent SHA256, recorded output SHA256, and row count for each materialized file.

Subset family	Evaluation rows
AIME, AMC, AIMO, BrUMO, CMIMC, HMMT	403
GSM8K	8,792
DAPO-Math-17K dedup	17,176
MATH-500, Minerva-Math	772
Omni-MATH, OlympiadBench	5,083

Table 1. Compact grouping of the 15 canonical subsets. The manifest is the authoritative per-file accounting.

3. Parser audit boundary

The Omni-MATH source snapshot has 4,428 rows. Of these, 4,409 are materialized as evaluation rows. Nineteen records are retained without alteration in `canonical/omni_math/issues.jsonl` because the current parser cannot reliably process their answers. The recorded categories are nine empty answers, six malformed LaTeX answers, one truncated answer, one unsupported piecewise-text answer, one verification error, and one row with multiple boxed answers. Keeping these rows separate makes the exclusion reviewable rather than silently repairing or discarding source data.

The release notes an audit baseline in which all 32,226 included canonical answers passed the current math-eval parser’s boxed strict/soft check consistently. That check describes compatibility with a named parser configuration; it is not a claim that every upstream answer is universally unambiguous.

4. Deterministic materialization

The canonical and DAPO dedup data-materialization scripts rely only on the Python standard library. The DAPO dedup script checks expected source and output row counts and identifier uniqueness. The canonical build checks row counts and canonical identifier uniqueness, then records parent and output SHA256 hashes in its manifest. This deliberately narrow implementation keeps the release inspectable: users can compare the manifest against the materialized files and their parents without relying on a hidden service or a third-party build framework.

For use, consumers can cite the versioned dataset artifact, inspect the relevant source or derived README, and load the canonical JSONL files into math-eval or another evaluator that understands the same fields. Consumers remain responsible for honoring the upstream terms attached to each dataset.

5. Licensing and provenance

Code and documentation in math-vault are released under the MIT License. The datasets are not relicensed: data remains subject to the license or terms documented for each upstream artifact, and an upstream absence of an explicit license is recorded as such rather than interpreted as public-domain permission. This boundary is central to the repository design: a uniform technical interface does not erase provenance or rights information.

How to cite

Dataset documentation report. Xinmu Ge. *math-vault: Curated, Traceable Snapshots of Public Mathematical Reasoning Datasets*. Independent dataset documentation report, 31 August 2026.

Dataset artifact. Xinmu Ge. *math-vault: Curated, Traceable Snapshots of Public Mathematical Reasoning Datasets*. Dataset v0.1.0, 2026. Related dataset DOI: 10.5281/zenodo.21411214.

References

- [1] Xinmu Ge. *math-eval: Reproducible Mathematical Reasoning Generation and Evaluation*. Software, v0.1.0, 2026. DOI: 10.5281/zenodo.21411208.
- [2] Xinmu Ge et al. *Towards Understanding On-Policy Distillation through the Lens of Test-Time Scaling*. arXiv:2608.11829, 2026. <https://arxiv.org/abs/2608.11829>.