

# Mercury Measurement Ledger and Result Governance

Chen, Ho Yiing (norika)<sup>1</sup>

<sup>1</sup>Independent Researcher, Taiwan, ORCID: [0009-0006-6816-9891](https://orcid.org/0009-0006-6816-9891), norika at  
charenix.com

Draft: June 18, 2026

## Abstract

Large interpretability projects can fail without a single mathematical error: they fail when partial runs are remembered as complete runs, when old parameters are mixed with new criteria, and when beautiful figures hide missing coverage. Mercury produced visual atlases for 38 model towers and recovered 51 traceable model/task result records, but not every record has the same evidentiary status. This paper formalizes a measurement ledger that separates formal fullgrid passes, partial or legacy runs, visual-only atlases, and failed attempts.

The core rule is severe: a result is not a result until its method, coverage, cell count, artifact hash, and acceptance threshold are written down. I define run admissibility, threshold validity, provenance completeness, archive safety, and paper-claim eligibility. The goal is not bureaucracy. The goal is to stop a research program from accidentally lying to itself.

## Author Note

This paper exists because I got angry at the number nineteen. The project had more than that: 38 towers in the city, 51 traceable result records, 34 formal fullgrid passes, and many partial runs that still matter. The mistake was not that the data was small. The mistake was treating unlike evidence as if it had one name.

That cannot happen again, because a count is never neutral once it enters a paper.

## 1 A Count Is A Claim

Thirty-eight answers the city question. Thirty-four answers the formal-gate question. Fifty-one answers the recovered-record question. A clean paper must say which question it is answering before it says the number.

## 2 The Ledger Problem

Interpretability work often talks about models, probes, layers, or neurons. It talks less about record keeping. That sounds boring until the archive is large enough that a partial shard, a legacy parameter run, and a formal pass can all look like “one measured model” in memory.

Mercury needs a ledger because the visual city is too large to trust informal memory. A tower may exist as an atlas, a dense pass, a head-only pass, a failed run, or a recovered backup. Those states are different, and a paper must not collapse them.

### 3 Run Object

Let a run be

$$r = (m, \theta, p, a, c, \gamma, h, t),$$

where  $m$  is the model or task name,  $\theta$  is the method configuration,  $p$  is the prompt or task family,  $a$  is the artifact path,  $c$  is the observed cell count,  $\gamma$  is coverage,  $h$  is an artifact hash, and  $t$  is a timestamp.

A run has no formal status until the tuple is recoverable. The existence of a folder is not enough. The existence of a screenshot is not enough either.

### 4 Admissibility

For a Mercury fullgrid result, define

$$Q(r) = \mathbf{1}[c(r) \geq C_{\min}] \mathbf{1}[\gamma(r) \geq \Gamma_{\min}] \mathbf{1}[h(r) \neq \emptyset] \mathbf{1}[\theta(r) \in \Theta_{\text{formal}}].$$

In the 2026 fullgrid criterion used here,

$$C_{\min} = 100,000,000, \quad \Gamma_{\min} = 0.90.$$

If  $Q(r) = 1$ , the run is admissible for formal coverage claims. If  $Q(r) = 0$ , the run may still be useful, but it must be labeled as partial, legacy, failed, visual-only, or diagnostic.

This is the line that prevents self-deception.

The run can still teach us something after it fails the gate. It just cannot wear the badge of a formal result.

The downgrade rule is part of the method. A partial run can support a caution, a failed-method comparison, or a historical note, but it cannot support a claim that says Mercury has fully measured a model under the current gate. That distinction sounds small until the archive becomes large enough that memory starts smoothing every run into the same story.

### 5 Claim Eligibility

Let a paper claim  $x$  depend on a set of runs  $R_x$ . The eligibility score is

$$E(x) = \frac{1}{|R_x|} \sum_{r \in R_x} Q(r).$$

The claim is formal only when

$$E(x) = 1.$$

If  $0 < E(x) < 1$ , the claim is mixed evidence. If  $E(x) = 0$ , it can only be a method note, failure report, visualization note, or hypothesis.

This rule would have prevented the nineteen-versus-thirty-eight confusion. The count depends on the claim being made.

### 6 Four Counts, Four Meanings

The Mercury archive needs at least four public counts:

$$N_{\text{atlas}}, \quad N_{\text{trace}}, \quad N_{\text{formal}}, \quad N_{\text{legacy}}.$$

In the recovered 2026 ledger, the working counts are

$$N_{\text{atlas}} = 38, \quad N_{\text{trace}} = 51, \quad N_{\text{formal}} = 34.$$

These numbers are not interchangeable. The atlas count answers “how many towers can be shown.” The trace count answers “how many records can be tied to artifacts.” The formal count answers “how many records satisfy the current fullgrid gate.”

Mixing those numbers damages both science and trust.

The ledger keeps the argument from getting drunk on its own archive.

## 7 Provenance Completeness

For each run, define provenance completeness:

$$P(r) = \frac{\mathbf{1}[a] + \mathbf{1}[h] + \mathbf{1}[\theta] + \mathbf{1}[c] + \mathbf{1}[\gamma] + \mathbf{1}[t]}{6}.$$

If  $P(r) < 1$ , the run should stay in the ledger but lose claim power:

$$Q'(r) = Q(r)P(r).$$

This weighted status is useful for old archives. It lets recovered data remain visible without pretending that damaged records are pristine.

## 8 Archive Safety

A run that exists only on a borrowed machine is not safe. Let

$$B(r) = \mathbf{1}[\text{local copy}] + \mathbf{1}[\text{external drive copy}] + \mathbf{1}[\text{remote copy}].$$

The archive safety score is

$$S(r) = \min\left(1, \frac{B(r)}{2}\right).$$

At least two locations are required because one disk, one VPS, or one borrowed GPU workstation can disappear. This is not paranoia. It is the normal failure mode of research done under pressure.

The data must survive the machine. If a result cannot survive a reset, it is not yet an asset.

## 9 Why Existing Practice Is Not Enough

Model cards describe intended use and evaluation context [1]. Datasheets for datasets describe dataset motivation, collection, composition, and maintenance [2]. Data statements add linguistic and demographic accountability for NLP resources [3]. Checklists for machine learning reproducibility push authors to report experimental details [4].

Mercury needs those ideas, but at the run ledger level. The project does not only release a dataset or a model. It releases a city of measurements, failed attempts, visual atlases, restored archives, and formal gates. The ledger is the thing that keeps those states from collapsing into one flattering number.

## 10 Falsification

The ledger can be audited. Pick a claim  $x$ , list the supporting runs  $R_x$ , and recompute  $E(x)$ . Then pick each  $r \in R_x$ , verify the artifact hash, cell count, coverage, and method configuration. If any term is missing, the claim must be downgraded.

A reviewer does not need to trust the author here. The formula tells the reviewer where to poke.

## 11 Limits

A ledger does not make weak measurements strong. It also does not rescue a bad method. What it does is more basic: it prevents the archive from being rewritten by mood, exhaustion, or memory.

That basic work matters. Without it, even a beautiful LLM city can become a rumor about itself.

## 12 Conclusion

Mercury Measurement Ledger and Result Governance is the guardrail under the city. The visual work can be spectacular, but every tower still needs a record, a threshold, a hash, a status, and a backup. If the evidence is partial, say partial. If it is formal, show the gate. The city becomes stronger when it stops trying to look cleaner than the work really was.

## References

- [1] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019. <https://arxiv.org/abs/1810.03993>
- [2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. Datasheets for Datasets. *Communications of the ACM*, 2021. <https://arxiv.org/abs/1803.09010>
- [3] Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 2018. <https://aclanthology.org/Q18-1041/>
- [4] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d’Alche Buc, Emily Fox, and Hugo Larochelle. Improving Reproducibility in Machine Learning Research: A Report from the NeurIPS 2019 Reproducibility Program. *Journal of Machine Learning Research*, 2021. <https://www.jmlr.org/papers/v22/20-303.html>