Experimental Attempts in Electronic Lab Notebooks: A Dataset Proposal for Scientific Debugging

Dmitrii Magas [1]

Abstract

Modern science has a publication bias: successful protocols, curated papers, and cleaned-up scientific narratives are often overrepresented. However, scientific AI agents must also reason about failed runs, deviations, and troubleshooting under incomplete and noisy evidence. Such events are often recorded in laboratory notebooks, yet public datasets of notebook-derived experimental attempts remain scarce despite the growth of electronic laboratory notebooks (ELNs) and open-notebook practices. We propose a dataset of run-level experimental attempt records with fields such as goal, hypothesis when available, run text, observations, status, deviation, response, and evidence spans for inferred labels.

1. Motivation

Public AI-for-science resources are skewed toward intended procedures and successful outcomes. This misses a central part of real experimental practice: diagnosing failed runs, tracing deviations, and deciding what intervention to try next. Recent real-world evidence makes this gap more concrete: in a randomized controlled trial on novice biology tasks, LLM assistance did not significantly improve the primary workflow-completion endpoint, despite modest gains on some intermediate measures (Hong et al., 2026). This is consistent with the view that practical lab assistance may depend not only on protocol recall, but also on debugging under incomplete and noisy evidence.

One promising response is to build a dataset of notebook-derived experimental attempt records that captures not only intended procedures, but also failed runs, deviations, observations, and troubleshooting responses . Laboratory notebooks are a core instrument of experimental science: they preserve what was attempted, what was observed, and how results were interpreted, supporting verification, reuse, and reproducibility (Nussbeck et al., 2014). Historically this role was served by paper notebooks, but modern research increasingly depends on digital files, software-driven analyses, and linked metadata, motivating electronic laboratory notebooks (ELNs) and related systems (Tremouilhac et al., 2017). Open-notebook science is one strand of the broader open-science movement (Bradley, 2007; Clinio & Albagli, 2017). It is conceptually well aligned with this proposal because it emphasizes sharing primary research records, including intermediate results and unsuccessful experiments. However, the literature suggests that open-notebook practice remains relatively niche rather than mainstream, with dedicated community efforts still limited in scale (Schapira et al., 2019; Harding, 2019).

2. Gap in Existing Scientific Data

We have already assembled pilot corpora from public scientific sources spanning notebook-like records, protocols, and reactions:

- **iGEM records** (iGEM Foundation): 4,000 lab notebooks and experiment details from synthetic biology projects with an open notebook publication model. These are closest to the desired dataset, but unstable formatting makes large-scale extraction difficult.
- **18,000 detailed protocols from protocols.io** (Teytelman et al., 2016): protocols from multiple scientific domains and levels of complexity, but centered on intended procedures rather than failure repair.
- **3,100 chemistry reactions from the Chemotion Repository** (Tremouilhac et al., 2020): step-by-step reaction procedures, but primarily successful records without troubleshooting or deviation notes.

These pilots suggest that public notebook-like sources can be aggregated at useful scale, but they also highlight a persistent gap: existing resources mainly capture successful procedures rather than experimental attempts with explicit failures, deviations, and follow-up actions. This is the level of record needed for training and evaluating scientific agents that must recognize failures, reason about likely causes, and respond appropriately under uncertainty.

[1]Independent Researcher, Berlin, Germany. Correspondence to: Dmitrii Magas <dmitrii@eamag.me>.

Accepted by ICML 2026 AI for Science Workshop.

3. Proposed Dataset

The target unit is a **run-level experimental attempt record**: a structured representation of one experimental run or troubleshooting episode. Each record links an intended goal or hypothesis to the executed run text, observations, outcome, and any follow-up response.

Each record will contain a compact set of fields: scientific domain, source type, goal, hypothesis when available, procedure or run text, observations, run status, deviation or anomaly, response action, and evidence spans supporting any inferred label. The dataset is intended to aggregate heterogeneous notebook-derived records into a common format that supports scientific debugging, retrieval, and evaluation.

4. Acquisition Roadmap

We will focus on an export-and-mapping pipeline for existing ELNs and notebook-like records rather than requiring contributors to adopt a new notebook template. The proposed tooling will support extraction from heterogeneous sources, mapping of native fields into a shared schema, and packaging into an interoperable release format.

Annotation will be layered on top of this export step. We will use LLM-assisted pre-annotation to suggest fields such as run status, deviation type, response action, and supporting evidence spans, followed by targeted human verification for quality control. This design reduces contributor burden, makes collection feasible across different local ELN systems, and creates a reusable path for future partner onboarding.

5. Metadata and Governance

The schema will specify field definitions, allowed values, and annotation provenance for all core variables, including status labels (*success*, *failure*, *partial*, *unclear*) and troubleshooting labels for detection, diagnosis, intervention, and outcome. Each annotation will record whether it was directly sourced, inferred, or LLM-suggested and later verified.

To support responsible sharing, records will also include provenance, licensing, and release constraints. We will support tiered release modes ranging from fully open text to restricted or metadata-only records, with de-identification, redaction of unpublished or hazardous procedural details, and partner-specific sharing controls where needed.

6. Acceleration Potential

A corpus of experimental attempts would support agents that detect when a run is drifting from the intended plan, identify the relevant deviation, diagnose plausible causes from partial evidence, propose a next intervention, and decide when uncertainty is high enough to ask a human or gather more evidence. These capabilities are difficult to learn from protocol collections alone because successful procedures rarely expose the reasoning required after a failed or ambiguous run.

Although our initial pilots are most likely to come from chemistry and biology, the schema is intended to generalize across experimental sciences wherever work is recorded as attempts, observations, deviations, and follow-up actions, including materials, engineering, and other laboratory workflows. This broader framing increases the value of the resource for multi-domain scientific assistants rather than domain-specific copilots alone.

The same resource can also serve as an evaluation benchmark for scientific debugging. Models could be tested on predicting run status, identifying the relevant deviation, ranking plausible causes, choosing a corrective action, or reconstructing the next experimental step from local notebook context. Such evaluations may be especially useful in educational, open-science, and resource-constrained settings, where better support for troubleshooting can lower the cost of failed runs and reduce dependence on constant expert supervision.

Impact Statement

This proposal aims to improve the realism and safety of AI-for-science by shifting attention from idealized successful procedures toward failed runs, deviations, and debugging. Such data could support more reliable scientific assistants and more informative evaluation. At the same time, laboratory records may contain sensitive procedural details, unpublished findings, or domain-specific hazards. Any resulting dataset should therefore use governed release, de-identification where necessary, and domain-appropriate screening before public distribution.

References

Bradley, J.-C. Open notebook science using blogs and wikis. *Nature Precedings*, 2007. doi: 10.1038/npre.2007.39.1.

Clinio, A. and Albagli, S. Open notebook science as an emerging epistemic culture within the open science movement. *Revue française des sciences de l'information et de la communication*, (11), 2017.

Harding, R. J. Open notebook science can maximize impact for rare disease projects. *PLOS Biology*, 17(1):e3000120, 2019. doi: 10.1371/journal.pbio.3000120.

Hong, S. Z., Kleinman, A., Mathiowetz, A., Howes, A., Cohen, J., Ganta, S., Letizia, A., Liao, D., Pahari, D.,

Roberts-Gaal, X., Righetti, L., and Torres, J. Measuring mid-2025 LLM-assistance on novice performance in biology. *arXiv*, 2026. doi: 10.48550/arXiv.2602.16703.

iGEM Foundation. Team wiki. https://competition.igem.org/deliverables/team-wiki. Accessed 2026-04-15.

Nussbeck, S. Y., Weil, P., Menzel, J., Marzec, B., Lorberg, K., and Schwappach, B. The laboratory notebook in the 21st century: The electronic laboratory notebook would enhance good scientific practice and increase research productivity. *EMBO Reports*, 15(6):631–634, 2014. doi: 10.15252/embr.201338358.

Schapira, M., Consortium, T. O. L. N., and Harding, R. J. Open laboratory notebooks: good for science, good for society, good for scientists [version 2; peer review: 2 approved, 1 approved with reservations]. *F1000Research*, 8:87, 2019. doi: 10.12688/f1000research.17710.2.

Teytelman, L., Stoliartchouk, A., Kindler, L., and Hurwitz, B. L. Protocols.io: Virtual communities for protocol development and discussion. *PLOS Biology*, 14(8): e1002538, 2016. doi: 10.1371/journal.pbio.1002538.

Tremouilhac, P., Nguyen, A., Huang, Y.-C., et al. Chemotion ELN: an open source electronic lab notebook for chemists in academia. *Journal of Cheminformatics*, 9(1): 54, 2017. doi: 10.1186/s13321-017-0240-0.

Tremouilhac, P., Lin, C.-L., Huang, P.-C., Huang, Y.-C., Nguyen, A., et al. The repository Chemotion: Infrastructure for sustainable research in chemistry. *Angewandte Chemie International Edition*, 59(50):22771–22778, 2020. doi: 10.1002/anie.202007702.