Artifact Inflation Without Epistemic Grounding:
Failure-Boundary Case Studies in AI-Assisted Non-Expert Research

Keiji Yoshimura
Independent Researcher

Version v0.1.2-ai-vixra-final – 2026-06-01

Companion repository: https://github.com/yokken0907/ai-artifact-inflation-claim-boundary-case-study

Abstract

Generative AI enables non-experts to rapidly produce papers, code, figures, repositories, quality-assurance documents, and submission metadata. This capacity can widen access to knowledge production, but it also creates a risk that artifact completeness is mistaken for domain validity. This manuscript presents a reflexive case study of two AI-assisted research attempts conducted by the author: a lithium-ion battery-model audit and an A10-derived Hubble-tension scaffold audit. In the battery case, an initially technology-oriented project was reduced to a residual-drift audit of simplified PyBaMM models against external discharge time series. Claims about safe charging, practical battery-management-system readiness, degradation reduction, lithium-plating avoidance, and real-cell validation were explicitly prohibited. In the Hubble case, a cosmological scaffold failed to establish a first-principles physical $E(z)/H(z)$ bridge and therefore stopped before real-data likelihoods, MCMC, posterior comparison, or evidence evaluation. Across two distinct domains, the same pattern appeared: AI inflated research artifacts, but strict grounding checks reduced them to claim-boundary and failure-boundary records. This manuscript does not argue that AI-assisted research is useless, nor does it evaluate expert-led AI science. It argues that non-expert AI-assisted research requires explicit domain-grounding, prior-art, external-data, implementation-or-experiment, expert-review-worthiness, and public-claim-boundary gates before public claims are made.

Keywords: generative AI; non-expert research; AI-assisted science; hallucination; claim boundary; failure-boundary archive; epistemic grounding; PyBaMM; Hubble tension

1 Introduction

Generative AI can now produce much of the visible surface of research: paper outlines, code, figures, tables, repository structures, quality-assurance notes, simulated reviews, and submission metadata. Automated research pipelines have already been proposed for parts of the scientific workflow, especially in computationally convenient machine-learning settings[1, 2]. Large language models are also widely discussed as tools for scientific discovery, literature handling, coding support, and hypothesis generation[3]. At the same time, hallucination remains a persistent reliability problem: language models can generate plausible but unsupported statements, and evaluation incentives may reward guessing rather than acknowledging uncertainty[4, 5].

The question in this manuscript is not whether AI can assist research. It can. The narrower question is what happens when a non-expert uses generative AI to build artifacts that resemble technical or scientific research in domains where professional grounding, experimental access, observational mapping, or implementation responsibility is missing.

Research question. When a non-expert uses generative AI to construct domain-specific research artifacts, how far does artifact completeness support domain validity, and where does it become artifact inflation without epistemic grounding?

The author has used generative AI to construct several research-like projects across cosmology, industrial technology, battery modelling, domestic-equipment noise, resource processing, and related topics. In several cases, the initial theoretical ambition was later reduced by validation attempts, claim-boundary review, and external-data checks. The present manuscript focuses on two representative cases: (1) a lithium-ion battery-model audit and (2) an A10-derived Hubble-tension scaffold audit.

The manuscript itself is AI-assisted. This is not incidental. The paper is therefore not an external criticism of AI-assisted research from a position outside the phenomenon. It is a reflexive record from inside the process: AI was used to generate research artifacts, those artifacts were later found to lack sufficient grounding for their initial ambitions, and AI is again used here to document the boundary failure. This recursive structure is unusual, but it is also the case material.

2 Background: AI-Assisted Research and Artifact Inflation

2.1 Automated and AI-assisted scientific workflows

Recent work has explored systems that generate research ideas, write code, run experiments, create figures, draft papers, and simulate review. The AI Scientist framework, for example, is presented as an end-to-end system for automated open-ended scientific discovery in machine-learning subfields[1]; later work extends this direction with agentic search and workshop-level paper generation[2]. These systems show that AI can participate not only in isolated writing tasks but in the production of complete research artifacts.

However, a complete artifact is not the same as validated knowledge. A paper-shaped object may contain correct-looking sections, references, equations, plots, and code while still lacking the connection to domain-specific assumptions that would make its central claim meaningful. This risk is amplified when the user is not able to independently audit the relevant domain.

2.2 Hallucination and epistemic grounding

Large language model hallucination is commonly defined as the generation of plausible but nonfactual or unsupported content[4]. In technical work, the risk is not limited to single false sentences. Unsupported content can be distributed across a whole artifact: an equation may be formally written but physically ungrounded; a simulation may run but not answer the claimed question; a repository may pass internal checks while failing to connect to real experimental or observational standards.

In this manuscript, *artifact inflation without epistemic grounding* refers to a process in which papers, repositories, figures, code, tables, manifests, review notes, and submission

metadata grow in apparent completeness, while the connection to domain knowledge, experimental access, observational likelihoods, implementation constraints, safety responsibility, and prior-art differences remains insufficient. The phrase does not mean that every AI-generated artifact is invalid. It identifies a specific failure mode in non-expert AI-assisted research.

3 Method: Reflexive Case Study

This manuscript is not a statistical study. It does not claim to generalize over all AI-assisted research. It examines two author-generated case archives and asks how each artifact moved from an initial ambitious framing to a bounded failure or pre-validation record.

The scope is limited as follows:

- The manuscript does not prove that AI-assisted research is impossible.
- It does not evaluate expert-led AI science, industrial research, laboratory research, or peer-reviewed AI-assisted research.
- It does not claim a practical battery technology, a battery-management-system method, a cosmological solution, or an observational result.
- It uses two self-produced AI-assisted archives as reflexive case materials.
- It proposes stopping gates for non-expert AI-assisted research before public technical claims are made.

The source materials were the author's own repository packages, including README files, claim-boundary notes, phase summaries, expert-handoff notes, and terminal decision records. These materials are not treated as independent external evidence of technical validity. They are treated as internal records of how the projects were constructed, constrained, and eventually downgraded.

4 Case 1: Lithium-Ion Battery-Model Audit

4.1 Initial framing

The first case began as an industrial-technology-oriented attempt concerning lithium-ion battery charging constraints, battery management, degradation, thermal limits, and feasibility boundaries. As the project developed, however, the final repository boundary reclassified it as a pre-validation and expert-review package rather than a battery technology.

The technical tools and data background were ordinary rather than proprietary. The project used simplified lithium-ion battery modelling through PyBaMM, an open-source framework for fast and flexible simulation of battery models[6]. The external data context was NASA's Li-ion Battery Aging Dataset, which includes charge, discharge, and electrochemical impedance spectroscopy profiles collected under different temperatures at the NASA Ames Prognostics Center of Excellence[7].

4.2 Final state: reduction to residual-drift audit

The final claim boundary allowed only modest statements. A reproducible computational workflow had been prepared; simplified PyBaMM SPM/SPMe simulations had been compared with external discharge time-series records; residual structures and residual drift across capacity or cycle-stage proxies had been audited; and model or parameter-set sensitivities had been recorded.

The prohibited claims were more important. The case did not support safe charging, practical BMS readiness, EV or automotive safety, degradation reduction, lithium-plating avoidance, SEI validation, lifetime extension, parameter-set correctness, real-cell validation, or superiority over existing battery-management protocols. The work therefore did not remain a technology proposal. It became a diagnostic record describing where a technology proposal could not yet be made.

In the final phase summary, model ranking changed depending on which quantity was emphasized: profile RMSE, late-region residual, endpoint or offset effects, thermal gap, or capacity-stage drift. Some earlier shape-oriented indicators made one model and parameter set appear preferable, while later residual-drift indicators changed the apparent balance. The result was not a stable battery method. It was an audit showing that the apparent conclusion was sensitive to the metric and that public technology claims should stop.

4.3 Meaning of the case

The battery case is not a battery-engineering result. It is a case in which generative AI helped produce a battery-technology-shaped artifact that, under stricter questioning, had to be reduced to a residual-drift audit of simplified models against external discharge data. The residual value is not a charging method but a stopping line: a record that indicates what could not be claimed.

This is a failure if judged by the initial ambition. It is also a partial success as claim-boundary control. Without the boundary process, the artifact could have expanded into unsafe claims about charging, degradation, or BMS performance. The final value lies in refusing those claims.

5 Case 2: A10-Derived Hubble-Tension Scaffold Audit

5.1 Initial framing

The second case examined whether an A10-derived cosmological scaffold could contribute to the Hubble tension. The Hubble tension is a major cosmological discrepancy between early-universe inferences under the standard cosmological model and late-universe local distance-ladder measurements. Planck 2018 reports precise cosmological parameters from cosmic microwave background observations under Lambda-CDM[8]; SH0ES reports a local Hubble constant measurement based on HST observations and distance-ladder calibration[9]. Reviews emphasize that proposed solutions must satisfy multiple observational and theoretical constraints[10].

The project initially carried the possibility of being a cosmological explanatory scaffold. Through successive audit phases, however, it was downgraded to a failure-boundary package.

5.2 Final state: no physical $E(z)$/$H(z)$ bridge

The terminal repository boundary states that the package is not a Hubble-tension solution, not a Lambda-CDM replacement, and not a report of real-data likelihoods, MCMC posteriors, evidence values, or validated cosmological predictions. The central terminal finding was the absence of a first-principles physical background expansion equation.

> The audited chain retained candidate proxies, internal diagnostic curves, contract-only scaffolds, and bounded hypothesis records. However, within the audited artifacts, no first-principles physical $E(z)$ or $H(z)$ bridge was established. Therefore the project stopped before real-data likelihoods, posterior comparison, evidence evaluation, or public cosmological solution claims.

The final decision matrix classified candidate proxies and contract-only bridge scaffolds as retained, but first-principles physical $E(z)$ derivation as not found, physical $H(z) = H_0E(z)$ as not established, real-data likelihood and MCMC/posterior/evidence as no-go, and public Hubble-solution claims as forbidden.

5.3 Meaning of the case

The Hubble case is not a cosmological result. It is a case in which generative AI helped build a theory-like scaffold, audit logs, decision matrices, expert-handoff notes, paper drafts, and repository materials, but the work could not establish the physical bridge required to enter observational cosmology. The final value is failure localization: the missing physical $E(z)$/$H(z)$ bridge was identified as the point at which the theory could not proceed.

Again, this is a failure relative to initial ambition. It is also an example of claim-boundary recovery. Instead of presenting a Hubble-tension solution, the project was reclassified as a bounded failure-boundary archive.

6 Cross-Case Pattern

The two cases differ strongly in domain. The first concerns industrial battery modelling and practical engineering claims. The second concerns cosmological modelling and observational-theory connection. Yet the failure pattern was similar.

Table 1: Cross-case correspondence

Item	Battery-model audit	Hubble-tension scaffold audit
Initial ambition	Industrial theory around charging constraints, BMS, degradation, and feasibility	Cosmological scaffold or Hubble-tension explanatory candidate
Artifact surface	Paper, PyBaMM comparisons, residual tables, figures, phase logs, expert dossier	Paper, audit logs, phase chain, decision matrix, expert handoff
Grounding barrier	Real-cell validation, calibration, BMS implementation, safety, degradation mechanism	Physical $E(z)$/$H(z)$ equation, observational mapping, real likelihood, model comparison

Item	Battery-model audit	Hubble-tension scaffold audit
Final state	Residual-drift audit of simplified models against external discharge data	Failure-boundary audit after physical bridge was not established
Forbidden claims	Safe charging, practical BMS, degradation reduction, real-cell validation	Hubble solution, Lambda-CDM replacement, real-data support, posterior/evidence claim
Residual value	Stopping line before technology claims and specialist-review questions	Localized failure boundary and specialist-handoff question

The observed sequence can be summarized as follows:

1. **Rapid artifact formation.** Generative AI helped produce theory names, paper sections, code, figures, repository structures, manifests, QA notes, and submission material.
2. **Plausibility through domain vocabulary.** The artifacts contained relevant terms, models, metrics, equations, and procedural language.
3. **Grounding deficit.** When judged against experimental, observational, implementation, or physical requirements, the central claims were unsupported.
4. **Boundary reduction.** The artifacts were not validated as technologies or theories; they were reclassified as claim-boundary and failure-boundary records.

This is the pattern named in this manuscript as artifact inflation without epistemic grounding.

7 Proposed Stopping Gates for Non-Expert AI-Assisted Research

The cases suggest that non-expert AI-assisted research should include explicit stopping gates before publication or public claim-making.

7.1 Domain-grounding gate

The work must connect to standard domain problems, variables, assumptions, measurement conditions, and responsibility structures. Use of domain vocabulary is not sufficient.

7.2 Prior-art difference gate

The work must state how it differs from existing research. If it only reuses existing tools or public data in a basic way, it should be labelled as reproduction, learning, audit, or preliminary organization rather than as a new technology or theory.

7.3 External-data gate

If external data are used, the origin, license, preprocessing, selection effects, measurement conditions, and redistribution boundaries must be stated. Derived CSVs, figures, or metrics cannot substitute for data provenance.

7.4 Implementation-or-experiment gate

For industrial technology, the work must connect to implementation, materials, manufacturing, safety, control, regulation, cost, and responsible operators before practical claims are made. For basic science, it must connect to physical equations, observables, likelihoods, and comparison models before theory claims are made.

7.5 Expert-review-worthiness gate

The work must be compressed into questions that a specialist can answer. "Is this theory correct?" is usually too broad. "Is this residual decomposition meaningful?" or "Can this scaffold be reformulated as a physical background equation?" is more reviewable.

7.6 Public-claim-boundary gate

Before public release, permitted and forbidden claims must be written explicitly. In AI-assisted work, artifact completeness can create pressure toward overclaiming. A claim-boundary document should therefore be treated as part of the artifact, not as an optional disclaimer.

Table 2: Stopping gates and failure handling

Gate	Required state	If the gate fails
Domain grounding	Connection to standard variables, assumptions, and responsibilities	Downgrade to learning note or preliminary organization
Prior-art difference	Clear difference from existing work	Downgrade to reproduction, review, or audit
External data	Clear provenance, preprocessing, and boundary	Stop as data-audit-preliminary material
Implementation or experiment	Practical or observational connection is possible	Forbid technical or scientific claims
Expert-review worthiness	Narrow questions a specialist can answer	Keep as internal archive or reformulate
Public claim boundary	Permitted and forbidden claims are explicit	Delay release or reclassify artifact

8 Discussion

8.1 Not an argument against AI

The cases do not show that AI-assisted research is worthless. On the contrary, AI made it possible to build and then audit artifacts that would otherwise have remained vague intu-

itions. It helped formulate claim boundaries, generate reproducible structures, identify missing bridges, and produce handoff questions.

The problem is not that AI produced nothing. The problem is that it produced a great deal. The abundance of papers, tables, code, and repositories can become misleading when the user lacks the expertise to distinguish artifact completeness from domain validity. In such cases, the most important research act may be stopping.

8.2 Failure-boundary archives

Immature AI-assisted research can be harmful if it is released as a technology, scientific solution, or validated result. However, deleting all failed artifacts may also erase information about where overclaiming began and where it was stopped. A failure-boundary archive is a middle category. It does not present failure as success. It records the failed premise, the missing bridge, the forbidden claims, and the narrow question that would remain for expert review.

The battery case and Hubble case both ended in such archives. Their value is not that they solve battery engineering or cosmology. Their value is that they show how AI-assisted artifacts can be downgraded before they become public overclaims.

8.3 Limitations

The sample size is two. Both cases were produced by the same author using similar AI-assisted workflows. The author is a non-expert in both domains. The cases have not undergone specialist peer review. Intermediate artifacts include both AI outputs and human decisions, and the separation between them is imperfect.

Therefore, the manuscript should not be read as a general theory of AI-assisted research. It is a bounded reflexive case report. Its contribution is descriptive and methodological: it identifies a recurring failure pattern in two author-generated cases and proposes gates that may help prevent similar overclaiming.

9 Conclusion

This manuscript examined two AI-assisted non-expert research attempts. The lithium-ion battery project was reduced from an initially technology-oriented ambition to a residual-drift audit of simplified models against external discharge data. The A10-derived Hubble-tension project was reduced from a cosmological scaffold to a failure-boundary archive after a physical $E(z)/H(z)$ bridge was not established.

Across both cases, generative AI rapidly inflated the surface of research: papers, figures, code, repositories, manifests, QA notes, and handoff materials. However, when the artifacts were checked against domain grounding, prior-art difference, external data, implementation or observational requirements, and expert-review-worthiness, the central claims could not proceed. The artifacts became boundary records rather than technologies or theories.

The appropriate response is neither blanket rejection of AI nor unbounded optimism. The appropriate response is gatekeeping at the level of claims. AI can accelerate intellectual work and produce impressive artifact surfaces. Precisely for that reason, non-expert AI-assisted research needs explicit mechanisms for saying what is not known, what is not validated, and what must not be claimed.

Acknowledgements and AI-use statement

The author acknowledges extensive assistance from generative AI systems in drafting, restructuring, translating, coding support, claim-boundary formulation, and document preparation. The author remains responsible for the final wording, claim boundaries, and the decision to publish or archive this manuscript. This manuscript itself is part of the reflexive case context it analyzes: it is an AI-assisted reflection on prior AI-assisted research attempts.

Conflict of interest

The author declares no institutional, corporate, or financial conflict of interest related to the technical validity of the two case studies. No practical battery technology, battery-management system, cosmological solution, or observational cosmology result is claimed.

Data and case-material boundary

The case materials are the author's own AI-assisted research archives: a battery-model expert-review dossier and an A10-derived Hubble-tension scaffold failure-boundary audit. Raw external battery data are not redistributed in this manuscript package. This paper summarizes internal documentation rather than introducing new battery experiments, real-cell validation, cosmological likelihood analyses, or MCMC results. A companion repository is identified for public archiving of the manuscript package and claim-boundary documents: https://github.com/yokken0907/ai-artifact-inflation-claim-boundary-case-study.

References

[1] Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. (2024). The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292.

[2] Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., and Ha, D. (2025). The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. arXiv:2504.08066.

[3] Zheng, T. et al. (2025). A Survey on Large Language Models in Scientific Discovery. Proceedings of EMNLP 2025, Main Conference.

[4] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Computing Surveys. DOI: 10.1145/3703155.

[5] Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. (2025). Why Language Models Hallucinate. arXiv:2509.04664.

[6] Sulzer, V., Marquis, S. G., Timms, R., Robinson, M., and Chapman, S. J. (2021). Python Battery Mathematical Modelling (PyBaMM). Journal of Open Research Software, 9(1), 14. DOI: 10.5334/jors.309.

[7] NASA Open Data Portal. Li-ion Battery Aging Datasets. NASA Ames Prognostics Center of Excellence.

[8] Planck Collaboration. Aghanim, N. et al. (2020). Planck 2018 results. VI. Cosmological parameters. Astronomy & Astrophysics, 641, A6. DOI: 10.1051/0004-6361/201833910.

[9] Riess, A. G. et al. (2022). A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km s^{-1} Mpc^{-1} Uncertainty from the Hubble Space Telescope and the SH0ES Team. The Astrophysical Journal Letters, 934, L7. DOI: 10.3847/2041-8213/ac5c5b.

[10] Di Valentino, E. et al. (2021). In the Realm of the Hubble Tension: A Review of Solutions. Classical and Quantum Gravity, 38, 153001. DOI: 10.1088/1361-6382/ac086d.

[11] Yoshimura, K. (2026). Battery Constraint Frontier Expert-Review Dossier. Author-supplied internal case archive, v0.1.0.

[12] Yoshimura, K. (2026). A10-Derived H0 Scaffold Failure-Boundary Audit. Author-supplied internal case archive, v0.1.0.

A Case Evidence Map

The following case-file categories were used to derive the two case summaries. They are not treated as independent external evidence of technical validity. They are the author's own project artifacts and are used as reflexive case materials.

Case	Key internal files
Battery-model audit	README; claim-boundary file; Phase 17 expert-review summary; specialist handoff note; Phase 16 findings appendix
Hubble-tension scaffold audit	README; claim-boundary file; terminal decision matrix; expert-handoff one-page summary