What the Records Don't Carry: A Position on Researcher-AI Co-Adaptation in Exploratory Laboratory Research

Zeyu Li [1] **Ziang Liu** [2] **Xiaoman Yang** [3] **Dan Luo** [1]

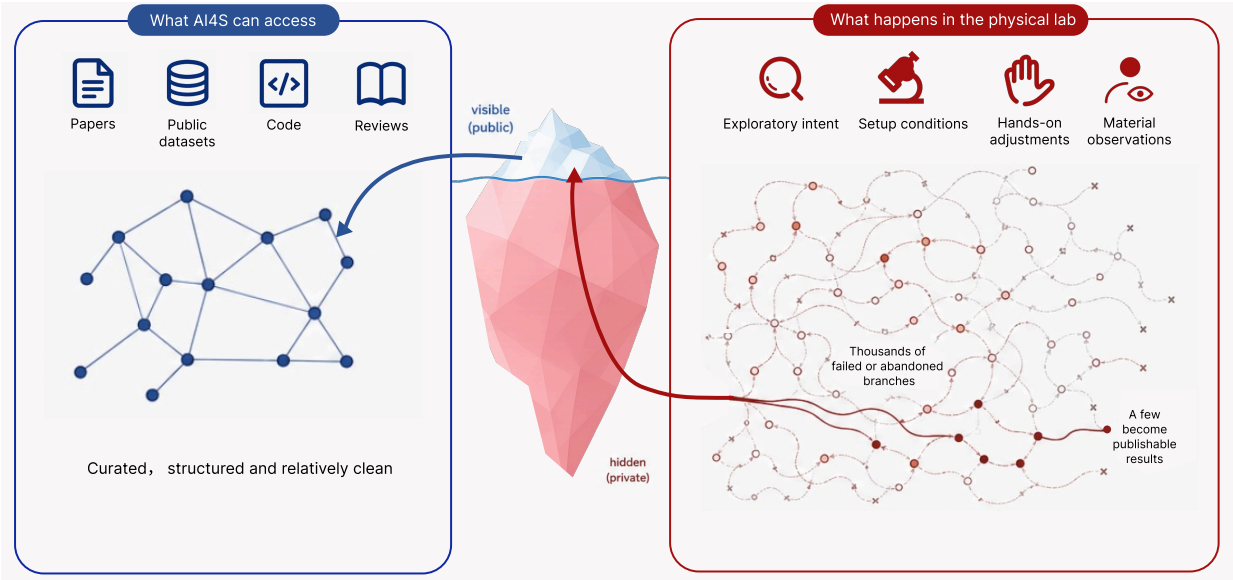

Figure 1. What AI for Science can access versus what happens in an exploratory physical laboratory. AI4S systems train on curated artifacts (papers, public datasets, code, and reviews) that represent only the fraction of laboratory work that survives to publication. The exploratory intent, hands-on adjustments, and material observations behind those artifacts mostly remain unrecorded, along with the many attempts that were tried and abandoned along the way.

Abstract

In exploratory research in physical laboratories, much of what guides discovery forms through direct work with materials and is only partly available to formal records. AI systems built on published literature and structured data have no way of knowing what these records do not carry. AI systems fill these gaps without recognizing them and produce output that only the researcher can evaluate. As researchers in a biomaterials and bioengineering laboratory, we demonstrate through a sustained collaboration with LLM-based AI systems how these difficulties compound in practice: judgment from handling materials did not transfer through the files AI received, and the reasoning that trained researchers apply by default was absent from AI's output. Correcting these gaps surfaced this judgment and reasoning, which had been invisible to both sides. We argue that bringing AI systems into exploratory research in physical laboratories beyond these fragments requires co-adaptation: as the researcher becomes better at recording what would otherwise stay invisible to AI, and AI becomes better at recognizing what records do not carry, both adapt through sustained interaction.

[1]Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, USA [2]Department of Computer Science, Cornell University, Ithaca, NY, USA [3]Design Technology, Cornell University, Ithaca, NY, USA. Correspondence to: Zeyu Li <zl788@cornell.edu>.

Accepted at the AI for Science workshop (ICML 2026).

1. Introduction

AI for Science has advanced rapidly in computational domains and in physical laboratories with well-defined optimization targets. But much of laboratory science is ex-

ploratory: researchers vary conditions without a settled theory (Steinle, 1997), and the judgment they develop through direct work with materials is only partly available to formal records (Polanyi, 1966). Researchers may document what they did and observed, but not all of the practical judgment that made those observations meaningful can be fully stated in a record. Current AI for Science systems, built primarily on published literature and structured data, have not yet extended to reach exploratory research in physical laboratories, where such judgment remains only partly available in formal records. AI built on these records can still produce output, but verifying it requires the same expertise as the research itself, so the researcher's effort goes to checking AI rather than advancing the work. Reaching this kind of research means AI's output is grounded enough in such judgment that the researcher can build on it instead.

We argue that AI for Science cannot reach exploratory research in physical laboratories without co-adaptation between researchers and AI systems through sustained interaction. Neither the researcher nor the AI system can close this gap independently; both must change while working together. We write from a biomaterials and bioengineering laboratory where this kind of research is part of our daily work. Drawing on a sustained collaboration with LLM-based AI systems, this paper describes where this gap appears, why it persists, and how these difficulties arose in practice and what the interaction began to surface.

2. Problem Setting

Much of science advances not through executing a fixed plan, but through ongoing exploration in which researchers vary conditions to discover how materials or processes behave, often without a settled theory. Philosophers of science call this exploratory experimentation (Steinle, 1997). Experimentation has been involved in 75% of all major scientific discoveries cataloged since 1900 (Krauss, 2024), and across physical laboratory sciences, much of that experimentation takes this exploratory form (Knorr Cetina, 1999), for example in biomaterials, tissue engineering, soft-matter chemistry, and many other fields. Yet in practice, this research rarely proceeds directly from design to data to publication. A researcher may try a condition, find that it fails, and return to exploring a different approach before any data are collected. Many experimental branches stop before they are ever measured because the researcher can tell from handling the material that a condition is not worth pursuing further. Only a fraction of attempts survive long enough to produce formal results.

AI for Science has made progress in physical laboratories when research goals are well defined. Self-driving laboratories can autonomously synthesize predefined target materials by proposing, executing, and refining experiments in a closed loop (Szymanski et al., 2023; Tom et al., 2024; Volk et al., 2023). AI has also begun to support exploratory research, but so far primarily in computational domains where the entire process exists in digital form (Lu et al., 2024).

When research is both exploratory and physical, the situation is different. AI has not yet reached this kind of work, where what guides the research forms through direct engagement with materials and only partly enters formal records. Researchers who do this kind of bench work report the same gap: AI remains contextually disconnected from their day-to-day work in the laboratory (Hou et al., 2026).

3. Why the Gap Persists

We believe the reason lies in how knowledge is obtained and transferred between the researcher and the AI when working on exploratory research in physical laboratories.

The researcher gains knowledge through direct sensory interaction with materials (Polanyi, 1966; Pickering, 1995). Much of this knowledge is bodily and sensory, resisting articulation in words. The observations that could be articulated face a different problem. In exploratory research, the significance of a given attempt is often not apparent at the time it occurs. A failed condition or an abandoned branch may turn out to matter only in light of what the researcher discovers next (Rheinberger, 1997). By the time the research direction settles enough to know what was important, many attempts have already passed without record. It is through these unrecorded attempts that the researcher's judgment forms: what is worth pursuing, what a result can support, and when something is going wrong. This judgment largely disappears from view before formal results are produced (Collins, 1985).

AI systems built on language models are now widely used by researchers (Liao et al., 2024; Liang et al., 2024). These systems acquire knowledge from formal records such as published text, structured databases, and documented protocols that have already been filtered and compressed. When a language model works with this text, it searches for similarities between what it encounters and patterns it has seen during training (Mirzadeh et al., 2025). This can be effective when the material fits familiar patterns. But when the material is novel or when the text contains gaps, the model does not recognize the boundary of what it can support. It fills the gap with a statistically likely completion, producing output that resembles structured reasoning but is governed by learned regularities rather than by access to what the text describes (Bender et al., 2021; Floridi et al., 2025). Studies show that language models consistently overestimate the probability that their outputs are correct (Kadavath et al., 2022; Tian et al., 2023; Sun et al., 2025).

In exploratory research, these two difficulties compound.

Much of what the researcher knows never enters any formal record and therefore never reaches AI. This is a fundamental property of this kind of research. Even between humans, this knowledge has never transferred through records alone. It has required sustained work at the bench, learning by doing experiments alongside someone more experienced (Collins, 1985). Between the researcher and AI, there is no shared practice. The researcher can offer records, but these carry conclusions without the experiential basis that formed them. AI receives these records and has no way of knowing that they are incomplete. It was never trained on the kind of knowledge that is missing (Collins, 2018; Lu, 2025; Li et al., 2026a). It processes the researcher's partial account as if it were the whole picture. The output it produces follows the conventions of scientific text (Kobak et al., 2025), and its hedging looks the same whether AI has sufficient basis or not (Yona et al., 2024). Only the researcher can judge whether AI's output is trustworthy because doing so requires the embodied judgment that AI never received. But verifying AI's work demands the same expertise as doing the research itself, so the researcher's effort goes to checking AI rather than advancing the work.

4. What Happens in Practice

The Collaboration. The observations that follow come from exploratory research in a physical laboratory in biomaterials and bioengineering, where the primary object of work is materials being designed, fabricated, and tested. In any given week, the researcher may be making samples, examining the properties they exhibit, and deciding whether those properties can support an intended application. Instrument readings rarely function as self-sufficient data to be analyzed. They confirm or challenge a judgment that is already forming through direct work with the material.

To prepare research writing from this prior work, the researcher worked with LLM-based AI systems over more than 130 conversation sessions. Whatever the researcher could access in the project record, AI could access. All working knowledge was maintained in a structured file system organized in multiple layers: source materials were processed into Markdown and JSON, organized through a navigation layer, and linked to writing process artifacts, including handoff documents that carried forward progress and decisions between sessions, and conversation records where the researcher supplied experimental context, corrected misinterpretations, and explained constraints that the files alone did not convey.

The collaboration remained difficult throughout, but the interaction itself gradually surfaced knowledge that neither side could have reached alone. The two episodes that follow show how these difficulties appeared in practice.

Case 1: Reading a Temperature. While preparing a section on how the gel behaved at different temperatures, AI worked with data from earlier experiments. The data file carried a stiffness measurement alongside a temperature of 90°C. To produce that measurement, a thin layer of gel had been placed on a heated plate that controlled the temperature while the instrument recorded how the material responded. AI read the 90°C label from the instrument-generated file and treated it as the sample's temperature. Nothing in the file suggested otherwise.

The researcher questioned it. The gel was mostly water, and if the sample reached 90°C, water would boil out, ruining the experiment. Following the instrument protocol, the researcher had sealed the sample under an oil layer to prevent evaporation. This kept the water in, but the oil layer also sat between the heating plate and the sample, blocking direct heat transfer. The researcher knew this from comparing two preparations: gels heated in a simple oven, where no oil layer sat between the heat and the sample, felt noticeably stiffer than gels heated on the instrument under the same intended conditions. The 90°C label in the file was the plate setting, not the sample temperature. The researcher carried this distinction by feel, along with dozens of similar awarenesses about what each label in each file actually referred to. AI arrived with the files but without these awarenesses.

When AI asked about this concern, the researcher explained the comparison, and AI recorded it as an acknowledged limitation. From that point, AI treated the temperature as unverified, and the writing reflected this. But the caveat by itself was thin. A colleague who had worked with this material at the same bench would have known from shared experience how much weight to give "it felt stiffer." AI received the sentence without that shared experience and could only treat it as a stated caveat. Throughout the task, similar losses appeared in protocol timing, washing procedures, instrument settings, and many other details where the record preserved what was done but not what it meant. Each loss was small on its own. Together, they shaped how the researcher read every result in the dataset, the same way most researchers with similar bench training would read it, because the same kind of experience tends to produce the same kind of judgment.

Case 2: A Sentence That Looked Like Reasoning. A different kind of gap appeared while preparing a discussion section that needed to explain a result. Two methods of making the same material had been compared (Li et al., 2026b), and one retained far less of a key component than the other. AI wrote a sentence explaining why:

"The substantial DNA loss likely occurred during the intermediate nanoparticle formation step, where incomplete NP solidification and DNA leaching into the ethanol collection bath would reduce the mass of DNA available for secondary

encapsulation."

However, no experiment had measured where the loss actually occurred. Words like "likely" and "would" carry meaning in scientific writing: they signal that the writer has weighed evidence and is offering an inference, not stating a fact. AI used the same language without having done that weighing, and the sentence looked like careful reasoning because the language of careful reasoning is what AI had learned to produce.

The researcher caught it immediately. AI did not reason about evidence the way a trained researcher does. For anyone with research training, sorting claims by what stands behind them is part of how thinking works: knowing whether a claim rests on a measurement, a published finding, an inference, or nothing at all. But AI produced the sentence without that discipline because generating text in the form of scientific reasoning does not require performing the reasoning itself. The researcher had never needed to state this sorting before because, among trained colleagues, it operates by default. Explaining this reasoning to AI pushed the researcher to put it into an explicit rule AI could apply to its outputs. Once stated that way, AI applied the sorting retroactively across the chapter and caught several prior sentences that made claims without evidence. The sorting became a working rule for subsequent sessions, though AI's compliance remained inconsistent, because rules in the context do not reliably constrain how language models generate text. Throughout the collaboration, similar patterns accumulated across evidence handling, procedural description, and data interpretation. Each pattern shaped how the research moved from data to claim, the same way most trained researchers would handle it, because the same kind of training produces the same kind of evidential discipline.

What Both Sides Would Need to Change. The first case surfaced small judgments that were never written down, which shaped how every measurement was read; the second surfaced reasoning disciplines that trained researchers apply without stating, shaping how results become claims. In exploratory research in physical laboratories, both are substantial. They are how the research moves, and the record does not carry them. However, this evidence comes from a single researcher at the end of the research cycle. The writing task was close to what AI already handles, and even there, the interaction could only recover small parts of what the researcher knew. Earlier in the cycle, where the research is more exploratory and less of it reaches any record, the gap is likely wider. If sustained interaction with AI had been present from those earlier stages, the knowledge available to both sides would have been richer.

If more researchers in these fields worked with AI throughout their research cycles, the knowledge that currently disappears between the bench and publication would begin to accumulate.

But this collaboration showed that reaching that point requires change on both sides. On the researcher's side, interaction with AI throughout the research cycle could change how knowledge is recorded, preserving what currently disappears before reaching any record. Researchers would also need to learn how to work with AI through sustained conversation, a skill that remained difficult even after 130 sessions in this collaboration, and that most researchers outside the AI field have had little occasion to build (Zamfirescu-Pereira et al., 2023; Tankelevitch et al., 2024). On the AI side, the researcher's effort was repeatedly directed toward showing AI what the research context required, rather than advancing the research itself.

If AI systems could seek information from the researcher when encountering gaps rather than filling them silently, that burden would shift. Further, exploratory research in physical laboratories relies on many instruments and examinations whose data currently reaches AI only after passing through the researcher's handling. Connecting them directly would reduce that dependency across the full scope of laboratory work.

5. Conclusion

Exploratory research in physical laboratories produces knowledge that is substantial but largely invisible to AI. This paper describes why, from within this kind of research, and shows through a sustained collaboration what the interaction between a researcher and AI could and could not surface. The gap follows from how knowledge works in this kind of research, and the cases suggest that narrowing it requires both sides to change through working together. As the researcher becomes better at working with AI and AI becomes better at reaching what the researcher knows, both adapt. That co-adaptation, through sustained interaction, is where this work points.

Impact Statement

This paper presents work whose goal is to advance responsible AI for Science. We highlight that AI systems for physical laboratory research should be developed and used with attention to the experimental context and judgment that formal records often do not preserve. We do not identify additional ethical concerns beyond those generally associated with deploying AI systems in scientific research.

References

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021*

ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), pp. 610–623. ACM, 2021. doi: 10.1145/3442188.3445922.

Collins, H. *Artifictional Intelligence: Against Humanity's Surrender to Computers*. Polity Press, Cambridge, 2018.

Collins, H. M. *Changing Order: Replication and Induction in Scientific Practice*. Sage, London, 1985.

Floridi, L., Morley, J., Novelli, C., and Watson, D. What kind of reasoning (if any) is an LLM actually doing? on the stochastic nature and abductive appearance of large language models. *arXiv preprint arXiv:2512.10080*, 2025. doi: 10.2139/ssrn.5901962. URL https://arxiv.org/abs/2512.10080.

Hou, I., Qin, A., Cheng, L., and Guo, P. J. Beyond the desk: Barriers and future opportunities for AI to assist scientists in embodied physical tasks. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, 2026. URL https://arxiv.org/abs/2603.19504. arXiv:2603.19504.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Knorr Cetina, K. *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press, Cambridge, MA, 1999. ISBN 978-0-674-25894-5. doi: 10.4159/9780674039681.

Kobak, D., González-Márquez, R., Horvát, E.-Á., and Lause, J. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813, 2025. doi: 10.1126/sciadv.adt3813.

Krauss, A. Redefining the scientific method: as the use of sophisticated scientific methods that extend our mind. *PNAS Nexus*, 3(4):pgae112, 2024. doi: 10.1093/pnasnexus/pgae112.

Li, L., Wang, R., Song, H., Mao, Y., Zhang, T., Wang, Y., Fan, J., Zhang, Y., Ye, J., Zhang, C., and Gong, Y. What papers don't tell you: Recovering tacit knowledge for automated paper reproduction. *arXiv preprint arXiv:2603.01801*, 2026a. URL https://arxiv.org/abs/2603.01801. Li and Wang are co-first authors.

Li, Z., Ramón, C. L., Bogdanowicz, S., Koeberle, A. L., Wang, D., Rueda, F. J., Cowen, E. A., Walter, M. T., Sethi, S. A., Lodge, D. M., Luo, D., and Andrés, J. A. Tracing environmental DNA transport in a large lake with synthetic DNA microparticles and hydrodynamic modeling. *Environmental Science & Technology*, 60(4): 3519–3531, 2026b. doi: 10.1021/acs.est.5c11071.

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., and Zou, J. Y. Mapping the increasing use of LLMs in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024. URL https://arxiv.org/abs/2404.01268.

Liao, Z., Antoniak, M., Cheong, I., Cheng, E. Y.-Y., Lee, A.-H., Lo, K., Chang, J. C., and Zhang, A. X. LLMs as research tools: A large scale survey of researchers' usage and perceptions. *arXiv preprint arXiv:2411.05025*, 2024. URL https://arxiv.org/abs/2411.05025.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. URL https://arxiv.org/abs/2408.06292.

Lu, J. Tacit knowledge in large language models. *The Review of Austrian Economics*, 2025. doi: 10.1007/s11138-025-00710-5. Published online 10 November 2025.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.05229. arXiv:2410.05229.

Pickering, A. *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press, Chicago, 1995. doi: 10.7208/9780226668253.

Polanyi, M. *The Tacit Dimension*. Doubleday, Garden City, NY, 1966.

Rheinberger, H.-J. *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press, Stanford, CA, 1997.

Steinle, F. Entering new fields: Exploratory uses of experimentation. *Philosophy of Science*, 64:S65–S74, 1997. doi: 10.1086/392587. Proceedings, Supplement, Part II: Symposia Papers.

Sun, F., Li, N., Wang, K., and Goette, L. Large language models are overconfident and amplify human bias. *arXiv preprint arXiv:2505.02151*, 2025. URL https://arxiv.org/abs/2505.02151.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y., and Ceder, G. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624 (7990):86–91, 2023. doi: 10.1038/s41586-023-06734-w.

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., and Rintel, S. The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 2024. doi: 10.1145/3613904.3642902. Best Paper, CHI 2024.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.330.

Tom, G., Schmid, S. P., Baird, S. G., Cao, Y., Darvish, K., Hao, H., Lo, S., Pablo-García, S., Rajaonson, E. M., Skreta, M., Yoshikawa, N., Corapi, S., Akkoc, G. D., Strieth-Kalthoff, F., Seifrid, M., and Aspuru-Guzik, A. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024. doi: 10.1021/acs.chemrev.4c00055.

Volk, A. A., Epps, R. W., Yonemoto, D. T., Masters, B. S., Castellano, F. N., Reyes, K. G., and Abolhasani, M. AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nature Communications*, 14 (1):1403, 2023. doi: 10.1038/s41467-023-37139-y.

Yona, G., Aharoni, R., and Geva, M. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7752–7764, 2024. URL https://aclanthology.org/2024.emnlp-main.443/. arXiv:2405.16908.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., and Yang, Q. Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg, Germany, 2023. ACM. doi: 10.1145/3544548.3581388.