It's not the Language Model, it's the Tool: Deterministic Mediation for Scientific Workflows

Marios Adamidis
Department of Materials Science and Technology, University of Crete
Heraklion, Greece
Institute of Electronic Structure and Laser, FORTH
Heraklion, Greece
m.adamidis@iesl.forth.gr

Danae Katrisioti
Department of Materials Science and Technology, University of Crete
Heraklion, Greece
Institute of Electronic Structure and Laser, FORTH
Heraklion, Greece
danae_kat@materials.uoc.gr

Yannis Tzitzikas
Computer Science Department, University of Crete
Heraklion, Greece
Institute of Computer Science, FORTH
Heraklion, Greece
tzitzik@ics.forth.gr

Emmanuel Stratakis
Institute of Electronic Structure and Laser, FORTH
Heraklion, Greece
Department of Physics, University of Crete
Heraklion, Greece
stratak@iesl.forth.gr

Abstract

Language models can produce convincing scientific analyses, but repeated generations on the same data do not guarantee the same result. A researcher may regenerate an identical query and receive a different fit, a different peak position or a different analysis procedure, without an obvious way to decide which output to trust. We propose *typed mediation*, a pattern in which the model orchestrates deterministic tools rather than generating analytical code. Each tool encodes one researcher's exact procedure for one instrument, ported through structured interviews. The model selects which tool to call and with what parameters. The tool produces the result. Regeneration does not change it. We evaluate this claim by running the same photoluminescence analysis on four platforms, including three commercial foundation models, four times each with the same prompt. The typed tool produces identical results across all runs. The commercial platforms either vary in numerical output and analytical methodology across runs, or fail to produce valid results on the task. We deploy this pattern on two instruments serving users over approximately six months, with very positive user feedback. Both cases are very challenging: they involve proprietary binary formats and per-seat licensed software, which force the tool to remain on local infrastructure alongside the data and the instrument it operates. We argue that deployment topology is not just a preference, but a structural requirement of scientific tool mediation. The result is a practical pattern for deploying language models in scientific workflows where reproducibility is mandatory, reducing analysis time from weeks to minutes while guaranteeing identical outputs across runs.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SETN 2026, Chania, Crete, Greece
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Computer systems organization** → *Architectures.*

Keywords

scientific AI, reproducibility, tool-mediated agents, deployment topology, Model Context Protocol

ACM Reference Format:
Marios Adamidis, Danae Katrisioti, Yannis Tzitzikas, and Emmanuel Stratakis. 2026. It's not the Language Model, it's the Tool: Deterministic Mediation for Scientific Workflows. In *Proceedings of 14th EETN Conference on Artificial Intelligence (SETN 2026).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Motivation: Reproducibility First. The integration of language models into experimental data analysis is accelerating, and with it the assumption that the model's output is reliable enough to act on. For instance, consider the case of a researcher, who after a long day of measurements, returns to the laptop to analyze a messy photoluminescence spectrum file[1]. The researcher uploads the file to the AI assistant they have used many times before and found efficient at producing what they consider passable. The model proceeds to work on the file via code execution and rapidly arrives at a result that appears well crafted and articulated. From the peak fit to the plot and the analysis, the output validates the researcher's hard work. There is only one simple issue: the researcher forgot to paste the right labels for the measurements, so they edit the query and press the regenerate button. This time, the peak position has shifted by 2 eV, the fit is just a bit different. Neither answer is obviously wrong. What is the principled way for the researcher to act at 9pm on a Thursday?

One important observation is that this failure *does not go away with a more powerful model.* The model's view of the project's

[1] A photoluminescence spectrum records the light emitted by a material after optical excitation, as a function of wavelength. It is a standard characterization technique in semiconductor physics.

semantics is governed by a generic system prompt, not tailored to any specific researcher's workflow. The model is capable enough to consistently produce convincing results. What it cannot do is produce the same result twice. The gap is determinism, and more model capability does not provide it [9]. Reproducibility is not a luxury in science; it is the standard that separates real discovery from mere claim.

This widespread issue appears as well in the low adoption of agents in academic research and medicine, sitting at low single digits of real-life deployment as shown in Anthropic's recent report [12]. These fields have been investigated greatly for the potential of AI-related integrations, but a lack of technical know-how, combined with the marketing pressure that the model is already capable enough on its own, has made it impractical for unsupported researchers to graduate from zero-shot prompting to task-strengthened, specialized harnesses [11].

Motivation: The Need for Privacy. A privacy concern also arises from the interaction. Many types of laboratory data appear uniformly consistent to the human eye, dampening the researcher's instinct towards what they will share with the cloud provider. A file containing a bad measurement is virtually indistinguishable from one that will lead to an impactful publication. Sharing raw data and results with the cloud provider creates a non-zero probability that the data shapes the next checkpoint's performance in that specific niche [2]. Cutting-edge scientific output has higher probability of being valuable on later rounds of training when compared to trivial information extracted from the web.

Approach. To address both requirements, we encode the researcher's exact manual procedure behind a tool surface and let the model orchestrate it. The tool then returns deterministic results that reproduce across regenerations, verification becomes quick, and the model's reasoning capacity is spent on the parts of the work that need it. Figure 1 contrasts the three arrangements: (a) the traditional method, in which the researcher manually operates particular software; (b) the current trend, in which an LLM performs the analysis and, operating stochastically, derives non-reproducible results; and (c) our arrangement, which leverages *both* the existing software and the LLM, "wrapping" the former in a way that is convenient for the latter to orchestrate deterministically. The pattern is not itself new: schema-gated tool use is established practice, and our contribution is operational rather than architectural. What we report is what it takes to make the pattern hold for one researcher's instrument workflow, and what happens when it is left running in a laboratory for six months. This framing also puts a practical question in front of us: when the workflow lives in the tool rather than the model, how much does model choice matter? We take it up in §5.1. A more refined illustration is given in Figure 2.

Evaluation. We have realized this approach as a platform, called FORTHought, we have deployed it in a research center, and the platform is in use by eleven researchers, across multiple instrument workflows, for a period of 6 months.

The users have provided very positive feedback, and FORTHought has now become part of their daily routine. In this paper, we shall focus on (a) the two deployments with the largest documented impact: a photoluminescence analysis pipeline and a scanning electron microscopy workflow, and (b) on the evaluation of *reproducibility* by running the same photoluminescence analysis on our platform and three commercial foundation models, four times each with the same prompt and the same data.

Contributions. In summary, our work makes four main contributions:

(1) we describe a *typed-mediation pattern* in which both the human and the language model address the laboratory software via the same typed interface, placing the deterministic core of the workflow in the tool rather than the model,
(2) we present *two real applications* ported through structured interview sessions with the researchers who own the workflows, deployed and actively used by researchers who verify outputs quickly and feed corrections back into the tools,
(3) we evaluate reproducibility by running the same analysis on four platforms four times each and show that the typed tool produces identical results across runs while code-generating approaches vary in both output and analytical methodology, and
(4) we argue that deployment topology is a structural requirement of scientific tool mediation, driven by both privacy concerns and the licensing constraints of most laboratory software, which together force the tool to live alongside the data and the instrument it operates.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our approach, i.e. the typed mediation pattern. Section 4 presents two deployment cases. Section 5 evaluates reproducibility across platforms. Section 6 concludes and outlines future directions.

2 Related Work and Novelty

Deployed laboratory agents remain rare. Hellert et al. report a synchrotron deployment where an agentic framework manages real-time operations across more than 230,000 control channels [10], while Vriza et al. find that, at a national X-ray nanoprobe facility, performance on the same task varies sharply across models [15], and Xie et al. show both the potential and the fragility of LLM-driven instrument control [17]. On the reproducibility side, Cissé et al. observe a single reasoning model producing outlier results below half its own average over repeat runs [3], and Cui and Alexander find considerable variation across 480 attempts on one analysis task even under identical configurations [4]. Within-model variance is thus a structural property of code-generating approaches, not a failure of any particular model.

Tool-mediated architectures address this by restricting the model to validated tool calls. Yang et al. report 100% tool-selection consistency across three models at maximum temperature [19] and Xu et al. 100% specification-level reproducibility when the model is limited to routing [18]; Pan et al. bring MCP-based mediation to national-scale cyberinfrastructure [13]. Strickland et al. formalize the principle as schema-gated orchestration, where nothing executes unless it validates against a machine-checkable specification [14], Doshi et al. extend MCP with labels for capability, confidentiality and trust [7], and Deng et al. present a skill-centric framework spanning ten categories of precision instruments [5].

Our contribution is operational rather than architectural. We take the schema-gated principle and apply it to a specific problem

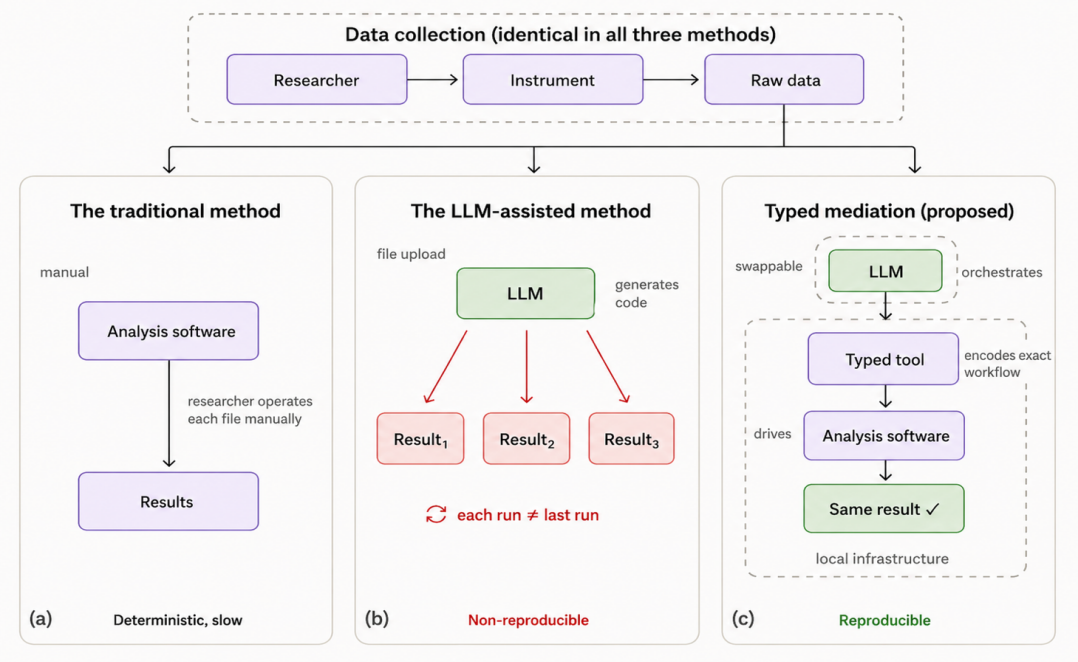

Figure 1: Three approaches to scientific data analysis. Data collection is identical in all cases. (a) Traditional: the researcher manually operates analysis software. (b) LLM-assisted: the model generates analytical code, producing different results on each run. (c) Typed mediation (proposed): the model orchestrates a deterministic typed tool that encodes the researcher's exact workflow on local infrastructure.

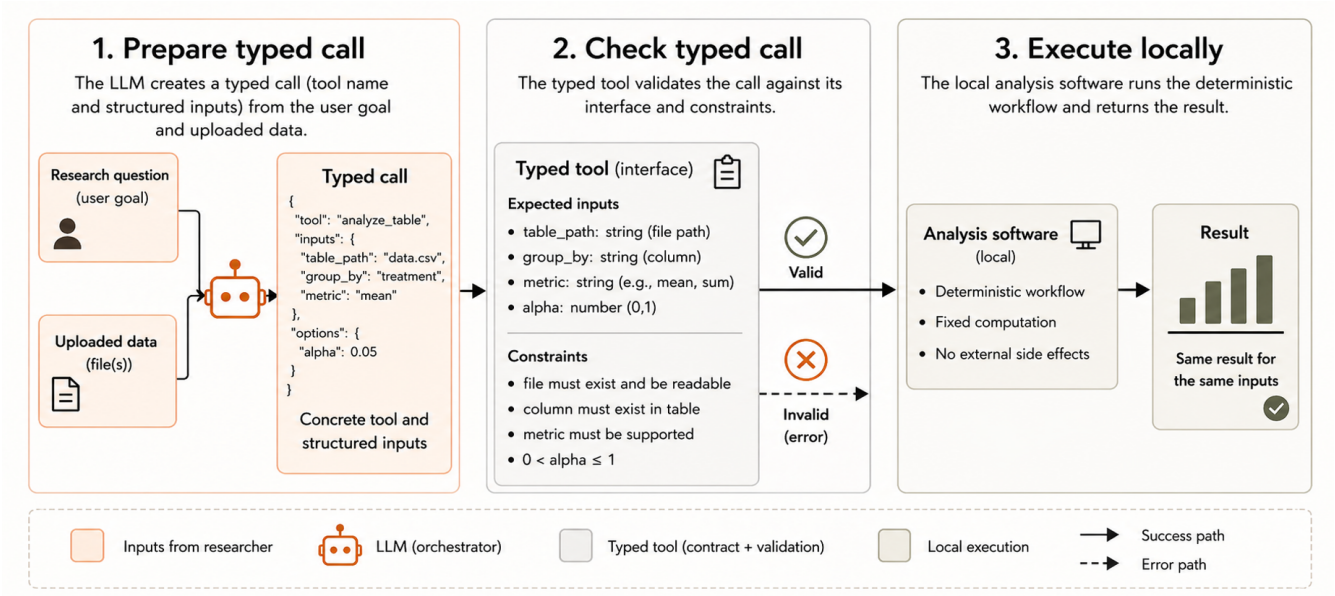

Figure 2: Typed mediation architecture. The model orchestrates tool calls through a typed schema. The tool encodes the scientist's workflow and drives licensed software on local infrastructure.

that none of these systems address: encoding one researcher's exact manual workflow, extracted through structured interviews, as a typed tool that any model can orchestrate.

Table 1: Key terminology of FORTHought

Concept	Definition
Software	Any application a human operates directly: a spreadsheet program, a licensed instrument suite, a standalone script. Designed for human interaction.
Tool	A function exposed to the language model through a programmatic interface. The model calls it; the user does not operate it directly.
Schema	A machine-readable specification of a tool's expected inputs and outputs: parameter names, types, allowed values and return format. It acts as the contract between the model and the tool.
Typed tool	A tool whose interface is defined by a strict schema. A call that violates the schema is rejected before execution. A call that conforms executes the same deterministic procedure every time, regardless of which model issued it.
MCP	The Model Context Protocol [1]: an open standard that exposes typed tools to language models. It defines how a model discovers available tools, reads their schemas, and issues calls.
Skill	A configuration document that defines the model's behavior for a specific instrument or researcher workflow. It specifies which tools are available, what the researcher's preferences are, and how results should be presented. Each deployment case has its own skill. An example is given in Appendix A.
Workflow specification	The output of the structured interview process: the complete set of analytical decisions a researcher applies to their data fitting model, spectral window, quality thresholds, preprocessing steps. Each typed tool encodes exactly one workflow specification. An example is given in Appendix B.

3 The FORTHought Approach

At first we should clarify terminology. The key concepts and their terms and descriptions, are given in Table 1. The reader is suggested to read it, to avoid ambiguities.

We could summarize the entire approach as follows:

> Each *typed tool* encodes a *workflow specification* extracted through *structured interviews* with the researcher, *wraps* the operations of *existing licensed software*, and is *exposed to the model* via MCP. A per-instrument *skill* tells the model which tools are available and how the researcher expects results to be handled. The model calls the tool, and the tool drives the software.

A Toy Example. Before turning to instrument data, it is worth seeing the failure in a setting with no physics in it at all. A user uploads a column of numbers and asks: *"give me the average, excluding outliers."* The request is unambiguous in intent and underspecified in method.

> **Data.** {4.9, 5.0, 5.0, 5.1, 5.1, 5.2, 12.0}. **Prompt.** "Average, excluding outliers."
> **Code-generating run 1.** The model writes code that discards points more than three standard deviations from the mean. Nothing is discarded: 12.0 sits only 2.3σ out, because it inflates the very σ it is tested against. Result: **6.04**.
> **Code-generating run 2.** The model writes code that discards points outside 1.5× the interquartile range. Here 12.0 is discarded. Result: **5.05**.
> **Typed mediation, every run.** The tool exposes `robust_mean(data, method, k)`, where the schema restricts `method` to `{iqr, sigma}` and the workflow specification fixes the researcher's own choice (`iqr`, $k = 1.5$). The model's only remaining job is to locate the data and issue a schema-valid call. Result: **5.05**, on every run.

Neither answer is a hallucination and neither is wrong: they answer different questions, differ by 20%, and nothing in the prompt distinguishes them. The model is not failing to recall a fact, it is *making a methodological choice*, and it remakes it on every regeneration. Typed mediation removes that choice from the model and places it in the tool, where the researcher made it once and can inspect it. The rest of this paper is this example at instrument scale: in the photoluminescence deployment (§4.1) `method` becomes the peak-fitting model (Voigt rather than Lorentzian) and k the despiking threshold and spectral window, with thirty such decisions rather than two. Section 5 reports what that costs.

3.1 Typed Mediation

Typed Mediation. We use the term *typed mediation* to refer to a pattern in which the language model does not perform the analysis itself, but selects and invokes a deterministic tool that encodes the researcher's exact procedure through a typed schema.

Determinism. Typed mediation places a deterministic tool between the model and the software the researcher operated manually. A system prompt and a per-experiment `Skill.MD` file (Appendix A) shape the model's behavior; the model then decides which tool to call and with what parameters, and the tool returns results for the model to serve or reason over. The interface follows the Model Context Protocol [1], which exposes each tool through a typed schema specifying inputs, outputs and execution requirements. The tool is not a generic API wrapper: it encodes one researcher's exact procedure for one instrument, for example the despiking parameters, spectral window and peak boundaries they use. The model picks which tool to call; the tool already knows how the work is done.

Architecture Overview. Figure 2 summarizes the typed mediation architecture. The model operates above the typed interface and is replaceable. The tool sits below it, anchored to the same machine as the licensed software and the instrument.

Why we need typed mediation. Language models are stochastic by design. That serves them well where recall would fail, but in scientific analysis the methodology directly affects the result and its relationship to data from other sessions. A deterministic tool surface lets the model spend its reasoning on scientific insight, grounded in the tool's output, rather than on what code to execute. Recent work

confirms this empirically: 100% tool-selection consistency across three models at maximum sampling temperature [19], and 100% specification-level reproducibility when the model is restricted to routing decisions [18].

The other path is to let the model generate the code required to satisfy the user's demands, call generic APIs or reason about the workflow itself. This is the current approach of most users, as reported by Anthropic [12]. On that path the choice of model becomes load-bearing, and measurably so; we return to the evidence in §5.1.

That volatility disappears when the model is restricted to selecting which tool to call. The typed schema accepts valid parameters and executes the same procedure regardless of which model issues the call. But the tool itself cannot move. Proprietary binary formats open only inside their licensed application, per-seat licenses bind that application to one workstation, and instrument-specific calibration files encode settings that apply to one physical device [6]. A typed interface exposed through MCP resolves this asymmetry: the tool remains where the work lives, and any model that can issue a valid call can orchestrate it, from bench instruments to national-scale cyberinfrastructure [13].

3.2 Problem Statement and Approach Formally

The problem of stochastic analytical variance. Let D be a scientific dataset and P be a natural language prompt describing an analytical procedure. In a code-generating LLM workflow (the "LLM-assisted method"), the analysis is treated as a stochastic function f_θ parameterized by the model weights θ. The output R is generated as: $R_i = f_\theta(D, P, \tau, s_i)$ where τ is the sampling temperature and s_i is the random seed for run i. The problem is defined by the *variance of methodology* $\mathrm{Var}(f_\theta(D, P)) \neq 0$, i.e. even when D and P remain constant, $R_i \neq R_j$ because the model reconstructs the analytical methodology (e.g., peak fitting models, background subtraction, spectral windows) stochastically at runtime. In a scientific context, this lack of determinism prevents R from being a "verifiable scientific result".

Proposed Solution. We adopt a transformation in which the LLM is restricted from generating code and is instead limited to *parameterizing* a deterministic tool. Let T be a *typed tool* that encodes a fixed analytical workflow W. The tool is governed by a schema S, which defines the valid input space $\mathcal{X}$.

In particular, the workflow W is defined by the *interview* (that extracts the researcher's methodology) and implemented (with the assistance of a developer) in the tool code (T), while the *skill file* defines S (the schema), i.e. it tells the model how to call the tool, not what the tool does internally (we could say that the skill file is the menu, not the kitchen). To summarize, the process is: *Interview* $\xrightarrow{\text{what analysis to do}} W \xrightarrow{\text{code impl.}} T$, and *Skill file* $\xrightarrow{\text{llm}}$ *interface* S of T. According to this approach, the process is decomposed into two distinct phases:

- Orchestration (Stochastic): The LLM acts as a *mediator* m_θ that maps (D, P) to a set of parameters $x \in \mathcal{X}$ such that x satisfies the constraints of S, i.e. we can write $x = m_\theta(D, P, S)$.
- Execution (Deterministic): The tool T executes the workflow W using parameters x, i.e. $R = T(D, x)$.

This **guarantees reproducibility** since T is a non-stochastic function implemented in local infrastructure, and thus it satisfies the property: $\forall i, j : x_i = x_j \Rightarrow T(D, x_i) = T(D, x_j)$. By enforcing x through a *typed schema* (Schema-Gated Orchestration), the system ensures that as long as the model's tool-selection x is consistent, the scientific result R is perfectly reproducible. This shifts the burden of "correctness" from the model's creative generation to a validated, researcher-defined tool surface.

This decouples reasoning (the LLM) from computation (the typed tool), and confines T to local infrastructure while m_θ may live in the cloud. Writing $\sigma^2_{total} = \sigma^2_{mediation} + \sigma^2_{execution}$, we have $\sigma^2_{execution} = 0$ by construction; the deeper finding is that a *tight enough* schema S also drives $\sigma^2_{mediation} \approx 0$ empirically, so the selected parameters (and hence the intensity exponent b) reproduce across regenerations.

4 Deployment Cases of FORTHought

As stated earlier, here we describe two deployment cases, that have significant impact. A few statistics about these two cases, i.e. number of typed functions, users, logged sessions, time (i.e. before and after FORTHought), validation, and other constraints, are given in Table 2.

Table 2: Deployment summary. Both tools have been in continuous use for approximately six months.

	PL spectroscopy	SEM periodicity
Typed functions	22	8
Active users	1 primary + group	3
Logged sessions	60+	15+
Manual time	~2 weeks	~2 days
Tool time	minutes	minutes
Validation	$\Delta \leq 0.02$	calibration-checked
Local constraint	binary format, per-seat license	device-specific calibration

4.1 Photoluminescence Analysis

The first deployment case is a photoluminescence analysis pipeline for semiconductor thin films. The proprietary software stores each excitation power level as a separate workbook inside a binary project file; a typical campaign produces more than twenty, each sharing a wavelength axis but holding a different emission spectrum. The researcher opens each one, applies the same sequence of operations, and combines the results into a single notebook. The initial tool automated this merging step. Figure 3 illustrates the full pipeline: the researcher describes the analysis in natural language, the model constructs a typed tool call, and the tool executes deterministically on local infrastructure.

The tool became useful, but it was not yet correct. To move from automation to fidelity, we conducted a structured interview of thirty questions across two sessions with the researcher who owns the workflow. An example is given in Appendix B. Each answer refined the tool. The fitting model was Voigt (a convolution of Gaussian and Lorentzian components that better captures peak

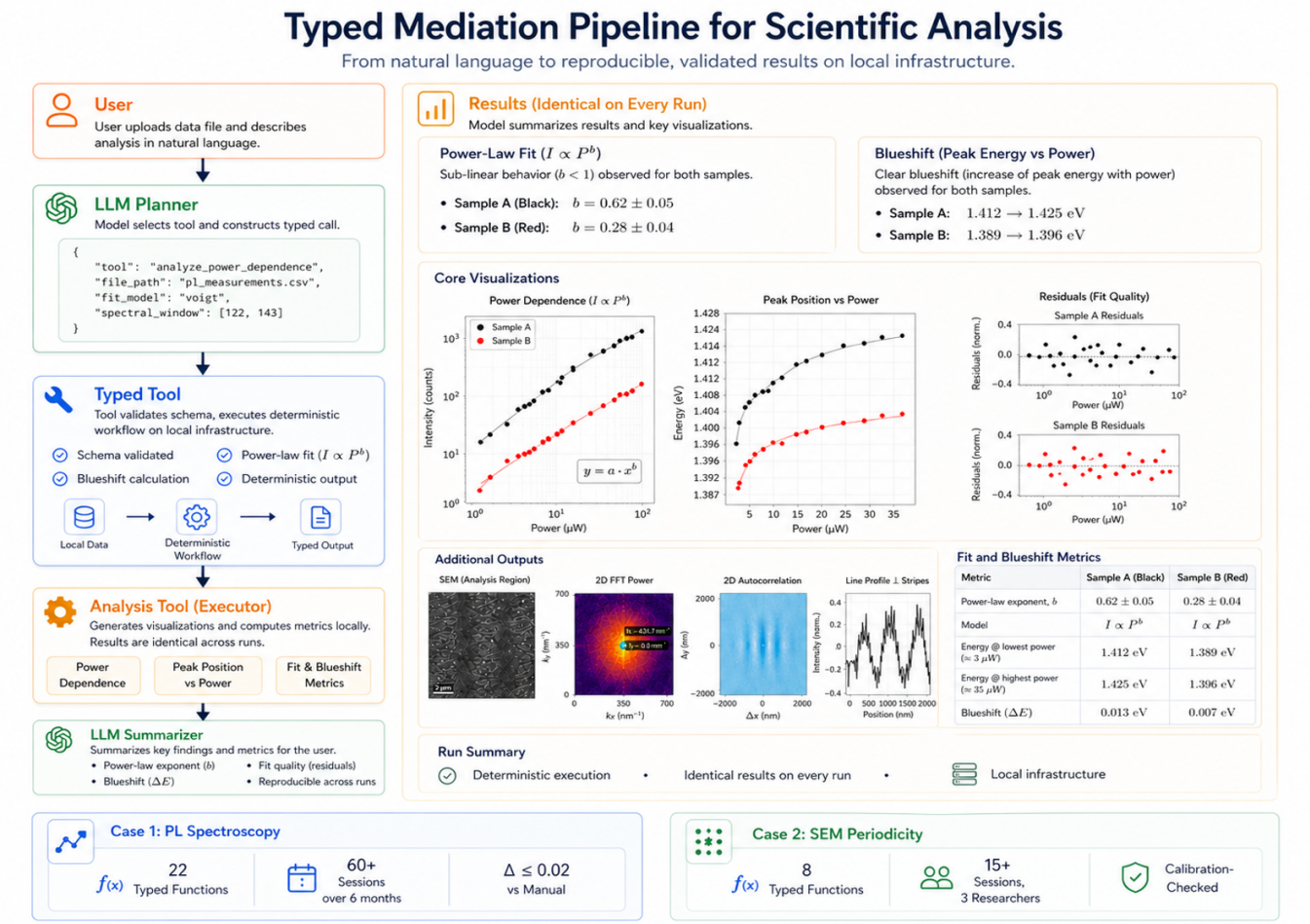

Figure 3: Typed mediation pipeline applied to photoluminescence spectroscopy. Left: the user describes the analysis in natural language; the model selects the tool and constructs a typed call; the tool validates the schema, executes the deterministic workflow on local infrastructure and returns identical results on every run. Right: core visualizations and fit metrics produced by the tool. The bottom strip summarizes both deployment cases (PL spectroscopy and SEM periodicity analysis).

shape in noisy spectra), not the pure Lorentzian we had assumed. The spectral window adapts per power level, narrowing around the emission peak as it shifts with excitation power, rather than staying fixed. The quality threshold is approximately $R^2 \geq 0.90$, evaluated visually, not the 0.95 cutoff we had implemented. Peak intensity is read as raw counts at the maximum, not as integrated area. For an example case, the power dependence uses two separate fits split at a saturation boundary near 10 μW, not a single allometric fit. None of these details were documented. They existed in the researcher's routine, refined over four years of doctoral work. The final tool encodes this routine across twenty-two typed functions. The right panel of Figure 3 shows the core visualizations and fit metrics that the tool produces from a single run.

The raw files use a proprietary binary format that opens only inside the licensed desktop application, whose license is bound to one workstation, and institutional policy forbids uploading raw experimental files to external services. This is the normal condition of many laboratories: the model cannot open the file, run the licensed software, or ask the researcher to export data she will not share. The tool has to run on the same workstation as the data and the software.

We did not build tools that imitate a result. We ported the exact workflow the scientist was already using into a typed tool that the model orchestrates.

Manual analysis of one campaign previously took about two weeks, spent on repeated file handling across twenty-one power levels, each needing the same operations inside the proprietary application. The ported workflow executes in minutes; the exponent from its automated split power-law fit matches the researcher's manual result within $\Delta \leq 0.02$, and regeneration returns the same numbers every time. She can verify the output in minutes because she knows what the correct answer should look like.

The researcher is a fourth-year doctoral student working on optical spectroscopy of two-dimensional materials. She was initially

skeptical because her existing workflow already worked. After the tool reproduced her method, she began using it regularly. Over approximately six months, she has accumulated more than sixty interaction sessions, including more than eleven sessions on the spectral analysis model. She has also introduced the platform to members of her research group. The deployment is now part of her working routine.

4.2 Scanning Electron Microscopy

The second case is a scanning electron microscopy workflow for periodic structure analysis on laser-processed surfaces. The tool encodes the calibration of one specific microscope, including pixel scale, magnification mapping and detector configuration. A different instrument of the same model would require different values.

The porting followed the same method. The researcher's analysis procedure was extracted through structured interviews and encoded as typed functions. The model selects the tool. The tool validates the input schema, executes the analysis locally and returns deterministic results. Three researchers now use this tool for periodic structure analysis on their respective samples.

This case is included to show that the pattern generalizes beyond spectroscopy. The instrument, the file format and the analysis are different. The architectural constraint is the same: the calibration binds the tool to the physical device, and the tool must run where the device is. Commercial platforms cannot replicate this workflow: the calibration files are device-specific and the analysis requires pixel-scale parameters that only the local tool encodes. The bottom of Figure 3 summarizes both deployment cases.

4.3 Application Methodology

Cost for interviews. The interview cost is small: in our experience a use case requires at most three sessions of approximately thirty minutes each.

Cost for software development. Producing the tool itself is quick (around one hour); the real cost is validation. Once the interview specifies what to build, the tool is straightforward to implement, but the researcher must then run it against real data and flag inaccuracies, which are corrected on the tool or prompt side. In the photoluminescence case (§4.1), seven of the thirty interview answers contradicted the initial implementation, each requiring a correction-and-validation cycle. Development cost therefore scales with the number of methodological corrections, not with interview duration or tool assembly. For the SEM case (§4.2), eight typed functions took about one day from interview to validated deployment.

Skills file. Writing the skill file takes minutes; it is a configuration document, not code, written last, after the tool is built and validated. It constrains the model's behavior during orchestration: which tools to call in what order, what parameter values to prefer, and how to respond when the prompt is underspecified or the tool returns an ambiguous error.

The content of the skill file matters more than its length. Modern language models are trained to be helpful, which means that when an error occurs they will attempt to use their reasoning and any available tools to find alternative solutions. In a scientific workflow this is dangerous because the alternative solution may produce a plausible but incorrect result. The skill file preempts this by defining predetermined behavior for failure modes, user preferences, and workflow-specific edge cases. For example, the PL skill file (Appendix A) specifies exact parameter names and rejects plausible-sounding alternatives that the typed schema would not accept.

A poorly written skill file is worse than having no skill file at all. Without one, the model still has the system prompt and the tool descriptions that accompany every call, and it can operate on those alone. Recent studies of over ten thousand MCP servers confirm that the quality of tool descriptions has a measurable impact on agent tool selection and argument passing [8, 16]. A bad skill file actively forces the model to pick tools and processes that do not match the user's pace or requests. This is why the skill file must be tailored to the specific user and their tasks.

The procedure is: (1) build and validate the tool until its outputs match the researcher's manual result, (2) observe where the model makes wrong choices during interactive testing with realistic prompts, and (3) write the skill to close those gaps, retesting until orchestration is stable.

5 Evaluating Reproducibility

Models Evaluated. To test whether typed mediation eliminates result variance in practice, we ran the same analysis on four platforms: (a) our system, (b) GPT-5.5 with Extended Thinking, (c) Claude Sonnet 4.6 with Adaptive Thinking, and (d) Gemini 3.1 Pro.

Evaluation Dataset and Workflow. We used the photoluminescence dataset from Section 4, exported from the proprietary analysis software as a CSV file. Each platform received the same file and the same natural-language prompt asking for fits across all power levels, extraction of peak intensity and position for each power, and allometric fits of the resulting power dependence. This is the typical workflow of an experimental scientist: export data from instrument software, upload to an AI assistant, describe the desired analysis in plain language. No prompt engineering, no parameter tuning, no specialized configuration. We ran each platform four times with no modifications between runs.

Evaluation Setup. We use four runs as a repeatability check under identical inputs, prompts and data. For the typed workflow, the analysis procedure is fixed in the tool, so repeated execution should return the same result unless the implementation or input changes. For the code-generating workflows, any change in fitting method, preprocessing step or numerical output across repeated runs shows that the procedure is being reconstructed at regeneration time. This is the risk we target: in scientific use, even one inconsistent rerun can affect which result a researcher decides to trust.

Evaluation Results. Table 3 and Figure 4 summarize the results. We can see that:

- Our platform produced identical results across all four runs: $b = 1.025$ with $R^2 = 0.995$. The output files from two of the four runs were byte-for-byte identical. A third differed by one digit in the fifth decimal place of a single coefficient due to floating-point arithmetic.
- GPT-5.5 produced valid outputs in all four runs but applied a different analytical procedure each time. Run 1 identified one emission peak. Runs 2 through 4 identified two. The background subtraction model changed between runs, with

Table 3: Reproducibility comparison across four platforms. Same dataset, same prompt, four runs each. Intensity b is the exponent from the allometric fit $I = aP^b$.

	Ours	GPT-5.5	Claude 4.6	Gemini 3.1
Valid results	4/4	4/4	4/4	0/4
Deterministic	4/4	0/4	0/4	0/4
b (range)	1.025	0.899–1.000	1.001–1.025	0.237†
σ_b	0.000	0.041	0.010	—
Intensity spread	<0.01%	54.8%	30.1%	N/A
Avg. time/run	0:34	4:21	5:15	1:50

†Fit $R^2 = 0.005$; centered on incorrect spectral region.

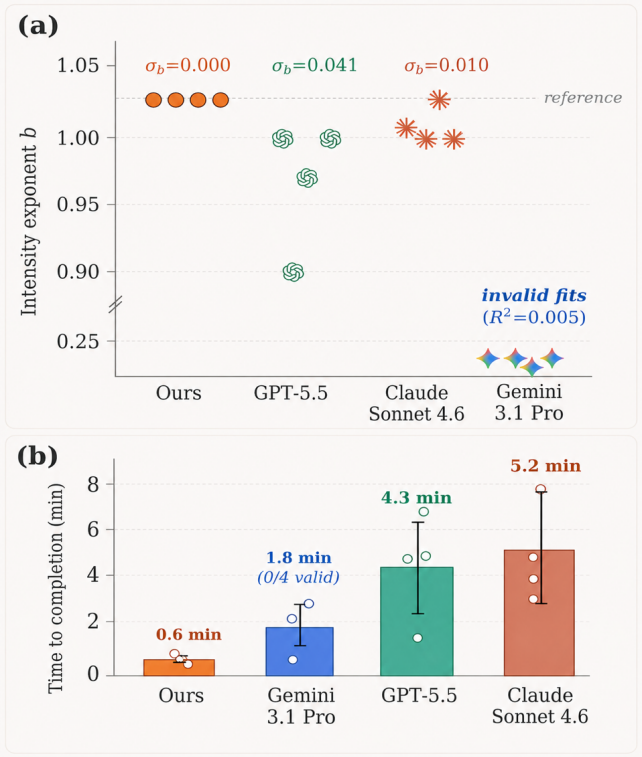

Figure 4: Reproducibility evaluation across four runs per platform. (a) Intensity exponent b from the allometric power-law fit. Our platform produces the same value across all runs ($\sigma_b = 0$); commercial platforms vary by an order of magnitude in spread. Gemini 3.1 returned values clustered near $b \approx 0.24$ from invalid fits ($R^2 = 0.005$); the y-axis is broken to show this. The dotted line marks the reference value from the researcher's own analytical procedure. (b) Time to completion, sorted by mean. Bars show mean across four runs, error bars show one standard deviation, dots show individual runs. Gemini 3.1 produced no valid results despite finishing fastest among commercial platforms.

linear, quadratic and polynomial variants appearing across the four runs. The intensity of the primary peak at the lowest excitation power varied by 54.8% across runs. These differences do not represent numerical fluctuations around a stable methodology. Each run constitutes a different experiment performed on the same data.

- Claude Sonnet 4.6 was qualitatively more consistent, always identifying one peak, though in most runs it selected the wrong emission feature rather than the primary peak. The fit window varied from 122 to 145 nm across the four runs, with no two runs using the same bounds. Intensity values at shared power levels disagreed by up to 30.1%.
- Gemini 3.1 Pro failed on all four runs. Every run missed a required preprocessing step (axis unit conversion in the exported data) and produced fits with $R^2 \approx 0.005$ centered on incorrect spectral regions.

All three commercial platforms were accessed through their respective paid subscription plans with all available features enabled, including code execution environments and extended context.

Our platform also completed each run in under one minute (mean 0:34), compared to 1:50–5:15 for the commercial platforms, reflecting the absence of code generation overhead.

The typed tool does not encode the expected answer. It encodes one researcher's analytical procedure: which fitting model to use, what spectral window to fit over, what despiking parameters to apply. If the input data changes, the result changes, but the procedure stays the same. An independent Lorentzian fit (deliberately different from the tool's Voigt profile) of the same dataset using a different spectral window and no despiking produces $b = 1.036$ ($R^2 = 0.997$), which differs from the tool's output by $\Delta b = 0.011$, within the fit uncertainty of ± 0.027. The value of b depends on methodology. What the typed tool guarantees is that the methodology does not change between runs.

5.1 Does Model Choice Matter?

We can now return to the question posed in the Introduction. The evaluation above answers it in one direction and the deployment answers it in another, and we are explicit that neither is a controlled cross-model experiment.

Under code generation, model choice is decisive and unstable. On the same dataset and the same prompt, the three commercial platforms disagreed with each other on the exponent, disagreed with *themselves* across runs, and one failed outright on all four attempts. Selecting a model in this regime does not select a procedure; it selects a distribution over procedures. This is what makes the question uncomfortable in practice: the researcher cannot inspect the distribution, and a single sample from it looks exactly like a result. The effect is not particular to our task. Vriza et al. evaluated code-generating agents at a national user facility and found that, on their hardest task, performance ranged from 0% on weaker models such as GPT-4o-mini, through 25% with high variance on Claude 3.5 Sonnet, to 100% on OpenAI's o3 [15]. Under code generation, then, model choice matters enormously, and that is precisely the problem: a dependence that strong on a component that changes without notice is not a foundation for a scientific procedure.

Under typed mediation, the question largely dissolves. The executed result is $R = T(D, x)$, an expression in which θ does not appear. Once a schema-valid call has been issued, the identity of the model that issued it cannot influence the number, because the model contributed no part of the computation. Model choice can

therefore act only through the mediation step m_θ, and a sufficiently constrained schema S leaves it little room to act: in our runs the selected parameters were identical across regenerations, which is why $\sigma^2_{mediation} \approx 0$ in §3.2. This is consistent with the tool-selection consistency Yang et al. report across three models at maximum sampling temperature [19] and the specification-level reproducibility Xu et al. report when the model is restricted to routing [18].

The deployment offers observational support, not proof. FORTHought orchestrates through a custom pipe backed by commercial Gemini models, and that pipe was repointed to different model tiers during the six-month deployment period for reasons of cost and latency unrelated to any analysis. The photoluminescence workflow continued to return results matching the researcher's manual procedure within $\Delta b \leq 0.02$ across those changes, and neither the tool nor the skill file was modified in response. We report this as operational experience rather than as a measurement: the changes were not controlled, not counterbalanced and not logged as experiments, and we decline to reconstruct them retrospectively as one.

What would settle it. A controlled study holding D, P and T fixed while varying θ across frontier and small locally-hosted models would turn this structural argument into a measurement, and would locate the point at which a model becomes too weak to populate the schema. That boundary, not the reproducibility of T, is where we expect model choice to reassert itself; we leave it to future work.

6 Concluding Remarks

We presented typed mediation, a pattern that places the deterministic core of a scientific workflow in the tool rather than the model, together with two deployed applications ported through structured interviews. On the same analysis run four times each, the typed tool reproduced identically ($\sigma_b = 0$) while three commercial foundation models varied in both output and methodology on every run. We further argued that deployment topology is not a preference but a structural requirement, since proprietary formats, per-seat licenses and data-privacy constraints force the tool to live alongside the data and the instrument it operates.

In the primary case, a photoluminescence analysis workflow that previously required approximately two weeks of manual processing per measurement campaign now executes in minutes and produces identical results across runs. In the second case, a scanning electron microscopy workflow for periodic structure analysis is now used by three researchers with calibration-checked results. The researcher who owns the PL workflow has used it in over sixty sessions across six months of continuous operation.

Limitations. Three limitations bound what this paper establishes. First, the reproducibility evaluation covers one dataset, one instrument workflow and one set of commercial platforms at a single point in time. We do not claim that code-generating approaches are inherently incapable of reproducibility, only that they did not achieve it on this task under standard usage conditions. Second, the adoption evidence is observational. Session counts, the reduction from weeks to minutes, and the researchers' reported satisfaction come from deployment logs and informal feedback rather than from a structured user study. We report no measured task-completion times, error rates or satisfaction scores against a baseline, and the time reduction is the workflow owner's estimate of her own prior practice rather than an instrumented measurement. Third, as set out in §5.1, the effect of the orchestrating model is argued structurally and supported observationally, but not measured. A structured user study and a controlled cross-model orchestration study are the two experiments this work most needs, and both are listed below.

Applicability. The relevant boundary is not between tool-assisted and tool-free tasks (modern deployments are almost never tool-free) but between tasks where the methodology should be fixed and tasks where variation is acceptable. Typed mediation offers diminishing returns for inherently exploratory work, for tasks where generic tool use is already consistently correct, and where approximate results suffice. It is most valuable where a non-trivial analytical procedure must be executed identically every time, which describes the majority of instrument-driven experimental workflows.

Issues for further work and research include: (i) extending the evaluation to additional datasets and instrument types, including a user study measuring task completion time and error rates with and without typed mediation, (ii) extending the pattern to additional instrument classes and evaluating cross-institutional deployment, and (iii) investigating how the interview-driven encoding process itself can be partially automated as the number of ported workflows grows.

Acknowledgment.

This paper has received funding from the European Union's Horizon Europe research and innovation programme GIANCE under Grant Agreement No 101119286. .

References

[1] Anthropic. 2024. *Introducing the Model Context Protocol.* https://www.anthropic.com/news/model-context-protocol
[2] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
[3] Abdoulatif Cissé, Max E. Cooper, Mengjia Zhu, Xenophon Evangelopoulos, and Andrew I. Cooper. 2026. Can We Automate Scientific Reasoning in Closed-Loop Experiments Using Large Language Models? *Digital Discovery* 5 (2026), 1132–1160. doi:10.1039/d5dd00520e
[4] Jiaxin Cui and Rohan Alexander. 2026. Same Prompt, Different Outcomes: Evaluating the Reproducibility of Data Analysis by LLMs. (2026). arXiv:2602.14349 Preprint..
[5] Han Deng, Anqi Zou, Hanling Zhang, Ben Fei, Chengyu Zhang, Haobo Wang, Xinru Guo, Zhenyu Li, Xuzhu Wang, Peng Yang, Fujian Zhang, Weiyu Guo, Xiaohong Shao, Zhaoyang Liu, Shixiang Tang, Zhihui Wang, and Wanli Ouyang. 2026. Owl-AuraID 1.0: An Intelligent System for Autonomous Scientific Instrumentation and Scientific Data Analysis. (2026). arXiv:2603.29828 Preprint..
[6] Zhuo Diao, Kouma Matsumoto, Linfeng Hou, Masahiro Ohara, Hayato Yamashita, and Masayuki Abe. 2026. Integrating Domain-Specialized Language Models with AI Measurement Tools for Deterministic Atomic-Resolution Experimentation. arXiv:2602.20669 [physics.app-ph] Preprint..
[7] Aarya Doshi, Yining Hong, Congying Xu, Eunsuk Kang, Alexandros Kapravelos, and Christian Kästner. 2026. Towards Verifiably Safe Tool Use for LLM Agents. In *Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER '26)*. ACM, Rio de Janeiro, Brazil, 1–5. doi:10.1145/3786582.3786839
[8] Mohammed Mehedi Hasan, Hao Li, Gopi Krishnan Rajbahadur, Bram Adams, and Ahmed E. Hassan. 2026. Model Context Protocol (MCP) Tool Descriptions Are Smelly! Towards Improving AI Agent Efficiency with Augmented MCP Tool Descriptions. (2026). arXiv:2602.14878 Preprint..

[9] Horace He and Thinking Machines Lab. 2025. *Defeating Nondeterminism in LLM Inference*. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/ Connectionism blog.
[10] Thorsten Hellert, Drew Bertwistle, Simon C. Leemann, Antonin Sulc, and Marco Venturini. 2025. Agentic Artificial Intelligence for Multistage Physics Experiments at a Large-Scale User Facility Particle Accelerator. *Physical Review Research* 8 (2025), L012017. doi:10.1103/jtqy-9jz1
[11] John R. Kitchin. 2025. The Evolving Role of Programming and LLMs in the Development of Self-Driving Laboratories. *APL Machine Learning* 3, 2 (2025), 026111. doi:10.1063/5.0266757
[12] Miles McCain, Thomas Millar, Saffron Huang, Jake Eaton, Kunal Handa, Michael Stern, Alex Tamkin, Matt Kearney, Esin Durmus, Judy Shen, Jerry Hong, Brian Calvert, Jun Shern Chan, Francesco Mosconi, David Saunders, Tyler Neylon, Gabriel Nicholas, Sarah Pollack, Jack Clark, and Deep Ganguli. 2026. *Measuring AI Agent Autonomy in Practice*. https://www.anthropic.com/research/measuring-agent-autonomy
[13] Haochen Pan, Ryan Chard, Reid Mello, Christopher Grams, Tanjin He, Alexander Brace, Owen Price Skelly, Will Engler, Hayden Holbrook, Song Young Oh, Maxime Gonthier, Michael Papka, Ben Blaiszik, Kyle Chard, and Ian Foster. 2025. Experiences with Model Context Protocol Servers for Science and High Performance Computing. (2025). arXiv:2508.18489 Argonne National Laboratory / Globus. Preprint..
[14] Joel Strickland, Arjun Vijeta, Chris Moores, Oliwia Bodek, Bogdan Nenchev, Thomas Whitehead, Charles Phillips, Karl Tassenberg, Gareth Conduit, and Ben Pellegrini. 2026. Talk Freely, Execute Strictly: Schema-Gated Agentic AI for Flexible and Reproducible Scientific Workflows. arXiv:2603.06394 [cs.AI] Preprint..
[15] Aikaterini Vriza, Michael H. Prince, Tao Zhou, Henry Chan, and Mathew J. Cherukara. 2026. Operating Advanced Scientific Instruments with AI Agents That Learn on the Job. *npj Computational Materials* (2026). doi:10.1038/s41524-026-02005-0
[16] Peiran Wang, Ying Li, Yuqiang Sun, Chengwei Liu, Yang Liu, and Yuan Tian. 2026. From Docs to Descriptions: Smell-Aware Evaluation of MCP Server Descriptions. (2026). arXiv:2602.18914 Preprint..
[17] Yong Xie, Kexin He, and Andres Castellanos-Gomez. 2026. Toward Full Autonomous Laboratory Instrumentation Control with Large Language Models. *Small Structures* (2026). doi:10.1002/sstr.202500173
[18] Yiqing Xu and Leo Yang Yang. 2026. Scaling Reproducibility: An AI-Assisted Workflow for Large-Scale Replication and Reanalysis. (2026). arXiv:2602.16733 Preprint..
[19] Chen Yang, Xianyang Zhang, and Jun Chen. 2026. ChatSpatial: Schema-Enforced Agentic Orchestration for Reproducible and Cross-Platform Spatial Transcriptomics. *bioRxiv* (2026). doi:10.64898/2026.02.26.708361 Preprint..

A An Example Skill Document for SEM Periodicity Analysis

Below is a lightly edited excerpt from the skill document for the SEM periodicity tool (Section 4). The model reads this document before interacting with the tool. Parameter names, types, and constraints constitute the typed schema S through which the model's output is channeled.

SEM Analysis — Periodicity & Particle Sizing

Workflow

(1) Read magnification from the SEM info bar (e.g. `x40000`).
(2) Identify the uploaded file from its metadata.
(3) Decide analysis type: periodicity, particle sizing, or both.
(4) Call the tool with exact parameters.
(5) Present results with embedded figures.

Analysis Type Selection

User request	`particle_analysis`
Periodicity, spacing, LIPSS, FFT	omit (default `false`)
Particle/grain size, distribution	`true`
General "analyze this"	`true` (runs both)

Tool Invocation

Periodicity only:

```
run("sem_fft", {
  "file_id":  "<UUID>",
  "mag_label": "x40000"
})
```

With particle sizing:

```
run("sem_fft", {
  "file_id":  "<UUID>",
  "mag_label": "x40000",
  "particle_analysis": true
})
```

Schema Constraints — Exact Parameter Names

Parameter	Type	Rejected aliases
`file_id`	string (UUID)	`image_id`, `id`
`mag_label`	string	`magnification`, `mag`
`particle_analysis`	boolean	`analyze_particles`

Extra parameters (`crop_bottom_px`, `roi`, `preset`) are not accepted unless the user explicitly requests them. The schema rejects any invocation that does not match these exact names and types.

The document instructs the model *what to observe* (Step 1: read magnification), *how to decide* (the analysis-type table), and *how to call the tool* (exact parameter names and types). The model cannot bypass the schema: a call with `magnification` instead of `mag_label` is rejected. This is the mechanism by which $x \in \mathcal{X}$ is enforced at runtime.

B Structured Interview Excerpt

The workflow encoded in each typed tool was extracted through structured interviews. Below is a representative subset of the thirty questions used in the photoluminescence case (Section 4); each targets one analytical decision, and in several the answer contradicted the assumption we had implemented, forcing a correction before deployment.

Q1. **What fitting profile do you use for the emission peak?**
Answer: Voigt.
Impact: We had implemented Lorentzian. Changed to Voigt with free Gaussian and Lorentzian widths.
Q2. **Is the fitting window fixed across all excitation power levels?**
Answer: No. Wider at high power, narrower at low power to avoid fitting noise in the wings.
Impact: We had used a single fixed window. Changed to adaptive windows.
Q3. **What is your R^2 quality threshold for accepting a fit?**
Answer: Approximately 0.90, evaluated visually by overlaying the fit on the data.
Impact: We had implemented a strict 0.95 cutoff. Relaxed to 0.85 to match real practice.
Q4. **How do you extract peak intensity?**
Answer: Screen reader — cursor placed on the peak maximum, raw counts read directly. Not from the integrated area under the fit.
Impact: We had used integrated fit area. Changed the extraction metric to peak height.
Q5. **Do you fit a single power law across the full excitation range?**
Answer: No. Two separate allometric fits ($I = aP^b$), split at a saturation boundary near $\sim$10 μW.
Impact: We had implemented a single fit. Added split fitting with configurable boundary.

Each answer encodes a decision that, guessed wrong, changes the numerical output. The full set of thirty covered preprocessing, model selection, quality control, parameter extraction and post-processing, and the resulting specification was validated against the researcher's manual results to $\Delta \leq 0.02$ on the exponent b.