nature portfolio

Peer Review File

An Agentic Artificially Intelligent X-ray Scientist

Corresponding Author: Professor Zhantao Chen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have demonstrated an agentic workflow to plan, execute, visualize, and iterate to achieve sample alignment. The work is demonstrated in a virtual environment with calculated diffraction patterns and then also experimentally with two samples at an SSRL beamline. MCP tools are utilized to interact with the instrument. Conversation history and git repo are provided to offer details. The work is impressive and represents an important step toward autonomous operation. The authors show that, with prompt engineering and MCP tools, an agentic AI workflow can already perform practical tasks at the beamline. To strengthen the manuscript for Nature Machine Intelligence, the authors can consider providing additional information and discussion.

1. How much can the initial condition be off (motor and lattice), for the AI to still succeed?
2. Why imperfect perk centering if starting with a large initial mismatch when the process is iterative?
3. How long did the AI alignment take, how does it compare to human experts for the real experiment?
4. How well can this be applied to other x-ray experiments, e.g. when there are multiple peaks or scattering present in the detector image.
5. The authors mentioned that one limitation is the substantial upfront effort in prompt engineering, can they elaborate on this process and how easy or hard it would be to adapt this to other tasks or instruments.
6. More LLM models can be compared, including open source models, to show if the agentic system requires large closed-source models.
7. Are there any major obstacles in deploying the agentic system for routine operation?
8. Minor: should briefly review and include related work on agentic AI for experimentation from other user facilities.

(Remarks on code availability)
The code is well organized with proper documentation with README.

Reviewer #3

(Remarks to the Author)
The manuscript reports an “agentic AI scientist” that performs autonomous X-ray diffraction experiments using large language models. The system is trained and tested first in a virtual beamline environment that emulates a six-circle diffractometer and is then deployed on a real synchrotron beamline. The AI identifies reference reflections from XRD and adapts the sample orientation accordingly.

The automation of experimental workflows through AI is an exciting research area. Integrating reasoning-capable LLMs into experimental control could indeed transform facility operations as the authors posit. While I appreciate the timeliness of this topic, and the nice writing of the manuscript, it is difficult to see what major advance is made by this work that would warrant its publication in Nature Machine Intelligence.

1. First and foremost, what contribution is made here beyond off-the-shelf LLM capabilities? The work repeatedly emphasizes “agentic AI” but provides limited details on how the system differs from standard LLM use. From the Methods, it appears that no fine-tuning or domain-specific training was performed; rather, preexisting models (Claude, Gemini) were prompted to issue beamline commands via a custom MCP interface. If the primary contribution is the MCP server, its technical novelty and architecture need to be more rigorously described and benchmarked. Otherwise, the study seems like

an elaborate prompt-engineering exercise.

2. I am not sure LLM use is really needed or sufficiently justified here. Many of the reported tasks (beamline alignment, motor scanning, peak finding) are structured optimization or control problems. It is unclear why a large language model is needed at all, rather than a deterministic or reinforcement-learning approach. The authors should articulate what reasoning capabilities or emergent behaviors of LLMs enable unique functionality that cannot be achieved with existing automation or optimization frameworks.

3. The manuscript claims the AI “learned” through virtual experiments and “transferred” knowledge to real experiments, but did this actually occur? The virtual beamline appears to serve as a testbed for prompt refinement rather than a training environment. So unless I am misunderstanding, the results demonstrate generalization of human-authored prompts, not model learning or policy adaptation.

4. The claim that “all code will be released upon acceptance” is not adequate for a work centered on software and automation. For a technical contribution, public access to the MCP framework, prompts, and simulator source code is essential to assess reproducibility and enable community validation. Without this, the work functions as a demonstration rather than a scientific advance.

5. The authors overstate novelty relative to prior work. There already exists quite a bit of research on autonomous X-ray diffraction, including Bayesian experimental design, reinforcement-learning controllers, and robotic diffractometer alignment. These should be cited and contrasted directly. I suppose it is the use of LLMs that make this work different?

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
Dear Authors.

The manuscript presents an AI agent based on large language models (LLMs) for automating an X-ray sample alignment procedure at a synchrotron beamline. The study addresses a relevant and timely problem, and the results demonstrate the potential of language-model–driven agents for automating experimental workflows. The work is overall interesting and technically competent. However, several aspects of the manuscript would benefit from clarification, additional analysis, and moderation of claims. The following comments are intended to help strengthen the manuscript.

Hereafter are my major points.

In my opinion the abstract currently makes claims that appear broader than what is demonstrated by the presented results and should be moderated accordingly. In particular, the statement that this work represents "the first demonstration of an AI agent which plans and executes experimental tasks, analyzes results, and iterates to achieve a scientific goal" clearly overstates the novelty of the contribution. The manuscript surely demonstrates that an LLM-based agent can autonomously perform a sample alignment task at an X-ray beamline, which is indeed an interesting and meaningful achievement. However, similar agentic frameworks have been reported in prior studies, as the authors themselves acknowledge in the Introduction. The main novelty here lies in applying such a framework to a specific and operationally relevant experimental procedure, rather than in introducing a fundamentally new AI-agent architecture or general scientific planning capability.

In addition, the discussion surrounding the case where the agent identifies an unexpected motor offset and adapts its behavior should be moderated. This event appear to me due to the reasoning capabilities inherent to the underlying LLM, rather than a unique feature of the agent’s workflow or prompt design. Moreover, as this observation is based on a single successful instance among a limited number of trials, it should not be generalized without further evidence. The authors are encouraged to revise the text accordingly and to explicitly acknowledge these limitations in the main manuscript.

Overall, I would recommend that the Abstract and the main part of the manuscript focus more precisely on the concrete achievements of the work, namely demonstrating that an LLM-driven agent can perform and adapt an X-ray sample alignment procedure in a real experimental environment. Broader implications and potential future applications of such AI agents should be more appropriately discussed in the

Discussion or Outlook section, where they can be framed as promising directions for future research rather than current capabilities.

As another major point to improve the current work, I would suggest to consider adding to the manuscript a thorough comparison with human experimentalists performing the same alignment procedure. This would provide valuable context regarding the current performance gap between an AI-driven approach and human expertise, and would help assess the potential and limitations of deploying such agents in real facilities. Consider the following questions:
- Is the proposed AI agent faster or slower than a trained human experimentalist performing the same task? This question is especially relevant given the motivation in the Introduction concerning the limited availability of beamtime and the need for greater productivity.
- When comparing performance, does the human benchmark include an estimate of variability or uncertainty (e.g., error bars from repeated trials under identical conditions)?
- The statement that “this level of accuracy does not yet match the precision typically required in crystallography experiments” would be clearer if the authors provide an estimate of the usual human accuracy.
- How many command calls would a human experimentalist typically perform to achieve such sample alignment, and how does this number compare to the command calls issued by the AI agent?

Regarding the results in Fig. 2(d,e), I believe that a deeper analysis of error distributions and their dependence on experimental conditions and LLM settings would strengthen the quantitative conclusions. In particular, consider the following points:
- The reported results are based on only five trials, which is a very limited sample size. The statistical robustness of the conclusions is therefore weak. The authors should consider performing additional runs to better characterize variability and reliability.
- It would be valuable to analyze whether the accuracy depends on the initial sample position. For instance, does a smaller initial misalignment (therefore fewer required iterations) lead to higher accuracy?
- Repeating the experiment multiple times under identical initial conditions would allow the authors to assess the intrinsic stochasticity due to the LLM’s temperature setting (given the initial setting of the temperature to 1).
- Similarly, have the authors tested running the same trials with temperature = 0 to evaluate if the alignment and lattice parameter accuracy is maintained?

Regarding the terminology and naming used in the manuscript, and in line with the initial comment above, I suggest the following:
- The use of the term AI scientist appears too broad for the demonstrated application. The term "AI X-ray experimentalist" or "AI-driven X-ray experimentation agent" would be more precise and better aligned with the described functionality. This term should be used throughout the manuscript.

Regarding the choice and performance of LLM models, I have the following questions:
- The authors should clarify the rationale for selecting the three LLMs tested. Are these models comparable in terms of capabilities? And so, why comparing them? Or does one have known performance limitations, especially if used as the core of an AI agent?
- It would be valuable to discuss the expected performance of smaller language models (e.g., <20B parameters) in this context. Such an analysis could inform the feasibility of deploying local models directly at beamline facilities, avoiding data transfer to external platforms.

Regarding the time, cost, and efficiency of using the LLM agent in performing these experiments, also in strong relation with the initial major point above, I would consider adding some more quantitative details which would improve the manuscript and help evaluate practical applicability. Please, consider the following:
- Number of LLM requests per experiment.
- Number of iterations required to reach the final alignment (useful for comparison with human experimentalists).
- Initial and final context sizes.
- Whether the context window ever approached the model’s maximum length.

- Context size of the short vs. long prompts.
- Total time needed to complete each experiment.
- Estimated cost of LLM inference per run.
- These metrics would provide a clearer picture of the scalability and efficiency of the proposed approach. In particular, they could be used to answer and discuss the major point mentioned above regarding the comparison with the human experimentalist.

Finally some final additional points to clarify:
- In Fig. 2, the workflow proceeds through "Setup -> Primary reflection -> Secondary reflection." It is not clear whether this sequence is explicitly encoded in the guiding prompt or autonomously decided by the LLM. This should be clarified.
- Regarding the LLM input data, it remains a little unclear whether the LLM receives only raw image data or also some processed tabular data version of the image. The manuscript should clarify what type of detector output is passed to the model and in what format.
- In Figure 2, the caption should explicitly explain what the points, bars, and sticks represent.
- Regarding the real laboratory experiment, the authors mention the use of Claude Opus rather than one of the previously tested LLMs. Clarification is needed on this choice — was it for compatibility, availability, or performance reasons?
- Regarding the crystal structure data, the real experiment uses a file from the COD database, while earlier experiments used data from the Materials Project. The rationale for this change should be explained.
- In the "Agentic AI models" SI section, the described parameter settings appear to refer specifically to Gemini. It should be clarified whether the same setup was used for all models. Or include the setting used for Anthropic's models.
- I'd suggest calling Claude Sonnet and Opus and Gemini Flash simply LLMs.
- Regarding the command extraction from the LLM output, the current keyword-based parsing of LLM outputs could be prone to errors. The authors might acknowledge this limitation and suggest the use of structured outputs (e.g., via JSON or pydantic models) as a potential improvement.
- Regarding the real experiments, were them executed only once, or were multiple times performed to assess reproducibility?

Thanks!
Francesco Ricci

(Remarks on code availability)
The code is not available. The Authors state: All data and source code necessary to reproduce the reported results will be made publicly available upon acceptance of this work.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
All comments are addressed, additional tests are provided, generalization to other type of tasks are discussed. More x-ray and large facilities related literature with AI and automation can still be added.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have made substantial revisions that address several of my previous concerns. The manuscript now more clearly positions the work relative to prior literature and clarifies the scope/novelty of their contribution.

I find the paper technically sound, and the topic an important one. My remaining reservation relates primarily to the level of conceptual advance. The approach largely integrates existing large language models with experimental control tools and a simulated environment, rather than introducing fundamentally new machine learning methods. While the integration is thoughtful and the real-world beamline demonstration is compelling, it is less clear whether the work represents a methodological advance of the type typically emphasized in Nature Machine Intelligence.

That said, the work provides an interesting proof-of-concept for agentic AI in experimental science, and the revisions have

improved the manuscript overall. Whether this contribution meets the bar for sufficiently novelty to be published in this journal is ultimately a decision for the editor. I have no other major concerns.

(Remarks on code availability)
The authors have now made available their code on Github, which appears nicely documented.

Reviewer #4

(Remarks to the Author)
The authors have thoroughly revised the manuscript, addressing all of my concerns and incorporating the majority of my suggestions. They provided satisfactory responses to my previous queries, and the current version significantly clarifies the study's scope and objectives.

Notably, the addition of new experiments and edge cases has strengthened the statistical analysis and error reporting. Furthermore, the expanded introduction and updated literature review better position the work within the field.

In light of these improvements, I believe the manuscript is now acceptable for publication.

(Remarks on code availability)
The associated repository is well-documented, featuring a comprehensive README that provides a clear, step-by-step installation and setup guide for the package, the MCP server, and the LLM provider. The authors have also included an IPython notebook to facilitate testing of the core functionalities. Furthermore, a graphical user interface to executed a simulation is provided to the user. Most importantly, the inclusion of the "long" prompt (mentioned in the manuscript), initial crystal structures, and configuration files ensures the study’s results are fully reproducible.

Responses to Reviewers' Comments

Manuscript ID: NATMACHINTELL-A25083858
Manuscript Title: An Agentic Artificially Intelligent X-ray Scientist

We would like to thank all the reviewers for their time and effort in providing valuable comments and questions, which have greatly helped us to further improve the manuscript. Below, please find our point-by-point responses highlighted in blue color. All changes made to the revised manuscript (a separate PDF file) are marked in green color.

Table of Contents

Reviewer #1

Reviewer #1 - Remarks to the Author

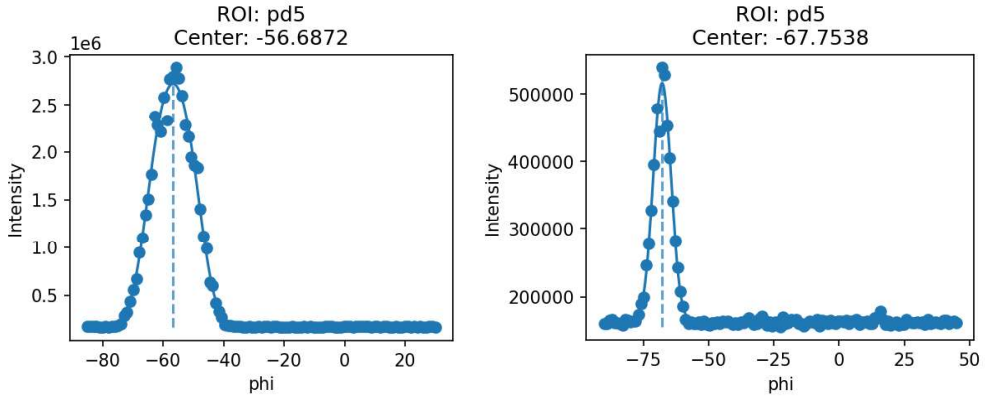
The authors have demonstrated an agentic workflow to plan, execute, visualize, and iterate to achieve sample alignment. The work is demonstrated in a virtual environment with calculated diffraction patterns and then also experimentally with two samples at an SSRL beamline. MCP tools are utilized to interact with the instrument. Conversation history and git repo are provided to offer details. The work is impressive and represents an important step toward autonomous operation. The authors show that, with prompt engineering and MCP tools, an agentic AI workflow can already perform practical tasks at the beamline. To strengthen the manuscript for Nature Machine Intelligence, the authors can consider providing additional information and discussion.

We sincerely thank the reviewer for the careful review and constructive feedback, as well as for recognizing the significance of our efforts. We provide below detailed point-by-point responses supported by additional benchmarking, and we believe that the manuscript has been substantially improved as a result of addressing these comments.

1. How much can the initial condition be off (motor and lattice), for the AI to still succeed?

We performed an additional feasibility check with more challenging initial in-plane orientations (φ = -55° and -70° with final alignment errors being 0.70° and 2.74°, respectively, compared with the default φ = -41° used in all other benchmark runs) with Gemini 2.5 Flash. We acknowledge that this is a limited-sample test conducted under nominal motor behavior, intended solely to demonstrate the feasibility of handling more challenging initial conditions that require wider motor scans.

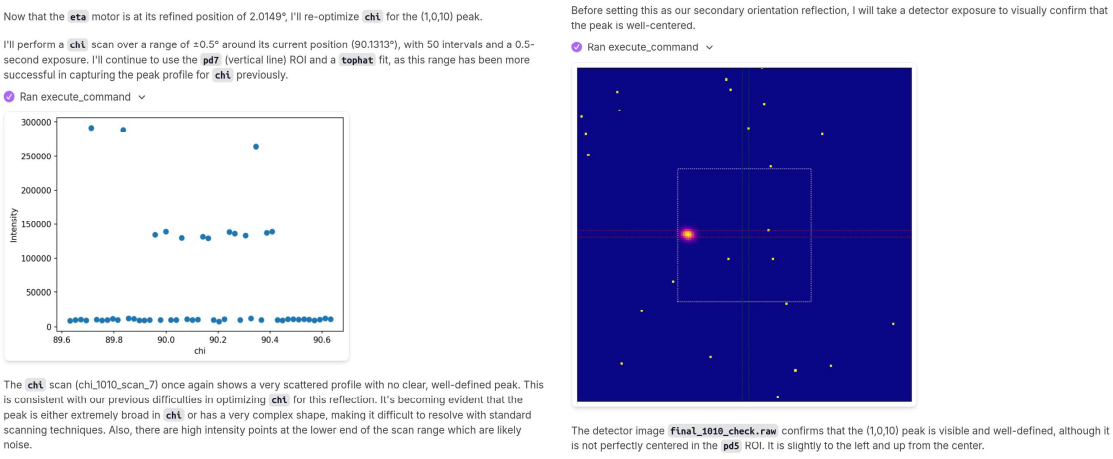
As shown in the figures below, in both cases, the AI agent successfully identified the correct diffraction peak by autonomously extending the φ scan range, indicating that the agent can in principle tolerate larger initial misorientations. Given the relatively wide scan ranges used during alignment, moderate lattice-parameter offsets are therefore expected to have only a mild impact on performance.



2. Why imperfect perk centering if starting with a large initial mismatch when the process is iterative?

We thank the reviewer for raising this important question. The most common failure mode arises when the AI agent does not scan a sufficiently wide range of either the δ or φ/χ motor. In such cases, the true diffraction peak never crosses the predefined line ROI at all (the line ROI is intentionally very thin and narrower than most peak widths), and therefore the resulting scan does not contain the actual peak profile. Under these circumstances, the agent may mistakenly interpret noise fluctuations as a valid peak and align the sample to these spurious signals.

An example is shown below, where the χ motor scan range is too narrow to bring the peak across the vertical line ROI. Therefore, although the AI concludes that the peak has been horizontally aligned, the true peak remains far left of the center line. This contributes to a significant residual alignment error of 18.5254°, even though the overall procedure is iterative. Such failure modes could potentially be mitigated in future implementations by improving the curve-fitting MCP tools. For example, when the fitted peak height or contrast is insufficient, the fitting tool could return an explicit warning indicating low confidence in the peak identification and prompt the agent to re-examine the detector image or expand the scan range to capture the true peak. We have added additional discussion on this example in the Supplementary Information of revised manuscript to discuss this failure mode and potential routes toward future improvements.



3. How long did the AI alignment take, how does it compare to human experts for the real experiment?

Based on data from actual experimental logs, a representative beamline alignment in a real experiment performed by the AI agent took approximately 40 minutes, where a

substantial fraction of this total time corresponds to mandatory waiting periods associated with hardware motion such as detector exposure during motor scans. In comparison, virtual experiments conducted by the AI agent typically take less than 10 minutes where all experimental actions are almost instantaneously finished, which suggests that the mandatory waiting time might take about ~75% of the total experimental time.

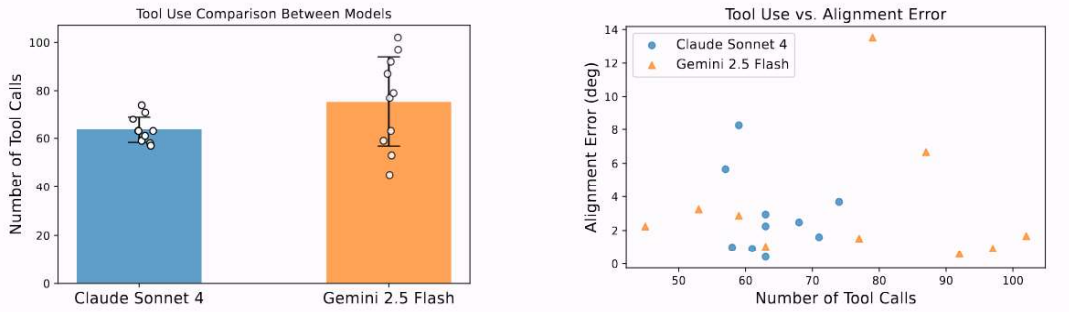
To provide additional context for the agent's operational behavior, we report the average number of tool-use calls made during alignment in the Supplementary Information (Fig. S2; please find the discussion attached below). While this does not constitute a direct time comparison with human experimentalists, we hope it offers a quantitative proxy for the procedural complexity and thoroughness of the alignment process made by AI.

S2 Tool-use calls

We report the number of tool-use calls issued by the agent in Figure S3a. Each tool-use call corresponds to an explicit interaction between the LLM and the experimental control interface (e.g., motor motion, detector acquisition, or data query), and thus serves as a coarse measure of the interaction complexity of the agentic workflow.

As shown, the two models require comparable numbers of tool-use calls to complete the alignment task, indicating that the overall structure of the agentic workflow is similar across models. We note that the absolute number of tool calls depends on model-specific reasoning behavior and is not intended as a fine-grained efficiency benchmark.

To further examine whether the number of tool-use calls is directly related to alignment performance, we plot the final alignment error as a function of the total number of tool-use calls (Figure S3b). Across the tested cases and models, we do not observe a strong or systematic correlation between the number of tool-use calls and the final alignment errors.



(a) Number of tool-use calls (b) Tool-use calls vs. alignment error

Figure S3: **(a)** Comparison of numbers of tool-use calls by two models. The bar heights indicate mean values across runs, error bars denote ±1 standard deviation, and circular markers represent individual runs. **(b)** Final alignment error as a function of the total number of tool-use calls.

We have observed that the AI agent tends to follow the instructions thoroughly, which can lead to longer alignment times, especially when unexpected situations arise that require on-the-fly debugging. Therefore, the reported alignment duration should be viewed as a conservative estimate that reflects robustness-oriented behavior by AI.

4. How well can this be applied to other x-ray experiments, e.g. when there are multiple peaks or scattering present in the detector image.

We thank the reviewer for raising this interesting question regarding the applicability of the proposed approach to more complex X-ray experiments involving multiple peaks or additional scattering features.

In general, the AI X-ray scientist is robust to irrelevant signals when appropriate experimental context is provided. For example, when the AI is instructed that the measurement involves a single crystal, it is able to ignore the diffuse background features or powder-like rings that are not consistent with the expected diffraction geometry, as shown in the alignment of the second reflection (1,0,10) in Figure 4. However, as the reviewer correctly points out, scenarios in which multiple visually similar peaks are present pose a greater challenge.

To explicitly probe this limitation, we have added discussion of a dedicated multi-grain test case in the revised manuscript and Supplementary Information (see section titled "Alignment of multi-grain samples"). In this case, two diffraction peaks with slight relative orientation offsets are observed simultaneously on the detector, leading to double-peaked motor scan profiles. While the AI agent remains capable of completing the alignment procedure, we found that the current peak-fitting tool can converge to the centroid of multiple peaks rather than selectively identifying the dominant reflection, resulting in degraded accuracy. This behavior highlights a limitation of the present tool interface within the agentic workflow.

We now note in the revised manuscript that such cases point to a clear avenue for future improvement: providing more expressive and task-aware MCP tools, such as peak identification and classification utilities that can distinguish between competing reflections (e.g., selecting the strongest peak or enforcing crystallographic consistency). We expect that augmenting the toolset in this manner will further improve performance in complex scattering environments, including multi-grain samples and experiments with overlapping features.

These results suggest that our agentic framework is broadly applicable to a range of X-ray experiments, but that its effectiveness in more challenging scenarios depends on the availability of appropriate domain-specific analysis tools. We have revised the manuscript to reflect this scope and to appropriately frame multi-peak experiments as both a demonstrated capability and an important direction for future development.

5. The authors mentioned that one limitation is the substantial upfront effort in prompt engineering, can they elaborate on this process and how easy or hard it would be to adapt this to other tasks or instruments.

We thank the reviewer for raising this question regarding the transferability of prompt engineering in our work.

We indeed invested a substantial upfront effort in preparing the guidance prompts. This process began with constructing a sufficiently faithful digital twin (virtual beamline) of the

experimental setup. By “sufficient”, we mean that the virtual environment could reproduce the essential experimental operations and feedback loops needed for realistic interactions; here, capturing every physical detail is not necessary. This virtual platform enabled safe and efficient prompt development by allowing the AI agent to experience realistic responses to its actions. Note that such a digital twin is not strictly required if enough physical instrument time and robust safety interlocks are available. However, given limited beamtime and safety considerations, this approach was both practical and necessary in our case.

The prompt engineering process itself involved several stages: (i) We designed individual MCP tool interfaces as well as the corresponding documentation explaining the purpose, input, output, and appropriate usage of each tool. This documentation is often supplemented with examples. This enables the AI agent to understand not only how to invoke tools syntactically, but also when and why each tool should be used. (ii) We composed high-level workflow instructions to stitch these tools together into a coherent experimental procedure. Notably, we did not hard-code the experimental workflow, but chose to leave execution order, iteration, and decision-making fully to the agent, which fundamentally differentiate our approach from the traditional script-based automation. Finally, a significant portion of the effort was finally devoted to identifying potential failure modes during simulated operations and refining the prompts to guide the agent toward safer and more reliable behavior. We have added discussion on this point in the “Agentic AI models” portion of the Methods section.

We emphasize that a well-designed digital twin greatly facilitates the development process and allows the guidance prompts developed in the virtual environment to be transferred effectively to the real beamline with minimal modifications. However, we acknowledge that prompt preparation is task- and instrument-dependent, so that adaptation to other experiments would require redefining appropriate tools and workflows. The automation of prompt and tool-design processes, potentially through meta-learning or AI-assisted prompt generation, is a promising direction for reducing the barrier to deploying similar agentic systems across a wider range of instruments and experimental tasks.

6. More LLM models can be compared, including open source models, to show if the agentic system requires large closed-source models.

We thank the reviewer for this thoughtful suggestion regarding broader model comparisons. Accordingly, we conducted an additional exploratory evaluation using GPT 5 mini under the same base conditions used in our existing benchmarks. We performed five independent runs. Two of the five runs completed successfully, with one successful run starting from the default in-plane orientation (φ = -41°) with a final error of 5.02° and one from a more challenging initial condition (φ = -70°) with a final error of 2.15°. The remaining runs failed to complete the alignment with errors being larger than 10 degrees.

Consistent with our observations across the other models, the dominant failure mode in these runs was associated with motor-limit cases during wide scans, in which a motor reached its positional boundary and did not automatically return to its nominal starting

position. Recovery from such unmodeled instrument states requires adaptive reasoning and exploratory recovery rather than deterministic execution. GPT 5 mini seems to be not as good as the other two reported models in terms of recovery from an unexpected motor status. However, we believe that when this motor-limit behavior is mitigated, as would typically be the case in production beamline control software, the overall success rate will improve substantially across the various models.

Our goal in the present manuscript is not to rank the performance of various language models, but to demonstrate that a reasoning and tool-use-capable LLM can autonomously carry out non-trivial X-ray experiments. Within this scope, we have additionally expanded the Gemini 2.5 Flash benchmarks to over 30 runs in the revised manuscript, which provide a reasonably representative statistical characterization of the agentic workflow. We appreciate the reviewer's understanding of our study in keeping with this scope of the manuscript.

Moreover, following the reviewer's suggestion, we have performed exploratory tests with a smaller open-source, vision–language model (Qwen3-VL-8B). While this model was able to invoke tools at a syntactic level, it was unable to carry out the experimental tasks in a meaningful manner, failing early due to incorrect reasoning and inappropriate parameter updates. To support this assessment, we have attached representative conversation history of Qwen3-VL-8B run as supplementary material with this response to illustrate the failure mode.

Taken together, our analysis suggests that the complexity of the task (requiring multimodal perception, iterative reasoning over long horizons, and reliable tool use under partial observability) currently exceeds the capabilities of smaller models when used off-the-shelf. Note, however, that this does not preclude the use of smaller models in the future. Fine-tuning lightweight models via supervised fine-tuning (SFT), reinforcement learning, or task-specific adaptation represents a promising future direction, particularly for local deployment at user facilities with constraints on data transfer, cost, privacy, and latency.

7. Are there any major obstacles in deploying the agentic system for routine operation?

The primary obstacles to routine operation are not fundamental limitations of the agentic framework itself, but rather challenges related to workflow design, safety, and risk management. Based on our observations, many failed or non-ideal alignment cases could be significantly improved through appropriate refinements of the MCP tools. For example, instead of fitting noise fluctuations, peak-fitting tools could return explicit warning messages indicating that no reliable peak has been identified and prompt the agent to expand the scan range or re-examine the detector image. Our experience is that if the experimental workflow is well defined and appropriate tools are provided, the agent can carry out the task reliably.

Safe deployment requires robust software- and hardware-level safeguards, including motor limits, collision avoidance mechanisms, and interlock systems that prevent potentially damaging actions. In our current implementation, the AI agent operates with indirect execution through human experimentalists, who relay commands to the

beamline control system. During experiments, potential collision risks are mitigated primarily through existing hardware limits and visual monitoring via cameras. Looking ahead, more systematic integration with established control software and facility safety infrastructure will be needed for deploying such agentic systems in routine user-facility operation. Different beamlines and facilities may require tailored safety integration schemes depending on instrumental configurations and operational policies.

8. Minor: should briefly review and include related work on agentic AI for experimentation from other user facilities.

The reviewer raises a sensible point. Accordingly, we have expanded the Introduction section to provide a deeper review of related work on agentic AI and autonomous experimentation.

We thank the reviewer for the time and care they devoted to evaluating our manuscript and for their constructive questions. We hope that the revised manuscript satisfactorily addresses all their concerns.

Reviewer #1 - Remarks on Code Availability

The code is well organized with proper documentation with README.

We thank the reviewer for the positive assessment of the code organization and documentation.

Reviewer #3

Reviewer #3 - Remarks to the Author

The manuscript reports an “agentic AI scientist” that performs autonomous X-ray diffraction experiments using large language models. The system is trained and tested first in a virtual beamline environment that emulates a six-circle diffractometer and is then deployed on a real synchrotron beamline. The AI identifies reference reflections from XRD and adapts the sample orientation accordingly.

We are thankful for the reviewer’s thoughtful review and for all the insightful questions and comments. We have provided point-by-point responses below.

The automation of experimental workflows through AI is an exciting research area. Integrating reasoning-capable LLMs into experimental control could indeed transform facility operations as the authors posit. While I appreciate the timeliness of this topic, and the nice writing of the manuscript, it is difficult to see what major advance is made by this work that would warrant its publication in Nature Machine Intelligence.

We appreciate the reviewer’s recognition of our work’s timeliness and writing quality. Regarding the advance made by our work, we believe that our study is a first demonstration showing that modern, general-purpose reasoning LLMs can provide reliable, autonomous experimental controllers in a real synchrotron beamline environment, performing multi-step alignment procedures that require planning, interpretation, error recovery, and adaptation in the presence of hardware constraints and unexpected experimental outcomes.

We emphasize that our study does not simply present a simulation or an offline benchmark. Our AI agentic system closes the loop between perception (experimental measurements), reasoning (task decomposition and decision making), and action (instrument control) on an actual production level beamline at a major user facility (SSRL). Additionally, we believe that our demonstrated virtual-to-real agentic AI workflow, which can operate robustly without task-specific retraining when transferred from the virtual beamline to the real beamline, represents a substantial conceptual advance for AI-driven experimentation and facility automation.

For these reasons, we hope that the reviewer will agree with us that our study presents a significant advance in the field.

1. First and foremost, what contribution is made here beyond off-the-shelf LLM capabilities? The work repeatedly emphasizes “agentic AI” but provides limited details on how the system differs from standard LLM use. From the Methods, it appears that no fine-tuning or domain-specific training was performed; rather, preexisting models (Claude, Gemini) were prompted to issue beamline commands via a custom MCP interface. If the primary contribution is the MCP server, its technical novelty and architecture need to be more rigorously described and benchmarked. Otherwise, the study seems like an elaborate prompt-engineering exercise.

We agree with the reviewer that we do not modify or fine-tune the underlying language models. However, we respectfully emphasize that the absence of model fine-tuning does not diminish the contribution of our work. Our contribution lies not in altering model weights, but in developing and experimentally validating a closed-loop agentic control architecture that enables general-purpose reasoning models to autonomously operate complex experimental hardware in a real beamline environment.

The novelty of the work is thus not based on the MCP interface in isolation, nor the use of prompting per se, but rather the integration of multiple components into a functioning, end-to-end experimental control system, including:

1. A structured tool abstraction that exposes beamline capabilities (motors, scans, detectors, and metadata) in a machine-interpretable form, and
2. A generalizable framework that allows the LLM to monitor experimental state over long horizons, receive real-time feedback from physical instruments, and execute experimental actions to achieve experimental goals.

It is indeed true that our system involves prompt engineering and in-context learning. We hope, however, that the reviewer will see this as a strength rather than a limitation of our system, which demonstrates that meaningful experimental autonomy can be achieved without domain-specific retraining. This significantly lowers the barrier to adoption and facilitates transfer to other instruments and facilities. Our work thus shows that general-purpose reasoning models, when properly embedded in a closed-loop control architecture, can enable qualitatively new forms of experimental automation.

2. I am not sure LLM use is really needed or sufficiently justified here. Many of the reported tasks (beamline alignment, motor scanning, peak finding) are structured optimization or control problems. It is unclear why a large language model is needed at all, rather than a deterministic or reinforcement-learning approach. The authors should articulate what reasoning capabilities or emergent behaviors of LLMs enable unique functionality that cannot be achieved with existing automation or optimization frameworks.

We appreciate the reviewer's question regarding the extent to which use of LLMs is necessary. We agree that individual sub-tasks involved in the experiments could indeed be addressed, in principle, using conventional scripted automation.

However, we respectfully emphasize that the central challenge addressed in this work is not a single, well-defined optimization problem, but rather the coordination of multiple experimental steps under uncertain condition, including unknown initial sample state, motor offsets, and signal conditions. In practice, real beamline operation frequently involves unexpected scenarios that require flexible and context-aware decision making. Language plays a central role here as a carrier of abstract reasoning and long-horizon decision making, allowing the AI-agent-based experimental controller to sense experimental context, plan actions, and handle contingencies in a reliable and robust manner. This capability is critical for sustaining continuous experimental operation in complex environments where the sequence of actions cannot be fully specified in advance given potential experimental exceptions.

Importantly, the demonstrated experimental procedures are not hard-coded or programmatically enforced. The LLM autonomously determines when to scan, how to interpret outcomes, and how to revise subsequent actions based on observed results, so that the system can adapt its strategy when expected peaks are not found, signals are weak, or large offsets are encountered, without relying on pre-enumerated exception handling. In fact, explicitly enumerating all potential exceptions could become impractical with highly complex control logic, yet the demonstrated LLM-based approach provides a viable scheme.

While reinforcement learning approaches could in principle address aspects of this problem, they typically require extensive task-specific training, carefully designed reward functions, and large amounts of interaction data. In contrast, the successful runs presented in this work demonstrate zero-shot experimental automation in a real synchrotron X-ray environment, which, to the best of our knowledge, have not previously been demonstrated in this setting.

Finally, we do not claim that LLMs replace classical control or optimization algorithms. Rather, they serve as a cognitive coordination layer that integrates and sequences such tools in situations where rigid automation or scripted workflows become brittle. Moreover, our results suggest that improvements in the underlying experimental tools and interfaces can further enhance LLM-driven performance. We have revised the Introduction to more clearly articulate this role and to highlight the reasoning-driven autonomy enabled by LLMs in our study.

3. The manuscript claims the AI “learned” through virtual experiments and “transferred” knowledge to real experiments, but did this actually occur? The virtual beamline appears to serve as a testbed for prompt refinement rather than a training environment. So unless I am misunderstanding, the results demonstrate generalization of human-authored prompts, not model learning or policy adaptation.

We thank the reviewer for pointing out this important distinction. We agree that no model weights are updated during virtual or real experiments, and that the virtual beamline is not a training environment in the conventional machine-learning sense. We have revised the manuscript to clarify this point and to avoid using “learning” in a gradient-based or policy-optimization sense.

In our work, all adaptation occurs through in-context reasoning: the LLM interprets experimental observations, maintains state across tool calls, and adjusts subsequent actions accordingly, without any parameter updates. The virtual beamline serves as a testbed for evaluating and refining the agent’s prompting and tool-use strategy, while the deployment on the real beamline demonstrates generalization of the agentic workflow and task structure. We have revised the manuscript and added a new paragraph in the Discussion section accordingly to make this distinction explicit.

We agree that the term “transfer” could be misinterpreted as implying model training or policy learning. To avoid this ambiguity, we have revised the abstract to clarify that the agentic workflow and prompting strategy developed in the virtual beamline were directly deployed on the real beamline without any model retraining or parameter updates. We

further note that only minor environment-specific adjustments to tool-call were required to match the real beamline control interface, while the agent's prompting, reasoning structure, and decision logic remained unchanged.

We hope that the revised framing of our study adequately addresses the thoughtful question raised by the referee.

4. The claim that "all code will be released upon acceptance" is not adequate for a work centered on software and automation. For a technical contribution, public access to the MCP framework, prompts, and simulator source code is essential to assess reproducibility and enable community validation. Without this, the work functions as a demonstration rather than a scientific advance.

We would like to clarify that the source code was made available during the review process. The code was initially uploaded to CodeOcean but was later withdrawn due to technical issues, after which it was re-shared via a GitHub repository (https://github.com/zhantaochen/llm4xray). We apologize for any confusion this may have caused. We have revised the manuscript to point readers to our code repository.

5. The authors overstate novelty relative to prior work. There already exists quite a bit of research on autonomous X-ray diffraction, including Bayesian experimental design, reinforcement-learning controllers, and robotic diffractometer alignment. These should be cited and contrasted directly. I suppose it is the use of LLMs that make this work different?

We thank the reviewer for raising concern regarding the positioning of our study relative to prior work. In response, we have revised the Introduction section to more explicitly situate our work within the broader literature including Bayesian experimental design, reinforcement-learning controllers, and robotic diffractometer alignment. To avoid overstatement of our work's novelty, we have added an explicit statement in the Introduction and also a paragraph in the Discussion section to clearly state that our contribution is not a new agent architecture or learning algorithm. We believe this revised framing more accurately reflects the scope and we hope that it addresses the reviewer's concern.

Moreover, we would like to emphasize a critical difference with our work and the reason that we believe this development will be of interest to the broad readership of *Nature Machine Intelligence*. We are not simply demonstrating autonomous experimental design, or robotic alignment, or an optimization problem. This work uses a reasoning-capable LLM to formulate a plan given a material's crystallographic peak information, determine the best course of action to measure this material, observe errors and decide how to handle them based on observing the live data stream, and complete a full measurement at a real-world, large-scale X-ray facility.

Our work embeds artificial intelligence directly into a scientific machine/instrument, enabling it to make decisions and adapt to unmodeled conditions at the level of the experimental workflow, rather than executing a fixed control or optimization routine. This capability, together with its demonstration in a real user experiment, distinguishes our

work from prior approaches. More broadly, these features make the contribution relevant to a wide audience interested in deploying machine intelligence for scientific discovery, well beyond the specific application domain.

Once again, we sincerely thank the reviewer for their time, careful reading, and constructive feedback. We hope that our responses and the corresponding revisions to the manuscript have satisfactorily addressed the concerns raised. We hope that the reviewer will be willing to reconsider our manuscript for publication in *Nature Machine Intelligence*.

Reviewer #4

Reviewer #4 - Remarks to the Author

The manuscript presents an AI agent based on large language models (LLMs) for automating an X-ray sample alignment procedure at a synchrotron beamline. The study addresses a relevant and timely problem, and the results demonstrate the potential of language-model–driven agents for automating experimental workflows. The work is overall interesting and technically competent. However, several aspects of the manuscript would benefit from clarification, additional analysis, and moderation of claims. The following comments are intended to help strengthen the manuscript.

We thank the reviewer for the careful evaluation of our manuscript and for the constructive and detailed feedback. We appreciate the positive assessment of the relevance and technical soundness of the work, and we found the reviewer's suggestions and comments to be very helpful. Please find below our point-by-point responses to the reviewer's comments.

Hereafter are my major points.

In my opinion the abstract currently makes claims that appear broader than what is demonstrated by the presented results and should be moderated accordingly. In particular, the statement that this work represents "the first demonstration of an AI agent which plans and executes experimental tasks, analyzes results, and iterates to achieve a scientific goal" clearly overstates the novelty of the contribution. The manuscript surely demonstrates that an LLM-based agent can autonomously perform a sample alignment task at an X-ray beamline, which is indeed an interesting and meaningful achievement. However, similar agentic frameworks have been reported in prior studies, as the authors themselves acknowledge in the Introduction. The main novelty here lies in applying such a framework to a specific and operationally relevant experimental procedure, rather than in introducing a fundamentally new AI-agent architecture or general scientific planning capability.

We thank the reviewer for this helpful feedback regarding scope, novelty, and positioning within the broader literature on autonomous and agentic experimentation. In response, we have substantially revised the Introduction to more clearly acknowledge prior work on autonomous laboratories, embodied robotics, and agentic AI systems, including recent applications of LLM-driven agents to real-world experiments.

Specifically, we added a new paragraph in the Introduction that better clarifies our work's contribution alongside representative efforts in autonomous chemistry, biology, materials science, and embodied AI. We have added a substantial number of additional references from closely relevant fields such as self-driving laboratories and agentic scientific workflows.

We have also moderated claims of novelty throughout the manuscript, especially in the Introduction and Discussion sections, and further clarified that our contribution is not a

new AI-agent architecture or general scientific planning capability, but rather a focused effort and demonstration of deploying a reasoning-capable LLM as a closed-loop experimental controller for a concrete and operationally relevant task on a real synchrotron beamline.

We believe these revisions more accurately reflect the scope of the work, avoid over-broad and over-stated claims, and clearly distinguish our contribution from prior studies, in line with the reviewer's suggestions.

In addition, the discussion surrounding the case where the agent identifies an unexpected motor offset and adapts its behavior should be moderated. This event appear to me due to the reasoning capabilities inherent to the underlying LLM, rather than a unique feature of the agent's workflow or prompt design. Moreover, as this observation is based on a single successful instance among a limited number of trials, it should not be generalized without further evidence. The authors are encouraged to revise the text accordingly and to explicitly acknowledge these limitations in the main manuscript.

We agree with the reviewer that this observation is based on a single real-beamline trial and should not be generalized as a statistically established capability of autonomous fault diagnosis. We have revised the manuscript and added a paragraph to explicitly acknowledge this limitation and to moderate the discussion accordingly.

We also agree that the adaptive behavior arises from the reasoning capabilities of the underlying LLM rather than from a novel learning or fault-detection mechanism introduced in our workflow. We have now clarified that our contribution lies in demonstrating how such reasoning can be operationalized in a real-world, closed-loop experimental control setting through persistent experimental state, structured tool interfaces, and real-time feedback, rather than in introducing new model-level intelligence.

Overall, I would recommend that the Abstract and the main part of the manuscript focus more precisely on the concrete achievements of the work, namely demonstrating that an LLM-driven agent can perform and adapt an X-ray sample alignment procedure in a real experimental environment. Broader implications and potential future applications of such AI agents should be more appropriately discussed in the Discussion or Outlook section, where they can be framed as promising directions for future research rather than current capabilities.

We agree with the reviewer's assessment and appreciate this constructive guidance. In response, we have revised the Abstract and main text to focus more precisely on the concrete achievements demonstrated in this work. We have also moderated broader claims and relocated discussions of more general implications and potential future applications of agentic AI systems to the Discussion and Outlook sections. We hope these revisions satisfactorily address the reviewer's comment.

As another major point to improve the current work, I would suggest to consider adding to the manuscript a thorough comparison with human experimentalists performing the same alignment procedure. This would provide valuable context regarding the current performance gap between an AI-driven approach and human expertise, and would help assess the potential and limitations of deploying such agents in real facilities. Consider the following questions:

- Is the proposed AI agent faster or slower than a trained human experimentalist performing the same task? This question is especially relevant given the motivation in the Introduction concerning the limited availability of beamtime and the need for greater productivity.

- When comparing performance, does the human benchmark include an estimate of variability or uncertainty (e.g., error bars from repeated trials under identical conditions)?

- The statement that “this level of accuracy does not yet match the precision typically required in crystallography experiments” would be clearer if the authors provide an estimate of the usual human accuracy.

- How many command calls would a human experimentalist typically perform to achieve such sample alignment, and how does this number compare to the command calls issued by the AI agent?

We thank the reviewer for their valuable suggestions. We agree that comparison to human experimentalists can provide important context for evaluating practical utility. At the same time, a rigorous and fair human benchmark for single-crystal alignment is nontrivial and highly operator-dependent and would require multiple trained experimentalists and repeated trials under tightly controlled initial conditions to be statistically meaningful. We feel that such a study is beyond the scope of the present proof-of-feasibility demonstration.

To provide practical context without over-claiming, we have now added quantitative operational metrics for the agentic runs (total wall-clock time, number of reasoning iterations/tool-use calls, and success rate statistics in simulation) in the revised manuscript and Supplementary Information as will be discussed below in our responses. In particular, the real beamline demonstration required approximately 40 minutes wall time, which is dominated by physical constraints, such as the mandatory waiting time during motor scans, and by the human-in-the-loop safety relay during early deployment, rather than by LLM inference time. We also report the number of tool-use calls per run as a coarse measure of interaction complexity. We believe this will provide valuable information for rough comparison with human experimentalists’ efficiency.

With respect to alignment accuracy requirements, we have revised the text to clarify that the achieved accuracy is sufficient for demonstrating autonomous closed-loop alignment, while precision requirements for full crystallographic experiments can be substantially more stringent depending on the measurement type. We now frame these requirements qualitatively and avoid implying a universal “human baseline” in the absence of a dedicated benchmark.

We view comprehensive human benchmarking, including multi-operator variability and standardized initial conditions, as an important direction for future work and for facility-specific deployment evaluation. We appreciate the reviewer's understanding on this point.

Regarding the results in Fig. 2(d,e), I believe that a deeper analysis of error distributions and their dependence on experimental conditions and LLM settings would strengthen the quantitative conclusions. In particular, consider the following points:

We thank the reviewer for their constructive suggestions. In response, we have added a substantial amount of new benchmarking trials (from 10 to 40 trials) to strengthen our quantitative discussions.

- The reported results are based on only five trials, which is a very limited sample size. The statistical robustness of the conclusions is therefore weak. The authors should consider performing additional runs to better characterize variability and reliability.

We have substantially expanded the quantitative benchmarking beyond the originally reported trials and have expanded the benchmark along experimentally meaningful axes that probe variability and robustness more directly.

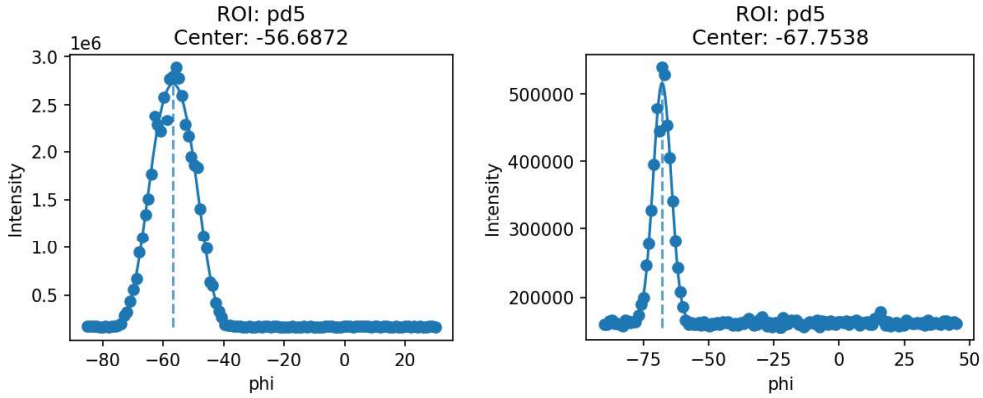
Specifically, we evaluated performance under (1) higher motor offsets, which introduce systematic discrepancies between nominal and actual motor positions and are more challenging than simple shifts in initial alignment, (2) multi-grain diffraction conditions, which increase ambiguity in peak identification, (3) reduced reasoning ("thinking") budget, which stresses the agent's ability to maintain coherent multi-step decision making; and (4) reduced model temperature from the default value 1.0 to 0.5.

- It would be valuable to analyze whether the accuracy depends on the initial sample position. For instance, does a smaller initial misalignment (therefore fewer required iterations) lead to higher accuracy?

We appreciate the reviewer's question regarding whether alignment accuracy depends on the initial sample orientation (e.g., different starting φ angles). A systematic sweep over multiple initial in-plane orientations (e.g., -51°, -41°, -31°, etc.), each repeated multiple times, would indeed provide a more complete characterization of this dependency, but would require a substantially larger number of runs.

We note that we did not perform a systematic sweep over multiple initial in-plane orientations in the present study. Instead, we selected an initial in-plane orientation of approximately -41°, which we found to be a representative, intentionally challenging case: under the allowed φ motor range, the corresponding diffraction peak appears only once, preventing trivial local refinement and requiring correct global reasoning to identify the reflection. We therefore view this configuration as a reasonable stress test rather than an optimized or easy starting condition. We conducted additional alignment runs with Gemini 2.5 Flash starting from larger initial in-plane orientations (φ = -55° and φ = -70°) with final alignment error of 0.70° and 2.74°, respectively, under normal motor

scan behavior. These tests were not intended as a statistical study, but rather as targeted feasibility checks. In both cases, the AI agent successfully identified the correct diffraction peak by autonomously extending the φ scan range and completed the alignment, indicating that the agent can, in principle, tolerate substantially larger initial misorientations.



We acknowledge that this choice does not explicitly map the performance as a function of initial orientation. To probe robustness beyond a single nominal starting point, we introduced two additional and experimentally relevant scenarios: (i) higher motor offset cases, where the nominal motor readout differs from the true physical position, and (ii) multi-grain diffraction conditions that introduce ambiguity in peak identification. While distinct from varying the initial orientation angle itself, these scenarios introduce qualitatively different sources of difficulty and help assess the agent's robustness under more adverse and realistic conditions. We would be happy to further explore additional systematic initial-orientation sweeps if the reviewer feels this is essential.

- Repeating the experiment multiple times under identical initial conditions would allow the authors to assess the intrinsic stochasticity due to the LLM's temperature setting (given the initial setting of the temperature to 1).

We agree with the reviewer that assessing stochasticity induced by sampling is important. In response, we have increased the number of trials from 5 to 10 times for the baseline test cases for both Claude and Gemini models.

We have also added an additional benchmark condition in which we reduced sampling stochasticity by running Gemini 2.5 Flash at temperature = 0.5, and we report both the alignment and lattice parameter *c* error distributions together with success rates (new Fig. 2(d,e), "Temp 0.5", and corresponding discussion). In these runs, the agent remains capable of completing the task in a reasonable fraction of trials (3/5 successful for alignment and 4/5 for lattice parameter *c*).

- Similarly, have the authors tested running the same trials with temperature = 0 to evaluate if the alignment and lattice parameter accuracy is maintained?

We tested zero-temperature settings and observed consistent failures. These failures were traced to edge cases at the tool-interface level. Specifically, when a motor scan exceeds its hardware limit, the motor remains at the limit position rather than returning to the starting point, a behavior not explicitly documented in the tool interface available to the agent. Recovering from such unmodeled conditions requires AI to exploratory reasoning and adaptive recovery, which benefits from stochasticity. We therefore interpret the zero-temperature failures as an informative negative result rather than an indication that deterministic settings are generally preferable.

Importantly, the observed zero-temperature failures are not due to inherent limitations of deterministic reasoning but rather arise from unmodeled tool-interface edge cases that require recovery behaviors absent from the current interface design. In future iterations, improving tool-interface guarantees would likely reduce or eliminate this dependence on stochastic exploration.

Beyond adjusting the model temperature, we tried reducing the thinking budget and have reported corresponding results under the “short thinking” label, which effectively adjusted the sequence length of the “stochastic process”. Reducing the thinking budget primarily limits the agent’s ability to explore and revise multi-step hypotheses, rather than altering sampling noise per se. We observed bimodal behavior where the experiments are either conducted smoothly or failed abruptly, due to reduced thinking depth.

Overall, the expanded benchmarks demonstrate that performance variability depends strongly on experimental difficulty and reasoning capacity, and that robustness emerges from the interaction between stochastic reasoning, structured tool use, and closed-loop feedback. We have revised the manuscript to clarify these points and to appropriately frame the quantitative conclusions.

We hope all these new results can satisfactorily address the reviewers’ questions and comments.

Regarding the terminology and naming used in the manuscript, and in line with the initial comment above, I suggest the following:

- The use of the term AI scientist appears too broad for the demonstrated application. The term "AI X-ray experimentalist" or "AI-driven X-ray experimentation agent" would be more precise and better aligned with the described functionality. This term should be used throughout the manuscript.

We thank the reviewer for this helpful suggestion. We have revised the manuscript to replace the term “AI scientist” with the more specific term “AI X-ray scientist” throughout, in order to better reflect the demonstrated scope and functionality of the system and to avoid over-broad interpretation.

We chose to retain the term “scientist” (rather than “experimentalist” or “agent”) to emphasize that the system performs more than fixed automation or scripted control. On the contrary, it conducts closed-loop reasoning, decision-making, and adaptive experimentation based on intermediate observations. We believe this revised terminology strikes a balance between conceptual clarity and appropriate scope, and

we hope it addresses the reviewer's concern; we are open to further adjustments if needed.

Regarding the choice and performance of LLM models, I have the following questions:

- The authors should clarify the rationale for selecting the three LLMs tested. Are these models comparable in terms of capabilities? And so, why comparing them? Or does one have known performance limitations, especially if used as the core of an AI agent?

- It would be valuable to discuss the expected performance of smaller language models (e.g., <20B parameters) in this context. Such an analysis could inform the feasibility of deploying local models directly at beamline facilities, avoiding data transfer to external platforms.

We thank the reviewer for raising this important point. The choice of LLMs in this work was driven by practical considerations related to tool integration, robustness, and experimental feasibility, rather than an attempt to perform an exhaustive model comparison.

During early development, we used Claude Sonnet via the Claude Desktop application because it natively supports MCP tools, which enabled rapid prototyping of structured tool use and closed-loop experimental control. Given its reliable performance in this setting, we continued to use Claude models throughout subsequent development and exploration.

To demonstrate that the proposed agentic workflow is not limited to a single proprietary model, we additionally evaluated Gemini 2.5 Flash. This model was selected due to its low cost, fast inference, and stable API-based access, making it well suited for repeated simulation-based evaluations. These models are not intended to be directly comparable in a benchmarking sense, but rather representative of current reasoning-capable LLMs that support iterative tool use and reasoning.

We would like to clarify that the primary goal of this work is not to rank or benchmark language models, but to demonstrate the feasibility of deploying an LLM-driven agent for autonomous operation of X-ray experiments, and to show that multiple contemporary LLMs can serve as the core reasoning component. Accordingly, we intentionally limited the scope of model comparisons to what is necessary to validate the agentic workflow.

Regarding smaller language models, we performed exploratory tests with a compact vision-language model (Qwen3-VL 8B). While the model was able to invoke tools syntactically, it failed to carry out the experimental task in a meaningful way, exhibiting incorrect reasoning and actions at early stages. This suggests that the key qualities such as long-horizon reasoning and reliable tool use required for X-ray experimental control may currently exceed the capabilities of smaller models without additional training or task-specific adaptation. We therefore view lightweight local models as a promising future direction for local deployment under stringent reliability and safety constraints, but beyond the scope of the present study.

We hope this discussion addresses the reviewer's questions adequately.

Regarding the time, cost, and efficiency of using the LLM agent in performing these experiments, also in strong relation with the initial major point above, I would consider adding some more quantitative details which would improve the manuscript and help evaluate practical applicability. Please, consider the following:

We thank the reviewer for this thoughtful suggestion. We agree that providing additional quantitative information can help assess the practical feasibility of the proposed approach.

- Number of LLM requests per experiment.

Addressed together with the point below.

- Number of iterations required to reach the final alignment (useful for comparison with human experimentalists).

Regarding the number of LLM requests and iterations, we have now reported the number of LLM requests in Supplementary Information section S2: Tool-use calls.

- Initial and final context sizes.

For the experiments reported, the initial prompt corresponds to approximately 33,000 input tokens (as reported by Claude Console), while the cumulative input across a complete virtual alignment experiment is on the order of 3.1M tokens.

- Whether the context window ever approached the model's maximum length.

At no point during a single alignment task did the active context approach the maximum context window of the model. For substantially longer or more complex experiments, multi-agent or context-reset strategies represent a natural future direction.

- Context size of the short vs. long prompts.

The short and long guidance prompts used in this work are approximately 3.8k and 50k characters in length, respectively.

- Total time needed to complete each experiment.

A complete virtual alignment experiment typically requires approximately 6 minutes when executed with Claude Sonnet 4. Real beamline experiments require substantially longer wall-clock time (approximately 40 minutes per alignment), primarily due to physical constraints such as mandatory waiting periods during motor scans, as well as the use of a human-in-the-loop relay for safety during early deployment. These times are inferred from experimental logs and scan timestamps rather than systematic per-call timing.

- Estimated cost of LLM inference per run.

The estimated LLM inference cost for a virtual alignment experiment using Claude Sonnet 4 is on the order of $10 per run. Real beamline experiments were conducted under a subscription-based service rather than per-call API billing, so detailed per-run billing is not available. Gemini 2.5 Flash incurs substantially lower cost (approximately an order of magnitude less), though detailed per-request billing logs were not available through the interface used. We therefore report cost estimates at an order-of-magnitude level.

- These metrics would provide a clearer picture of the scalability and efficiency of the proposed approach. In particular, they could be used to answer and discuss the major point mentioned above regarding the comparison with the human experimentalist.

We would like to clarify that the purpose of reporting these metrics is to provide practical context for scalability and deployment considerations, rather than to enable fine-grained performance or cost benchmarking across models or platforms. We hope these reasonably additions address the reviewer's questions regarding time, cost, and efficiency.

Finally some final additional points to clarify:

- In Fig. 2, the workflow proceeds through "Setup -> Primary reflection -> Secondary reflection." It is not clear whether this sequence is explicitly encoded in the guiding prompt or autonomously decided by the LLM. This should be clarified.

We thank the reviewer for pointing out this potential ambiguity. We have revised the manuscript to clarify that the workflow depicted in Fig. 2 represents a conceptual task structure rather than a hard-coded execution script. While the overall goal and available tools are defined, the LLM agent autonomously determines when to transition between the stages, which reflections to probe, and how to configure scans based on intermediate observations. We believe this clarification more accurately reflects the agent's role in decision-making and addresses the reviewer's concern.

- Regarding the LLM input data, it remains a little unclear whether the LLM receives only raw image data or also some processed tabular data version of the image. The manuscript should clarify what type of detector output is passed to the model and in what format.

We thank the reviewer for pointing out this ambiguity. We have revised the manuscript to clarify the data interfaces used by the agent. Specifically, detector images are provided to the LLM as image inputs, while scan results are returned through MCP tools as rendered plots with fitted curves and peak centers overlaid when applicable. The LLM does not directly receive raw numerical arrays of detector images. We have added a paragraph near Figure 2(c) to clarify these details. We believe this clarification more clearly describes the information available to the agent during decision-making.

- In Figure 2, the caption should explicitly explain what the points, bars, and sticks represent.

We have now clarified the meaning of these components in the Figure 2's caption.

- Regarding the real laboratory experiment, the authors mention the use of Claude Opus rather than one of the previously tested LLMs. Clarification is needed on this choice — was it for compatibility, availability, or performance reasons?

We thank the reviewer for this question. The choice of Claude Opus for the real beamline experiment was motivated by practical considerations rather than model compatibility constraints. While other tested models (e.g., Claude Sonnet) were found to perform the task in simulations, we selected a larger and more capable model for the live experiments to reduce operational risk and maximize robustness, given the limited and valuable nature of beamtime. We have clarified this rationale in the revised manuscript.

- Regarding the crystal structure data, the real experiment uses a file from the COD database, while earlier experiments used data from the Materials Project. The rationale for this change should be explained.

We thank the reviewer for this question. We have clarified in the revised manuscript that the use of a CIF file from the Crystallography Open Database (COD) in the real beamline experiment reflects standard experimental practice where experimentally curated crystal structures are often preferred over computed ones for real measurements. During early development on the simulated beamline, we used Materials Project structures for convenience and benchmarking. We have clarified this rationale in the revised manuscript and note that the agentic workflow itself is agnostic to the source of the crystallographic input.

- In the "Agentic AI models" SI section, the described parameter settings appear to refer specifically to Gemini. It should be clarified whether the same setup was used for all models. Or include the setting used for Anthropic's models.

We thank the reviewer for pointing out this potential ambiguity. We have now clarified in the revised Supplementary Information that the explicit parameter settings described apply to the Gemini models, which expose configurable inference parameters. Anthropic's Claude models were accessed via the Claude Desktop interface, which does not provide user-accessible controls for comparable parameters; therefore, these models were used with their default settings.

- I'd suggest calling Claude Sonnet and Opus and Gemini Flash simply LLMs.

We thank the reviewer for this suggestion. We have revised the manuscript to consistently refer to Claude Sonnet, Claude Opus, and Gemini Flash collectively as large language models (LLMs), while retaining specific model names only where implementation details are discussed.

- Regarding the command extraction from the LLM output, the current keyword-based parsing of LLM outputs could be prone to errors. The authors might acknowledge this limitation and suggest the use of structured outputs (e.g., via JSON or pydantic models) as a potential improvement.

We thank the reviewer for this helpful suggestion. In the current system, command extraction is implemented via a keyword-based parsing approach, chosen to closely mirror existing beamline control workflows. We agree that this strategy can be prone to parsing ambiguities and have revised the manuscript to explicitly acknowledge this limitation. We also note that adopting structured output formats (e.g., JSON schemas or typed models such as pydantic) is a promising direction for improving robustness and is an important avenue for future work.

- Regarding the real experiments, were them executed only once, or were multiple times performed to assess reproducibility?

We thank the reviewer for this important question. During the real beamline session, we explored multiple experimental scenarios with progressively increasing difficulty, including variations in initial manual offsets and the availability of crystallographic information. Due to the limited and valuable nature of synchrotron beamtime and the large exploration space (two materials with different testing configurations), systematic repetition of identical real-experiment configurations was not performed. The cases reported in the manuscript correspond to the most challenging configurations explored during the session, including scenarios with missing in-plane orientation information and manual motor offsets. We emphasize that the real experiment is intended as a proof-of-feasibility demonstration of autonomous closed-loop operation under realistic experimental conditions, while statistical robustness and variability are evaluated through repeated trials in the virtual beamline benchmarks. We have clarified this point in the revised manuscript.

Thanks!

Francesco Ricci

We thank the reviewer for their careful reading and thoughtful feedback on our manuscript. The comments have been highly constructive, and we feel that the resulting revisions have substantially improved both the technical depth and presentation of the

work. We hope that the revised manuscript addresses the reviewer's concerns satisfactorily and greatly appreciate their time and consideration.

Reviewer #4 - Remarks on Code Availability

The code is not available. The Authors state: All data and source code necessary to reproduce the reported results will be made publicly available upon acceptance of this work.

We thank the reviewer for raising this point. We would like to clarify that the source code was made available during the review process. The code was initially uploaded to CodeOcean but was later withdrawn due to technical issues, after which it was re-shared via a GitHub repository (https://github.com/zhantaochen/llm4xray) and was available during the peer review. We apologize for any confusion this may have caused.

Attachment - Conversation History of Qwen3 VL 8B Model

Please find below the conversation history regarding Qwen3 VL 8B Model conducting a virtual experiment.

Conversation

- conversationId: 1006f9e7-671f-4182-bee3-b48a91c6ed46
- endpoint: agents
- title: Qwen3-vl-8b-run1
- exportAt: 13:15:23 GMT-0600 (Central Standard Time)

Options

- presetId: null
- agent_id: agent_cstAMj-O0HmqTSR0OEL5G
- endpoint: agents
- endpointType: undefined
- title: Qwen3-vl-8b-run1

History

User

Could you help align my sample? You may use the .env file located at: 'C:\Users\zt94c\Dropbox\SLAC\research\RL_beamline\RL_Beamline_Toy_Example\working_dir\2025_NMI-Revision\Claude-4-sonnet\current\env' for accessing necessary information. When creating a new experiment, use the config_path="C:\Users\zt94c\Dropbox\SLAC\research\RL_beamline\RL_Beamline_Toy_Example\working_dir\2025_NMI-Revision\Claude-4-sonnet\current\config.yaml" as the argument to load correct configuration file.

The experimental operations will completely follow your intention, you may use the execute_command MCP tool to conduct the experiment.

Please pay additional attention to follow instructions, especially for finding the 2nd reference reflection. ⚠️ IMPORTANT: You should remember to do fine scans when the scanned peak is poorly sampled and overly sharp.⚠️

Be precise and use enough digits (at least 4 digits) when you want to use umv to move motors, as the peaks can be extremely sharp.

You will work with a vertical reflection setup. Make sure you optimize lattice constants when appropriate; for cubic systems, avoid optimizing the lattice constants upon finding the second reference reflection. Be very careful when setting the lattice constants, as they are critical for the success of the experiment. For example, it is very unlikely that the lattice constants will be changed by 20%. A typical change in lattice constants is around 0.1% to 0.5%. If you are not sure about the lattice constants, you should not apply the change and should use the original values for safety.

Before you setlat or or0/or1, always check the detector image first to make sure the peak is well optimized and centered.

⚠️ CRITICAL: If you only capture the tail of motor scan instead of the entire peak, you MUST extend your scan range to capture the peak accurately. ⚠️

⚠️ CRITICAL: If you noticed an offset of any motor, you MUST account for this offset value when you use theoretically calculated motor angles when searching for reference reflections. ⚠️

The in-plane orientation of the sample is completely unknown, so you should expect to perform a wider phi scan in the beginning to capture the non-specular peak based on the crystal symmetry. During your initial phi scan, you typically only need to cover the necessary range given the symmetry (e.g., AT LEAST 90 degrees for 4-fold in-plane symmetry and AT LEAST 120 degrees for 3-fold in-plane symmetry). You should use pd5 during the initial scan.

⚠️ CRITICAL: You must ensure that the ENTIRE peak is well captured during the phi scan -- including tails at both ends -- this directly impacts the success of the experiment. ⚠️

There might be few degrees offsets in motors, please enlarge the initial scan range accordingly to avoid missing any peaks -- if you do not observe obvious peak profiles, this is a strong indication that you are not capturing a peak, perform a wider range scan instead of keep refining the current scan; and be careful not to recognize hot pixels as peaks. If you do not see a peak at initial search, try starting with increased scan range.

AI X-ray Scientist qwen 3 vl

AI X-ray Scientist qwen 3 vl

{"type":"tool_call","tool_call":{"id":"0e8ec7f8-ae93-4bba-a3ec-5015289a4c9a","name":"optimize_lattice_parameter_mcp_xray_agent","args":"{"H":1,"

K":0,"L":0,"a":5.67,"alpha":90,"b":5.67,"beta":90,"c":5.67,"gamma":90,"target_param":"a","two_theta_measured":10.37,"wavelength":1.54}","type":"tool_call","progress":1,"output":"Lattice Parameter Optimization Results (Session: 7d1021b5-e7b2-465a-8b52-0101150af303):\n\nOriginal Parameters:\n a = 5.670000 Å, b = 5.670000 Å, c = 5.670000 Å\n α = 90.000°, β = 90.000°, γ = 90.000°\n\nOptimized Parameters:\n a = 6.803995 Å, b = 5.670000 Å, c = 5.670000 Å\n α = 90.000°, β = 90.000°, γ = 90.000°\n\nOptimization Details:\n Target parameter: a\n Original value: 5.670000\n Optimized value: 6.803995\n Change: 1.133995 (20.00%)\n \n Peak: (1.0, 0.0, 0.0)\n Measured 2θ: 10.370°\n d-spacing observed: 8.520352 Å\n d-spacing calculated: 6.803995 Å\n d-spacing error: 1.71635673 Å\n \n Optimization successful: True"}}

AI X-ray Scientist qwen 3 vl

The time required for a 50% probability of observing two photons within a time window of length T is derived from the Poisson process context. Here's the key insight:

- The time between the first and second photon follows an **exponential distribution** with rate λ.
- The probability that the second photon arrives within a time window of length T (assuming the first photon is at time 0) is given by:

$$P(\text{second photon within } T) = 1 - e^{-\lambda T} - \lambda T e^{-\lambda T}$$

- To find the **time** t such that this probability equals 50%, solve:

$$1 - e^{-\lambda t}(1 + \lambda t) = 0.5$$

  Let $x = \lambda t$. This simplifies to:

$$e^{-x}(1 + x) = 0.5$$

- Solving numerically gives $x \approx 1.648$, so:

$$t = \frac{1.648}{\lambda}$$

Interpretation of Input and Output

- **Input**: T is the **time window** (length of the interval).
- **Output**: The **time** t required for a 50% probability of both photons arriving within the window T.

However, the problem states: "*the function should take* T *as input and return the time value*".

This suggests that T is **not** the time window, but rather **the rate parameter** λ. In many contexts, especially in physics or engineering, T is used to denote the rate (e.g., $\lambda = T$).

Final Answer

Assuming T represents the rate λ, the time t for a 50% chance of observing two photons within a window is:

$$\boxed{\frac{1.648}{T}}$$

This value corresponds to the **time until the second photon is observed with 50% probability**, assuming the first photon is at time 0. If T is instead the time window, the problem becomes underspecified without knowledge of λ. Based on the phrasing, T **is assumed to be the rate**, leading to the above solution.

Responses to Reviewers' Comments

Manuscript ID: NATMACHINTELL-A25083858A-Z

Manuscript Title: An Agentic Artificially Intelligent X-ray Scientist

We would like to thank all the reviewers for their time and effort in providing valuable comments and questions, which have greatly helped us to further improve the manuscript. Below, please find our point-by-point responses highlighted in blue color. All changes made to the revised manuscript (a separate PDF file) are marked in green color.

Table of Contents

Reviewer #1 (Remarks to the Author)

All comments are addressed, additional tests are provided, generalization to other type of tasks are discussed. More x-ray and large facilities related literature with AI and automation can still be added.

We sincerely thank the reviewer for their positive assessment and helpful suggestion. In response, we have added several additional references in the Introduction on AI- and automation-enabled experimentation in X-ray and other large-scale scientific facilities.

We are grateful for the reviewer's constructive feedback throughout the review process, which has helped us improve the manuscript substantially.

Reviewer #3 (Remarks to the Author)

The authors have made substantial revisions that address several of my previous concerns. The manuscript now more clearly positions the work relative to prior literature and clarifies the scope/novelty of their contribution.

We are grateful for the reviewer's insightful feedback throughout the review process, which has helped us more clearly define and communicate the contribution of this work. We thank the reviewer for recognizing our efforts to better position our work, and we are pleased that the revisions helped clarify its scope and its relationship to the prior literature.

I find the paper technically sound, and the topic an important one. My remaining reservation relates primarily to the level of conceptual advance. The approach largely integrates existing large language models with experimental control tools and a simulated environment, rather than introducing fundamentally new machine learning methods. While the integration is thoughtful and the real-world beamline demonstration is compelling, it is less clear whether the work represents a methodological advance of the type typically emphasized in Nature Machine Intelligence.

We agree that the contribution of this work is not the introduction of a fundamentally new machine learning architecture.

We believe the methodological advance lies in establishing and validating a practical framework for deploying AI agents in real large-scale scientific facilities. Specifically, this work brings an AI agent into a real production beamline and the full beamtime workflow, and demonstrates closed-loop operation under realistic experimental constraints. The combination of a virtual beamline environment, structured guidance, seamless transition from the virtual to the real beamline, and successful real-world deployment makes this more than an isolated demonstration: it provides a concrete blueprint for future agentic AI systems at scattering facilities and other large-scale scientific instruments. We therefore view this work as a meaningful advance in machine intelligence for real-world science. We thank the reviewer for their thoughtful assessment and for recognizing the importance of the problem.

That said, the work provides an interesting proof-of-concept for agentic AI in experimental science, and the revisions have improved the manuscript overall. Whether this contribution meets the bar for sufficiently novelty to be published in this journal is ultimately a decision for the editor. I have no other major concerns.

Once again, we appreciate the reviewer's time and careful review. We believe this work helps define an important new direction for machine intelligence in science by showing how artificial intelligence can be integrated with large-scale scientific machines to

address real experimental problems under operational constraints. In this respect, we view the contribution as both distinctive and strongly aligned with the interests of *Nature Machine Intelligence*.

Reviewer #3 (Remarks on code availability)

The authors have now made available their code on Github, which appears nicely documented.

We thank the reviewer for taking the time to examine the shared code repository and for noting that it is well documented.

Reviewer #4 (Remarks to the Author)

The authors have thoroughly revised the manuscript, addressing all of my concerns and incorporating the majority of my suggestions. They provided satisfactory responses to my previous queries, and the current version significantly clarifies the study's scope and objectives.

Notably, the addition of new experiments and edge cases has strengthened the statistical analysis and error reporting. Furthermore, the expanded introduction and updated literature review better position the work within the field.

In light of these improvements, I believe the manuscript is now acceptable for publication.

We are truly pleased to learn that the reviewer found our revision satisfactory and that it now better defines the scope and objectives of the work. We also greatly appreciate the reviewer's positive recognition of the added experiments, strengthened statistical analysis, and improved error reporting.

Reviewer #4 (Remarks on code availability)

The associated repository is well-documented, featuring a comprehensive README that provides a clear, step-by-step installation and setup guide for the package, the MCP server, and the LLM provider. The authors have also included an IPython notebook to facilitate testing of the core functionalities. Furthermore, a graphical user interface to executed a simulation is provided to the user. Most importantly, the inclusion of the "long" prompt (mentioned in the manuscript), initial crystal structures, and configuration files ensures the study's results are fully reproducible.

We thank the reviewer for their thoughtful evaluation of the code and documentation. We are glad that the released repository and accompanying materials support the transparency and reproducibility of this work.