Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An agentic AI neutron scattering data scientist at the China Spallation Neutron Source

Rong Du

durong@ihep.ac.cn

Institute of High Energy Physics, Chinese Academy of Sciences

Dongbo Xiong

Institute of High Energy Physics, Chinese Academy of Sciences

LI Qingmeng

Institute of High Energy Physics, Chinese Academy of Sciences

Jiajun Zhong

Institute of High Energy Physics, Chinese Academy of Sciences

Liujun Han

Institute of High Energy Physics, Chinese Academy of Sciences

Jvping Xu

Institute of High Energy Physics, Chinese Academy of Sciences

Huaican Chen

Institute of High Energy Physics, Chinese Academy of Sciences

Jingyu Li

Dongguan University of Technology

Peng-Fei Liu

China Spallation Neutron Source https://orcid.org/0000-0002-9170-5238

Yiqing Sun

Institute of High Energy Physics, Chinese Academy of Sciences

Fazhi Qi

Institute of High Energy Physics, Chinese Academy of Sciences

Zhengde Zhang

Institute of High Energy Physics, Chinese Academy of Sciences

Junrong Zhang

Institute of High Energy Physics, Chinese Academy of Sciences

Hao Wang

Institute of High Energy Physics, Chinese Academy of Sciences

---

Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11103192/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

An agentic AI neutron scattering data scientist at the China Spallation Neutron Source

Hao Wang [a, b], Dongbo Xiong [a], Qingmeng Li [a], Jiajun Zhong [a], Liujun Han [a], Jvping Xu [a, b], Huaican Chen [a, b], Jingyu Li [d], Pengfei Liu [a,b], Yiqing Sun [a], Fazhi Qi [a], Zhengde Zhang [a, *], Junrong Zhang [a, b, c, *], Rong Du [a, b, c, *]

[a] Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China

[b] Spallation Neutron Source Science Center, Dongguan 523803, China

[c] University of Chinese Academy of Sciences, Beijing 100049, China

[d] Dongguan University of Technology, Dongguan 523000, China

* Corresponding author. Email address: durong@ihep.ac.cn (Rong Du), jrzhang@ihep.ac.cn (Junrong Zhang), zdzhang@ihep.ac.cn(Zhengde Zhang)

Abstract:

The interpretation of neutron scattering data remains a major bottleneck in materials discovery, as it demands an iterative cycle of data reduction, modeling, and refinement. That is too time-consuming for experts and too complex for novices. Here we introduce DrSai-Rongzai, an autonomous agent built on a large language model (LLM) that perfoms joint neutron powder diffraction (NPD) and pair-distribution-function (PDF) analysis without human intervention. Deployed at the China Spallation Neutron Source (CSNS), the agent orchestrates established refinement software through a hierarchical "Reasoning + Acting" framework. Its core innovation, Agentic Context Operation (ACO), maintains a coherent 'working memory' of file paths, intermediate results, and domain-specific rules across entire processing steps, even when the underlying large language model undergoes context compression. Across reference materials (Si, CsAgTe and Bi2MoO6), the agent refined structural parameters statistically indistinguishable from expert-derived solutions, while requiring markedly fewer iterative cycles. In an exhaustive 18-run factorial ablation study on two high-entropy alloys, we systematically evaluated the MemoWrite and CoreSkillWrite in ACO mechanisms with and without 50k-token context compression, quantifying the necessity of each component. Memo improved both mean composite performance and repeatability even without compression (+20.2% and +11.3%, respectively). Under compression, Memo improved these metrics by +10.7% and +17.0%, and retaining CoreSkill provided a further +12.5% and +6.6% improvement, yielding a mean composite score of 0.9316 and repeatability stability of 0.9905 for full ACO. The agent is integrated with a permission-controlled CSNS data interface, enabling secure, on-demand analysis of user-proprietary data with automatic archiving. By combining rigorous benchmarking with practical, facility-ready deployment, DrSai-Rongzai demonstrates that autonomous AI agents can move beyond assisted recommendations to become reliable partners in complex structural science, potentially accelerating the rate of discovery at pulsed neutron sources worldwide.

Keywords: DrSai-Rongzai; Agentic AI; Neutron scattering; Rietveld refinement; Pair distribution function; Context management; CSNS

Introduction

Large-scale scientific facilities, including synchrotron radiation sources, spallation neutron sources, and free-electron lasers, have become indispensable platforms for probing matter across atomic to mesoscopic scales. These facilities operate numerous beamlines serving thousands of users annually, driving discoveries in materials science, condensed matter physics, chemistry, and biology [1–4]. In recent years, high-throughput and in situ experiments generate data at unprecedented rates. A single experiment can now produce terabytes of data, far exceeding the capacity of manual analysis. However, the interpretation of such data remains a deeply human-centric endeavor. It demands iterative reasoning, domain-specific judgment, and the orchestration of heterogeneous software tools—skills acquired only through years of hands-on experience. This bottleneck between data generation and data interpretation is now widely recognized as the primary limitation on scientific productivity at these facilities, yet few solutions exist to bridge this gap.

Recent advances in AI have introduced large language model (LLM) agents capable of autonomously executing complex professional workflows, opening a promising path toward full automation at large-scale facilities. On the instrumental side, multiple agentic systems have been demonstrated in instrument operation [5,6], even multi-stage physics experiments [7]. For example, Chen et al. developed an LLM-based agent that autonomously performs X-ray sample alignment on a synchrotron beamline [5]. On the data analysis side, specialized agents have emerged for tasks such as small-angle scattering (SasAgent) [8], single-crystal neutron crystallography (NeuDiff Agent) [9], and powder diffraction Rietveld refinement [10]. However, these systems share a fundamental limitation: they are single-task, single-modality by design. A complete analysis of target sample often needs complementary datasets. Diffraction patterns reveal average structure, pair distribution function probes local structure, and small-angle scattering captures mesoscale morphology. Together, they tell a complete structural story. The joint interpretation of these multimodal data is not merely a convenience but a scientific necessity for understanding complex materials. Achieving this requires an agent that can orchestrate multiple analytical techniques, each itself a complex workflow, within a unified decision-making framework. As shown in Fig. 1(a), analysis of a single modality already constitutes a “Reasoning + Acting” (ReAct) workflow of

observation, planning, and action [11]. Joint analysis of multiple modalities demands an additional outer loop that orchestrates these individual analyses, resulting in a hierarchical ReAct structure shown in Fig. 1(b). Beyond this structural complexity lies a second, equally critical challenge: the long-horizon context management problem. Each inner loop involves iterative model fitting, an ill-posed inverse problem that often requires 20+ iterations of parameter adjustment, constraint application, and convergence diagnostics. These iterations rapidly inflate the agent's context window with tool outputs and decision traces [10,12–18]. When automatic context compression is triggered, the agent risks losing two types of information that are essential for task completion: the expert knowledge embedded in refinement skills that guides each decision, and the file paths to structured data and parameter files that tools require for execution. The loss of either can be catastrophic, eroding decision quality or causing complete workflow failure. While context management is a recognized challenge in long-running agentic tasks [19,20], it becomes particularly acute in scientific data analysis, where iterative refinement is not the exception but the norm. Existing systems, designed for single-pass or short-horizon tasks, are fundamentally unequipped to handle these dual challenges of hierarchical orchestration and sustained context fidelity.

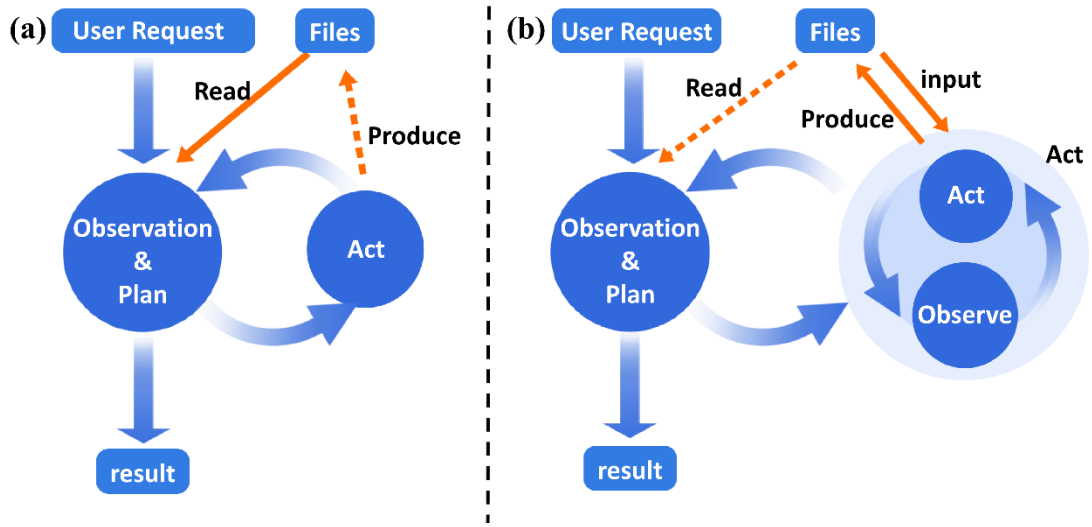


Figure 1. (a) the general "ReAct" workflow of LLM-based agent. (b) the Hierarchical ReAct workflow in large-scale facility data analysis. The real arrow line means absolutely necessary step. The dashed arrow means the step might exist or not.

To overcome these dual limitations, the structural challenge of hierarchical multimodal orchestration and the context challenge of long-running iterative

refinement, we introduce DrSai-Rongzai, an autonomous LLM-based agent for neutron scattering data analysis at the China Spallation Neutron Source (CSNS). This work extends our previous Rongzai agent, which achieved expert-level Rietveld refinement for neutron powder diffraction (NPD) [10], in two significant dimensions. First, we demonstrate for the first time that LLM-based autonomous pair distribution function (PDF) fitting can achieve expert-level quality. This technique is fundamentally different from NPD refinement. Our agent produces results that matches human-refined patterns, R_w values, and parameter consistency. Second, we go beyond single-technique automation. We integrate NPD refinement and PDF fitting into a unified framework, enabling joint multimodal analysis through a hierarchical ReAct workflow. The key enabler for robust execution is a novel Agentic Context Operation (ACO) mechanism. It allows the agent to autonomously store long-term expert knowledge and critical information, including file paths, in a dedicated context block that is preserved throughout long tasks and exempted from compression. The agent can delete such context when it is no longer needed, saving tokens and reducing cognitive load. Drawing on the Agent Skills concept from Anthropic [21], we refactor the refinement capabilities of the previous Rongzai agent [10] into skill-description and MCP-tool-based modules with a standardized interface and timeout-aware automatic polling, facilitating the addition of new analytical capabilities. We benchmark the agent against expert-level manual fits for standard materials (Si, CsAgTe, Bi_2MoO_6) and demonstrate successful joint analysis on high-entropy alloys ($AgMnPbSbTe_4$, $AgMnGePbSbTe_5$), where the ACO mechanism proves essential for stable performance. Finally, we connect the agent to the CSNS Data Organization Management Access Software (DOMAS) database [22] through a permission-gated interface, providing secure, on-demand access to user data and making DrSai-Rongzai a production-grade analytical workhorse rather than a research prototype. We believe this work validates the effectiveness of LLM agent technology for data analysis at large-scale facilities, represents a key step toward fully intelligent user facilities, and holds substantial value for future developments in intelligent facilities.

Results

Architecture of DrSai-Rongzai

Fig. 2 shows the interactions among three parties: the user, the beamline

scientist, and the agent. The user provides samples to the facility; the beamline scientist conducts experiments; and the resulting data are stored in the database. Upon user authorization, the agent can retrieve and analyze the user's data, returning results to both the user and the database as part of the scientific record.

The agent’s harness consists of three functional layers. (1) The first layer is the system prompt, which includes the agent’s role definition, general constraints, skill descriptions, base-tool schemas, and dedicated context block for ACO. (2) The second layer contains the full skill content and base tools. Each skill embeds the expert knowledge required for a specific data-analysis task, along with the Application Programming Interface (API) specifications of the corresponding Model Context Protocol (MCP) tools. Only after reading the skill can the agent acquire the correct calling methods for the refinement-related MCP tools. Among the base tools, two are most important. The permission-checking tool verifies the current user's access rights to specified data. The database content is hard-coded to be inaccessible to the agent, and only after invoking this tool and obtaining permissions can the agent access the authorized data. The ACO tools comprise three components: TodoWrite (for recording the plan), memoWrite (for recording key information and file paths), and coreSkillWrite (for promoting or demoting skills). Among these, TodoWrite is a general-purpose task-tracking tool, while memoWrite and coreSkillWrite are novel components specifically designed to address the unique challenges of neutron scattering data analysis. These challenges include the preservation of expert knowledge and critical file paths under long-horizon, iterative refinement loops, where automatic context compression would otherwise cause irreversible information loss. Together, these three components constitute the ACO mechanism, which is a core methodological contribution of this work. (3) The third layer comprises the MCP tools, exposed via MCP servers. Their APIs and internal implementations are not directly exposed to the model; instead, they are described in the skill layer. The current MCP tools include multiple toolkits developed by our team in collaboration with the beamline scientist at the Multi-Physics Instrument (MPI) of CSNS. These

toolkits are based on commonly used open-source neutron scattering software, including a NPD refinement toolkit based on GSAS-II and a neutron total-scattering PDF fitting toolkit based on PDFgui [23,24].

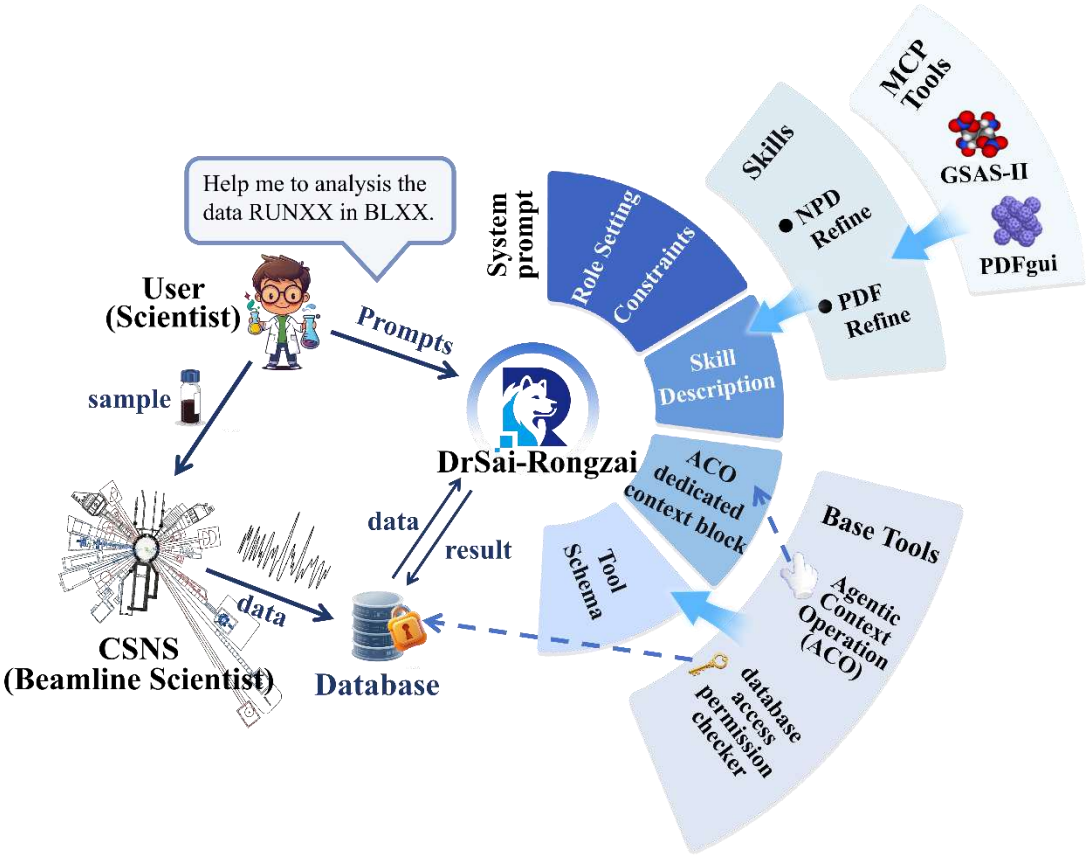
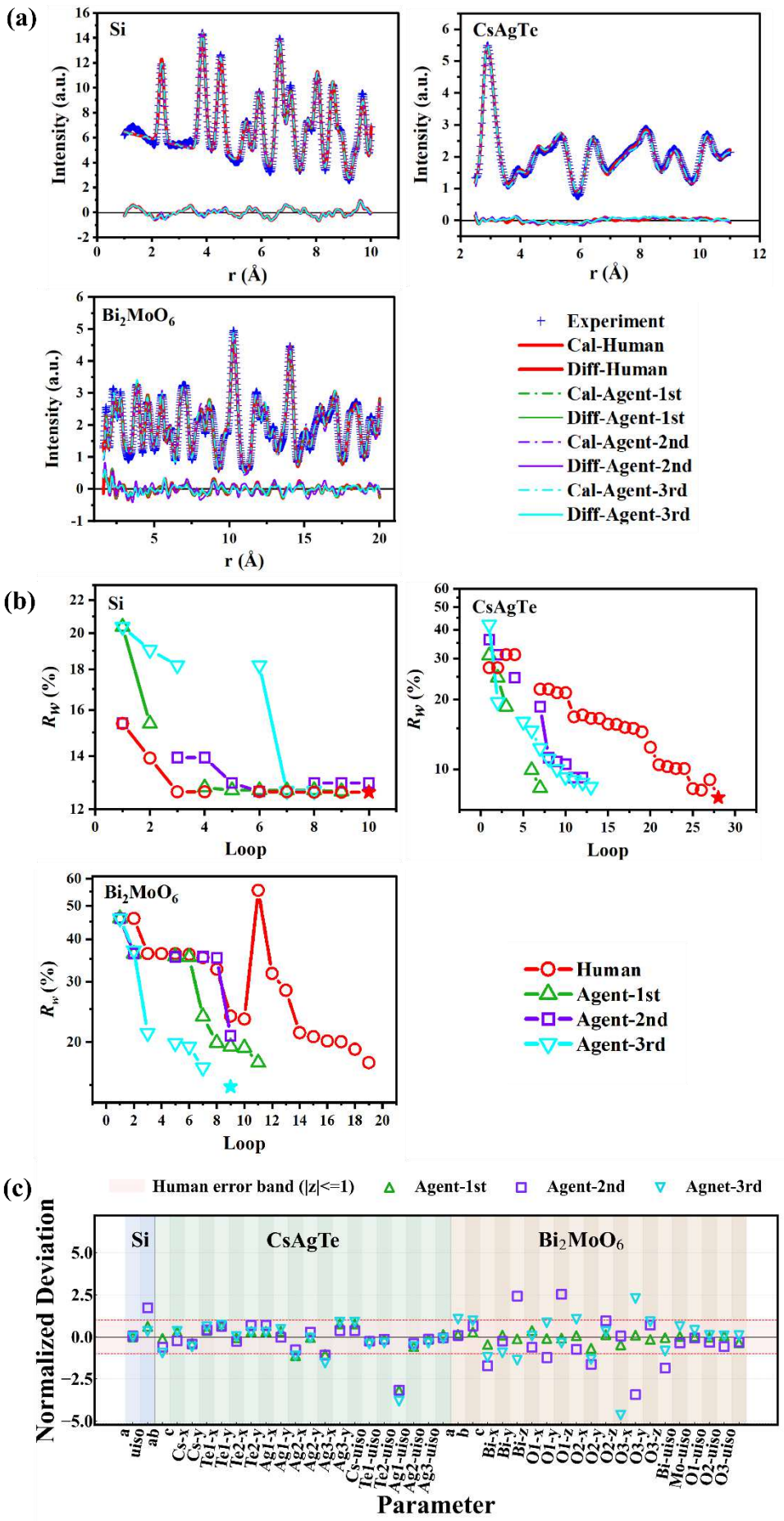


Figure 2. System architecture of the DrSai-Rongzai agent and its role within the CSNS data workflow, highlighting the three-layer harness and the permission-controlled database connectivity.

Expert-level Performance in Neutron PDF Fitting

To evaluate whether the agent can reach expert-grade analytical quality, we benchmarked autonomous PDF fitting against reference fits produced by experienced human analysts. The benchmark dataset comprised three standard samples: Si (a reference material), CsAgTe (a thermoelectric material) [25], and Bi_2MoO_6 (a photocatalytic material) [26] . All were measured at the MPI of CSNS. The PDF fitting toolkit was developed, validated, and integrated into the agent's MCP toolset primarily by the MPI beamline scientist, ensuring that the toolchain reflects real-world experimental conditions and domain expertise. The agent performed three independent fits for each sample, and the results were evaluated across four

complementary metrics: (1) visual agreement of the fitted patterns (qualitative comparison against human fits), (2) the weighted-profile R factor (R_w), (3) the number of refinement iterations required to reach convergence, and (4) the consistency of refined structural parameters relative to the human-reference values (quantified by the fraction of parameters within one standard deviation of the reference). The results demonstrate expert-level performance across all four metrics. First, as shown in Fig. 3(a), all agent-produced patterns are nearly indistinguishable from the human benchmark fits, indicating that the agent captures the essential structural features. Second, Fig. 3(b) demonstrates that the agent achieved R_w values comparable to the human reference across all three samples, and notably obtained a better R_w for Bi_2MoO_6, suggesting that the agent can, in some cases, surpass human performance. Third, the agent required substantially fewer refinement loops than the manual reference in nearly every case (shown in Fig.3(b)), demonstrating superior efficiency without sacrificing quality. Fourth, approximately 84% of the agent-refined parameters fell within one standard deviation of the human-reference values (shown in Fig.3(c)). It is important to recognize that PDF fitting, like all ill-posed inverse problems, does not admit a unique ground-truth solution. Different refinement strategies can yield equally valid results. Therefore, the 84% agreement rate, combined with the visual and R_w evidence, confirms that the agent-derived parameters are physically reasonable and consistent with expert practice. Collectively, these findings demonstrate that tool-grounded autonomous agent can match or, in some respects, exceed the performance of expert manual workflows in routine PDF fitting scenarios, establishing a solid foundation for more complex multimodal analysis tasks.


(a)
Si
CsAgTe
Bi2MoO6
Intensity (a.u.)
r (Å)
Experiment
Cal-Human
Diff-Human
Cal-Agent-1st
Diff-Agent-1st
Cal-Agent-2nd
Diff-Agent-2nd
Cal-Agent-3rd
Diff-Agent-3rd
(b)
Si
CsAgTe
Bi2MoO6
Rw (%)
Loop
Human
Agent-1st
Agent-2nd
Agent-3rd
(c)
Human error band (|z|<=1)
Agent-1st
Agent-2nd
Agnet-3rd
Si
CsAgTe
Bi2MoO6
Normalized Deviation
Parameter

1

Figure 3. Expert-level performance of the LLM agent in neutron PDF fitting compared with human experts. (a) Visual comparison of the fitted PDF patterns for Si, CsAgTe, and Bi_2MoO_6. (b) Comparison of the R_w decay trends and the number of refinement iterations required to reach convergence. (c) Comparison of normalized deviations of key refined parameters relative to human reference values. The red shaded area in (c) represents the human error band.

ACO Ablation Study on Multimodal Neutron Scattering Data

To evaluate the effectiveness of the proposed ACO mechanism in multimodal neutron scattering data analysis tasks, we conducted a factorial ablation study using real neutron diffraction and PDF data from two high-entropy alloy samples, $AgMnPbSbTe_4$ and $AgMnGePbSbTe_5$ [27]. The task required the agent to independently fit four neutron datasets (two NPD and two PDF), evaluate the fitting results, and draw analytical conclusions. We designed a composite scoring formula that integrates R_w and conclusion correctness metrics to score each attempt. We evaluated the three ACO configurations (Memo + CoreSkill, Memo only, and neither) both without context compression and under a 50k-token context-compression setting. Each of the six resulting conditions was repeated three times.

Before comparing the ablation conditions, we established the evaluation framework for the agent-generated conclusions. Table I lists the complete conclusion vocabulary observed across all runs and its human annotation. The first four conclusions are findings explicitly reported for these high-entropy alloys in the original study [22] and form the primary literature-grounded reference for the hit-rate calculation. The additional TRUE-labelled conclusion was independently supported by the human assessment of the same dataset. The remaining statements were designated as FALSE or Uncertain when the available evidence did not support a definitive positive conclusion. Thus, the TRUE-labelled statements collectively define the target set used to calculate hit rate, whereas the first four provide its directly literature-grounded core.

Table I. Complete set of conclusions produced by the agent in the ablation study and their human annotations. The first four TRUE-labelled conclusions are explicitly reported in the original high-entropy-alloy study [22] and constitute the primary literature-grounded reference for the hit-rate metric. "Uncertain" denotes insufficient evidence to determine correctness.

Conclusion	Ground Truth	Basis for hit-rate evaluation
$AgMnPbSbTe_4$ belongs to Fm-3m	TRUE	Original-study result
$AgMnGePbSbTe_5$ belongs to Fm-3m	TRUE	
$AgMnPbSbTe_4$ has a local structural disorder	TRUE	
$AgMnGePbSbTe_5$ has a local structural disorder	TRUE	
Ge doping induces the shrinkage of cell	TRUE	Human-supported result from the evaluated dataset
The doping of Ge induces stronger local structural disorder	Uncertain	Insufficient supporting evidence
Ge doping reduces the disorder of Te, making the cell more stable	Uncertain	
Ge doping reduces the disorder of Pb	Uncertain	
Ge doping improves the disorder of Ag	Uncertain	
The doping of Ge reduces the local structural disorder	Uncertain	
Ag is almost expelled from the main phase after Ge doping	Uncertain	
Ge doping improves the occupies of Sb	Uncertain	

Fig. 4 compares the two operating regimes directly. Panels (a–c) show the three no-compression conditions, whereas panels (d–f) show their counterparts under context compression. In the no-compression regime, enabling memoWrite improves the composite scores substantially relative to the baseline without ACO, demonstrating that maintaining an explicit working record remains beneficial even when automatic compression is not invoked. Adding coreSkillWrite produces no further measurable gain in this regime. This is consistent with the necessary refinement knowledge remaining available in the active context window when compression is absent. Under compression, however, the distinction between the three configurations becomes pronounced. The full-ACO condition (brown: Compression & Memo & CoreSkill) yields the three highest composite scores, followed by the memo-only condition (purple: Compression & Memo & No CoreSkill), while the condition without either mechanism (red: Compression & No Memo & No CoreSkill)

performs the worst. Retaining the comprehensive skill content incurs an initial token overhead, but it preserves the expert guidance required for reliable refinement decisions after compression is triggered. Furthermore, removing memoWrite forces the agent to recover lost file paths, increasing token expenditure and, in one run, leading to incomplete PDF refinements and missing R_w values in panel (f).

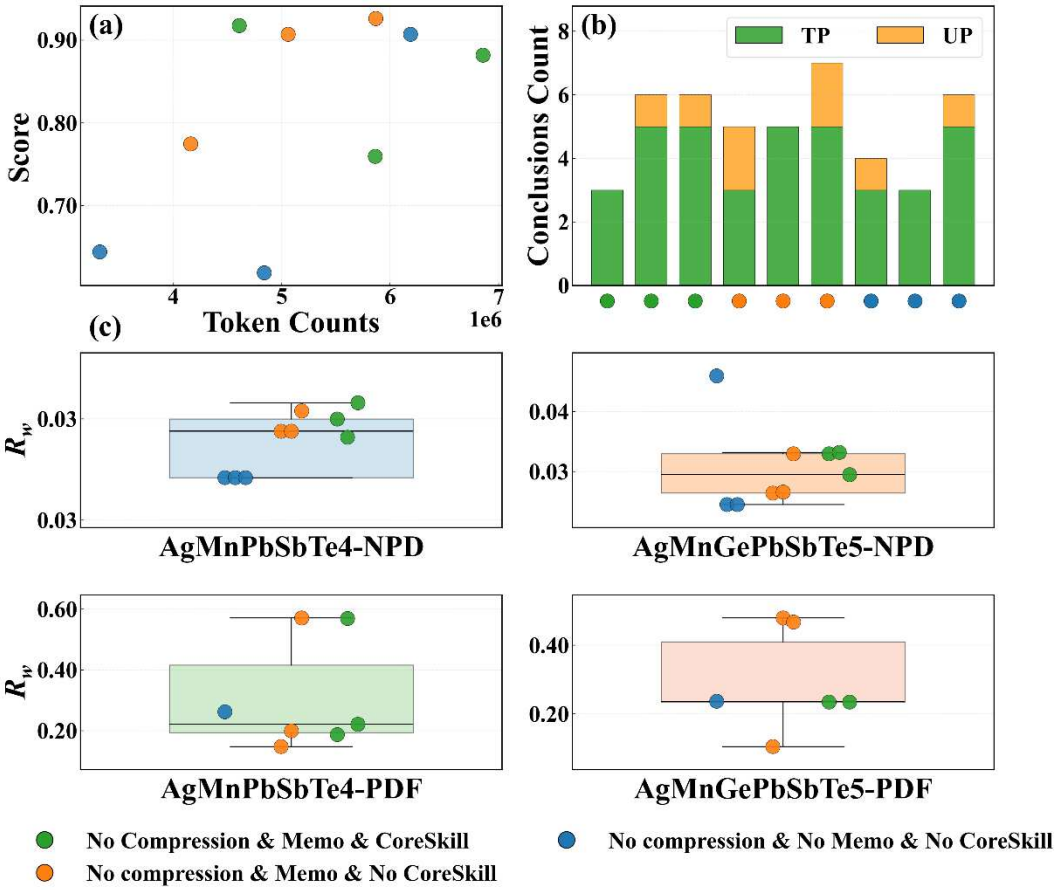
To quantify condition-level performance and reproducibility, Table II reports the mean composite score and repeatability stability across the three independent runs. The no-compression comparison isolates the contribution of the two ACO components when the context window is intact. Relative to the no-ACO baseline, memoWrite improves the mean composite score from 0.7234 to 0.8694 (+20.2%) and repeatability stability from 0.8079 to 0.8990 (+11.3%). Adding coreSkillWrite in this no-compression setting produces no additional gain: the mean score changes slightly from 0.8694 to 0.8531 (−1.9%), while stability changes from 0.8990 to 0.8947 (−0.5%). This absence of benefit is consistent with the refinement knowledge remaining in the active context window without compression.

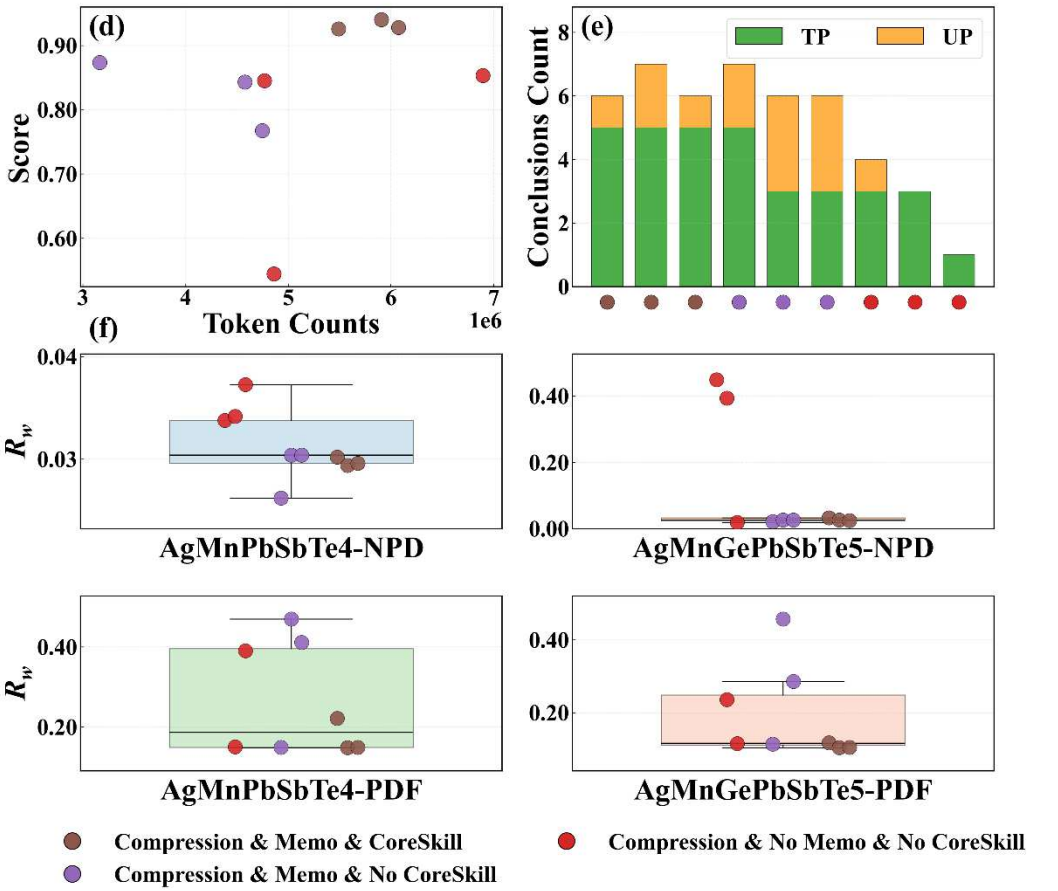
Under compression, both ACO components provide distinct and complementary improvements. Relative to the compressed no-ACO baseline, memoWrite improves the mean composite score from 0.7479 to 0.8282 (+10.7%) and repeatability stability from 0.7942 to 0.9294 (+17.0%). Retaining CoreSkill in addition to Memo then improves the mean score from 0.8282 to 0.9316 (+12.5%) and repeatability stability from 0.9294 to 0.9905 (+6.6%). Overall, full ACO improves the mean composite score by 24.6% and repeatability stability by 24.7% relative to the no-ACO baseline. The repeatability stability metric provides information beyond the mean score alone: it shows that Memo reduces run-to-run variation even without compression, whereas preserving CoreSkill is crucial for achieving near-identical outcomes across repeated long-horizon workflows after compression. Together, these effects show that Memo provides a general benefit by preserving task-specific operational state, while CoreSkill becomes decisive when compression would otherwise remove the domain knowledge needed for reliable refinement decisions.

Table II. Condition-level composite performance and repeatability across three independent runs. Larger values are better for both metrics.

Context setting	Memo	CoreSkill	Repeats	Mean composite score	Repeatability stability
No compression	No	No	3	0.7234	0.8079
No compression	Yes	No	3	0.8694	0.8990
No compression	Yes	Yes	3	0.8531	0.8947
Compression	No	No	3	0.7479	0.7942
Compression	Yes	No	3	0.8282	0.9294
Compression	Yes	Yes	3	**0.9316**	**0.9905**

The conclusion breakdowns in panels (b) and (e) are evaluated against the ground-truth labels in Table I. No incorrect conclusions were observed in any attempt, so each bar comprises only True Prediction (TP) and Uncertain Prediction (UP) statements. The full-ACO condition consistently produces the most correct conclusions, whereas the condition without ACO produces the fewest. Panels (c) and (f) display the four R_w values for the nine no-compression and nine compression attempts, respectively. Without compression, the three configurations return broadly comparable NPD refinements, while memoWrite prevents the reduced quality observed in several PDF fits in the no-ACO baseline. Under compression, the full-ACO condition maintains low R_w values for all four datasets, while removing CoreSkill degrades the PDF fits, and removing both mechanisms yields the poorest fitting quality and incomplete PDF refinements. Together with the condition-level metrics, these results show that ACO improves both performance and repeatability, with the role of CoreSkill becoming especially important when long-horizon analysis is subject to context compression.


(a)
Score
Token Counts
1e6
(b)
TP
UP
Conclusions Count
(c)
R_w
AgMnPbSbTe4-NPD
AgMnGePbSbTe5-NPD
AgMnPbSbTe4-PDF
AgMnGePbSbTe5-PDF
No Compression & Memo & CoreSkill
No compression & Memo & No CoreSkill
No compression & No Memo & No CoreSkill


(d)
Score
Token Counts
1e6
(e)
TP
UP
Conclusions Count
(f)
R_w
AgMnPbSbTe4-NPD
AgMnGePbSbTe5-NPD
AgMnPbSbTe4-PDF
AgMnGePbSbTe5-PDF
Compression & Memo & CoreSkill
Compression & Memo & No CoreSkill
Compression & No Memo & No CoreSkill

1

Figure 4. Factorial ablation study of the ACO mechanism with and without a 50k-token context-compression setting. (a–c) No-compression conditions; (d–f) corresponding compression conditions. (a,d) Composite evaluation score versus token consumption for each attempt. (b,e) Breakdown of selected conclusions into correct (TP) and uncertain (UP) statements; no incorrect conclusions were produced. (c,f) R_w values for the four fitted datasets across the three conditions × three repeats in each context setting. Missing data points in panel (f) indicate incomplete refinements due to lost file paths under full ablation.

Discussion

In this work, we present DrSai-Rongzai, an autonomous LLM-based agent that integrates domain-specific MCP tools with a novel ACO mechanism. Our results demonstrate that the agent achieves expert-level performance in both Rietveld refinement and PDF fitting across multiple material systems, while maintaining robust execution under long-horizon, multi-step iterative tasks, a regime that typically challenges standard LLM workflows. To our knowledge, this is the first reported deployment of an autonomous agent for joint analysis of multiple neutron data types within a production beamline environment at a spallation neutron source.

The core methodological contribution of this work lies in the ACO mechanism, which directly addresses the intertwined challenges of context inflation and critical information loss under automatic compression in scientific workflows. Unlike conventional static prompting or summarization, ACO enables the agent to actively manage its own context by storing expert knowledge (coreSkillWrite) and transient file paths (memoWrite). The ablation study (Fig. 4) offers compelling evidence for this design. Notably, while retaining skill content increases initial token consumption, it substantially improves overall cost-efficiency in long-horizon tasks. Disabling coreSkillWrite degrades PDF refinement quality, while disabling memoWrite forces the agent to spend substantially more tokens reconstructing lost file paths, occasionally resulting in complete fitting failures. Importantly, the repeatability analysis demonstrates that these effects are not confined to mean performance. ACO also constrains run-to-run variability, making the agent's analytical behavior more predictable under repeated execution. This counter-intuitive result suggests that

targeted context retention is not merely beneficial for quality but is also more resource-efficient in the long run by avoiding expensive recovery actions.

Beyond methodological novelty, DrSai-Rongzai represents a paradigm shift in facility data analysis workflows. By decoupling algorithmic kernels (GSAS-II/PDFgui) from orchestration semantics via MCP and a standardized JSON snapshot structure, the system provides unprecedented reproducibility and auditable intermediate states for autonomous analysis. Combined with a permission-gated database interface, the agent transitions from a research prototype to a legitimate participant in the facility's data analysis workflow.

We acknowledge several limitations. The current benchmark, while demonstrative, covers a limited set of well-characterized materials. Generalizability to highly complex or poorly understood phases remains to be tested. Furthermore, the agent operates within predefined tool sets and does not autonomously propose novel structural models or modify underlying fitting algorithms. The ACO mechanism, while effective, introduces an initial token overhead when coreSkillWrite is active. Future work will explore adaptive strategies that dynamically decide which skill fragments to retain. We also plan to integrate human-in-the-loop fallback and uncertainty quantification for low-confidence situations, and to extend the agent to time-resolved in-situ datasets. As foundation models continue to evolve, we anticipate that the agentic paradigm will become an indispensable component of future smart laboratories, accelerating the scientific return of every beam-hour.

Methods

System Prompt Design

The system prompt defines the agent's role, general constraints, skill descriptions, base-tool schemas, and reserved slots for dynamic content insertion. To ensure stable long-horizon behavior, the constraints follow a hierarchical priority: system prompt > user input > skill > tool feedback. This prevents low-priority information from overriding core instructions. The role definition and general constraints are as follows:

1) Role definition:

You are a professional analytical AI agent for neutron scattering data: Rongzai.

2) General constraints:

(1) For any specific tasks, first identify and read the relevant skill. Never skip the skill and execute directly based on experience. (2) Promote a skill to a CoreSkill if it involves multi-step tasks referenced across multiple decision rounds. Only one CoreSkill can exist at a time. (3) Downgrade a CoreSkill immediately when no longer needed. (4) Before each tool call, clearly state “what to do next” and “why do it”. (5) Use only currently available tools. When information is insufficient, clarify with the user. Do not fabricate data. (6) Automatically execute the full workflow by default. Every output must include a tool call to prevent the system from stopping. Stop only when the task is concluded or user input is necessary. (7) Forging tool calls or returning empty or whitespace-only text is strictly prohibited.

3) Skill descriptions and elevation logic:

(1) NPD refinement skill. Name: gsas2-refine. Description: Rietveld refinement skill for neutron powder diffraction using the GSAS-II engine. It can't be used for X-ray data. It also covers obtaining and processing files required for refinement. This skill must be read first whenever the user requests data refinement using GSAS-II. This skill should be elevated to a CoreSkill to guide your refinement-related decisions and operations continuously. (2) PDF fitting skill. Name: pdf-refine. Description: PDF refinement skill for pair distribution function fitting using the current PdfFit snapshot workflow. This skill must be read first whenever the user requests PDF data refinement. This skill should be elevated to a CoreSkill to guide refinement-related decisions and operations continuously.

4) Base-tool information:

The agent is directly equipped with a fixed set of local tools, plus dynamically registered MCP proxy tools. The full tool names and purposes are listed in Table III. The agent’s system prompts also contain the base tools’ schema about the accurate usage, which are too lengthy to present here.

Table III. The base tools and their purposes in the agent.

Category	Tool Name	Purpose
Skill Management	Skill	Load a skill by name and inject its content into runtime context
ACO	TodoWrite	Create or update the active structured todo list for complex workflows
	CoreSkillWrite	Promote one loaded skill into high-priority core runtime rules or remove a promoted core skill when it is no longer needed to conserve resources
	MemoWrite	Maintain a compact runtime memo of key facts, file paths, and constraints
MCP Integration	ConnectMcpServer	Connect an external MCP server and register all exposed tools into current runtime
	mcp_<remote_tool_name>	Call a specific remote MCP tool after MCP connection (dynamically registered)
Vision	AnalyzeImageWithVision	Delegate image understanding to a dedicated vision model
Shell	run_bash	Execute a shell command and wait for completion
File Operations	run_read	Read a local text file by line range
	run_write	Write full content into a file
	run_edit	Replace the first exact occurrence of old_text with new_text in a file
	serve_files_for_user	Package files as downloadable artifacts for frontend delivery

MCP Tools Implementations

The MCP tools layer exposes the underlying analytical backends to the agent through standardized interfaces. For NPD, the Rietveld refinement toolkit (based on GSAS-II) exposes task-oriented tools such as initialization, refinement, and structure manipulation. For PDF analysis, the fitting toolkit (based on PDFgui) was developed, validated, and integrated specifically for this work. The toolkit exposes task-oriented

MCP tools for PDF fitting through the same interface as the NPD refinement toolkit.

To robustly support long-running, multi-step refinements under agent orchestration, we implement a generic, tool-agnostic subprocess polling controller within the agent's harness. The controller initiates a computational task as a child process and repeatedly checks its status at fixed intervals until completion or timeout. If a task exceeds its deadline, the tool returns a structured pending snapshot and explicitly instructs the agent to continue with the same mission and task identifiers. This enables non-blocking, multi-turn continuation instead of a single blocking remote procedure call (RPC). At the agent-connection layer, server-sent events (SSE) read timeouts are extended to prevent transport-level premature disconnects during heavy workloads, thereby preserving task-level continuity. This architecture decouples the algorithmic kernels (GSAS-II and PDFgui routines) from the orchestration logic, namely stateful task identity, polling, and continuation, greatly improving the reproducibility and recoverability of autonomous scientific workflows.

Database Permission and Security Workflow

The facility data interface is implemented as a permission-gated workflow rather than as a general database or filesystem browsing capability. The agent is not initialized with unrestricted access to the CSNS data archive. Each data analysis request must pass through a DOMAS authorization gate before any proposal data path is exposed to the agent.

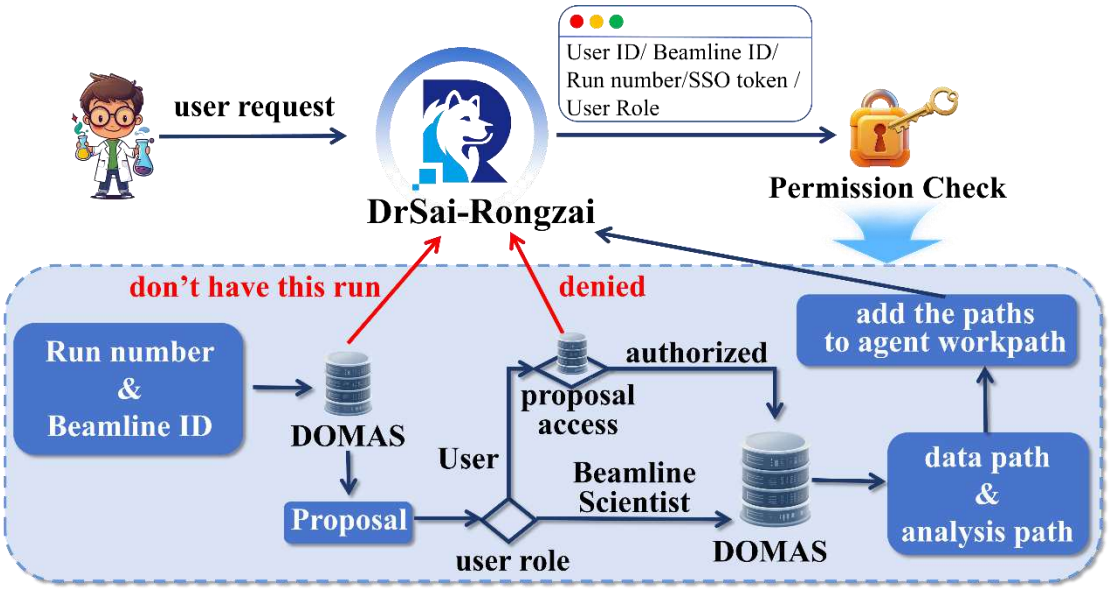


Figure 5. Permission-gated DOMAS data workflow used by DrSai-Rongzai to couple user authentication, run-level authorization, and scoped filesystem access.

As shown in Fig. 5, authentication metadata is supplied by the web interface during the session handshake. Only a whitelisted set of browser cookies, including the (Single Sign-On) SSO token, authentication provider, and user role, is copied into the session record and refreshed by the agent at the beginning of each permission check. The permission-checking tool then binds three pieces of information: the authenticated user identity, the beamline identifier, and the normalized run number. The SSO token is used to query the user-accessible beamlines, proposals, and runs through the DOMAS permission provider. For beamline scientists, the requested run can be resolved directly to its proposal through the facility service; for ordinary users, the agent verifies whether the requested run appears under one of the user's authorized proposals. This process exposes a single logical result to the agent: either a structured failure message explaining why access is denied or incomplete, or a successful authorization record containing the proposal identifier and the reduction directory associated with the requested run.

Only after this authorization succeeds is the agent granted file-level access to the relevant data locations. The workflow checks that the proposal contains a reduction directory and that at least one subdirectory matches the requested run number. If multiple matching run folders are found, the agent must ask the user to disambiguate before analysis proceeds. Once the run is uniquely identified, the corresponding analysis directory is created if it does not already exist. This directory, together with the authorized reduction directory, is added to the agent's allowed working paths. This design keeps the model isolated from unauthorized proposals by default. File operations remain confined to the normal workspace until the permission tool explicitly expands the accessible path list for a verified user-run-proposal tuple.

The same gate also defines the output policy. Once permission has been granted and analysis has produced an improved result, the agent is required to write the analysis report and important derived files back to the proposal-level analysis directory using stable filenames. These filenames follow a standardized naming

convention and are overwritten on update, preventing the proliferating of multiple uncontrolled copies. In this way, authorization, data retrieval, analysis execution, and result archiving are coupled into a single auditable workflow. The permission layer therefore serves not only as an access filter but also as the boundary that converts the agent from a general conversational assistant into a controlled participant in the facility data lifecycle.

Evaluation Metrics

We employed two quantitative evaluation frameworks, corresponding to the analyses presented in Fig. 3 and Fig. 4.

1) Normalized-Deviation Analysis for Parameter Agreement (Fig. 3c)

To assess consistency between agent-refined parameters and human-reference values, we defined the normalized deviation for each parameter i and agent attempt j as:

$$Z_{i,j} = \frac{X_{i,j} - X_{i,\mathrm{human}}}{\sigma_i} \tag{1}$$

where $X_{i,j}$ is the value predicted by agent attempt j for parameter i. $X_{i,human}$ is the human reference value. σ_i is the uncertainty associated with the human reference value. A value of $|Z| \leq 1$ indicates that the agent's result falls within one human uncertainty, considered statistically consistent. The fraction of parameters satisfying this condition was reported as a summary statistic.

2) Composite Evaluation Metric for Multi-Data Analysis (Fig. 4a)

To quantitatively evaluate the refinement quality across all datasets, we utilized the standard Weighted Profile R-factor (R_w), defined as:

$$R_w = \sqrt{\frac{\sum_i \omega_i (y_i^{obs} - y_i^{clac})^2}{\sum_i \omega_i (y_i^{obs})^2}} \times 100\% \tag{2}$$

where y_i^{obs} and y_i^{calc} are the observed and calculated intensities at point i in the diffraction/PDF profile, respectively, and w_i is the statistical weight (typically $w_i = 1/y_i^{\mathrm{obs}}$). For the ACO ablation study, we designed a composite score to holistically

rank each attempt, integrating conclusion correctness, conclusion recall (hit rate), and refinement quality from the four fitted data (two NPD and two PDF).

Conclusion-correctness score. Let P_j be the set of conclusion IDs returned by the agent in attempt j. Based on a pre-established human-annotated conclusion set, we partition all possible conclusions into three ground-truth categories: TRUE (supported by human analysis), FALSE (contradicted), and Uncertain (insufficient evidence). We define:

$$\mathrm{T}P_j = \left| P_j \cap TRUE \right| \tag{3}$$

$$\mathrm{F}P_j = \left| P_j \cap FALSE \right| \tag{4}$$

$$\mathrm{U}P_j = \left| P_j \cap Uncertain \right| \tag{5}$$

The correctness score is:

$$S_{correct,\,j} = \frac{TP_j}{\max(1,\, TP_j + \mathrm{F}P_j + \alpha\, \mathrm{U}P_j + \varepsilon)} \tag{6}$$

Where α=0.3 is the penalty coefficient assigned to uncertain conclusions, and $\varepsilon=10^{-9}$ is introduced for numerical stability. Under this formulation, false conclusions are penalized strongly, whereas uncertain conclusions incur only a weak penalty.

Human-conclusion hit rate – These measures recall over the true conclusions:

$$S_{hit,\,j} = \frac{TP_j}{TRUE} \tag{7}$$

Refinement-fit score. For each of the four data, the R_w value (denoted x_m) is transformed using an exponential decay:

$$S_{Rw_m} = e^{-x_m} \tag{8}$$

If an R_w value is NaN (e.g., due to non-convergence), it is assigned $S_{Rw_m} = 0$. The overall fit score is the average of the four:

$$S_{fit,\,j} = \frac{1}{4} \sum_{m\,=\,1}^{4} S_{Rw_m,\,j} \tag{9}$$

Final composite score. The three components are combined as:

$$S_{total,\,j} = 0.28\, S_{correct,\,j} + 0.12\, S_{hit,\,j} + 0.60\, S_{fit,\,j} \tag{10}$$

where the 60% weight allocated to refinement quality reflects its primary importance

in structural analysis, as errors in the physical fit render any subsequent analytical conclusions unreliable.

Condition-level ablation metrics. For each experimental condition c, the composite scores from the three independent repetitions $S_{c,1}$, $S_{c,2}$, and $S_{c,3}$ were summarized by their arithmetic mean:

$$\overline{S_c}=\frac{1}{3}\sum_{r=1}^{3} S_{c,r} \tag{11}$$

To assess reproducibility, we calculated a repeatability-stability index from the mean absolute pairwise difference among the three scores:

$$R_c=1-\frac{\left|S_{c,1}-S_{c,2}\right|+\left|S_{c,1}-S_{c,3}\right|+\left|S_{c,2}-S_{c,3}\right|}{3} \tag{12}$$

Because the composite score is bounded in ([0,1]), this index also lies in ([0,1]); larger values indicate more reproducible performance, and $R_c = 1$ denotes identical scores across all three repetitions. The condition-level values reported in Table II were calculated from the non-excluded attempts and exported automatically by the analysis script.

Acknowledgments

This work is supported by the National Key R&D Program of China (Grant No. 2023YFA1610000). Rong Du is grateful to Rongzai for her valuable inspiration and spiritual support.

Author contributions

R.D., Z.Z. and Junrong Z. conceived and designed the study. H.W. implemented the DrSai-Rongzai agent and the ACO mechanism, performed the benchmark and ablation experiments, and drafted the manuscript. Q.L., D.X., L.H., Y.S. and Jiajun Z. contributed to the implementation of the harness and the ACO mechanism and developed the NPD refinement skill and the GSAS-II MCP toolkit. J.X. and H.C. developed, validated and integrated the PDF fitting toolkit, performed the neutron scattering measurements, and contributed to data analysis and interpretation. J.L. and P.L. provided the high-entropy alloy and CsAgTe samples, respectively. F.Q. contributed to data management and storage infrastructure. Z.Z., R.D. and Junrong Z. supervised the project. H.W., R.D., Z.Z. and Junrong Z. revised the manuscript. All authors discussed the results and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The NPD and PDF data analyzed in this study were measured at the MPI beamline of the CSNS and are subject to the data-management policy of the CSNS. The raw beamline data are user-proprietary and are therefore not publicly available. The processed data supporting the findings of this study, including the reduced NPD

and PDFs, the refined structural parameters, the R_w values, and the refinement input and output files (GSAS-II project files and PDFgui fitting snapshots), are available from the corresponding author (Dr. Rong Du, durong@ihep.ac.cn) upon reasonable request, subject to approval by the CSNS.

References

[1] W. Zhu, M. Ma, Q. Sun, F. Liao, J. Wu, X. Che, K. Feng, Z. Yang, H. Huang, J. Zhong, T. Cheng, Y. Liu, M. Shao, Z. Kang, M. Ni, Interface-induced tensile strain stabilizes ruthenium oxide for high-performance acidic water oxidation, Nature Communications (2026). https://doi.org/10.1038/s41467-026-77521-0.

[2] M. Soda, E.M. Forgan, E. Blackburn, E. Campillo, V. Ryukhtin, I. Hoffmann, A. Kikkawa, Y. Taguchi, H. Yoshizawa, H. Kawano-Furukawa, Asymmetric slow dynamics of the skyrmion lattice in MnSi, Nature Physics 19 (2023) 1476–1481. https://doi.org/10.1038/s41567-023-02120-5.

[3] S. M., C. P., H. P.F., T. V.P., W. M.T., W. C.C., Determining hydrogen positions in crystal engineered organic molecular complexes by joint neutron powder and single crystal X-ray diffraction, CrystEngComm 16 (2024). https://doi.org/10.1039/c3ce42070a.

[4] M.W. Terban, S.J. Billinge, Structural analysis of molecular materials using the pair distribution function, Chemical Reviews 122 (2022) 1208–1272.

[5] Z. Chen, A.N. Petsch, A.J. Israelski, R. Plumley, L. Shen, C. Wang, C. Peng, Y. Ni, A. Bansil, S. Chowdhury, M. Li, J.B. Thayer, V. Thampy, J.J. Turner, An agentic artificially intelligent X-ray scientist, Nature Machine Intelligence 8 (2026) 1075–1086. https://doi.org/10.1038/s42256-026-01261-5.

[6] M. Cherukara, A. Vriza, H. Chan, T. Zhou, V. Katti Sastry, M. Prince, Calms: Context-aware language model for science, US Department of Energy (DOE) Software (2024) 69.

[7] T. Hellert, D. Bertwistle, S.C. Leemann, A. Sulc, M. Venturini, Agentic AI for multi-stage physics experiments at a large-scale user facility particle accelerator, arXiv E-Prints (2025) arXiv-2509.

[8] L. Ding, C. Do, Sasagent: Multi-agent ai system for small-angle scattering data analysis, arXiv Preprint arXiv:2509.05363 (2025).

[9] Z. Xiao, L. Zhang, G. Zhang, X. Wang, NeuDiff Agent: a governed AI workflow for single-crystal neutron crystallography, Applied Crystallography 59 (2026).

[10] Q. Li, H. Wang, D. Xiong, J. Zhong, W. Ji, H. Hu, Y. Zhang, B. Zhang, H. Wang, Y. Zhu, Rongzai agent: A Large Language Model-Based Autonomous Assistant for Rietveld Refinement of Neutron Diffraction Data, arXiv Preprint arXiv:2605.13911 (2026).

[11] M. Zhang, Y. Yang, R. Xie, B. Dhingra, S. Zhou, J. Pei, Generalizability of Large Language Model-Based Agents: A Comprehensive Survey, ACM Comput. Surv. 58 (2026). https://doi.org/10.1145/3794858.

[12] P. Tian, W. Zhou, J. Liu, Y. Shang, C. Farrow, P. Juhás, S. Billinge, SrRietveld: a program for automating Rietveld refinements for high-throughput powder diffraction studies, J. Appl. Crystallogr. 46 (2013) 255–258.
[13] C.M. Biwer, Z. Feng, D. Finstad, M. McDonnell, M. Knezevic, M. McKerns, D.J. Savage, S.C. Vogel, Spotlight: efficient automated global optimization in rietveld analysis of diffraction data, Scientific Reports 15 (2025) 8358. https://doi.org/10.1038/s41598-025-92452-4.
[14] Z. Feng, Q. Hou, Y. Zheng, W. Ren, J.-Y. Ge, T. Li, C. Cheng, W. Lu, S. Cao, J. Zhang, Method of artificial intelligence algorithm to improve the automation level of Rietveld refinement, Comput. Mater. Sci. 156 (2019) 310–314.
[15] A. Aimi, Kenjiro Fujimoto, Development of an automatic, high-throughput structural refinement method using Rietveld analysis, ACS Combinatorial Science 22 (2019) 35–41.
[16] Y. Ozaki, Y. Suzuki, T. Hawai, K. Saito, M. Onishi, K. Ono, Automated crystal structure analysis based on blackbox optimisation, Npj Computational Materials 6 (2020) 75.
[17] X. Cui, Z. Feng, Y. Jin, Y. Cao, D. Deng, H. Chu, S. Cao, C. Dong, J. Zhang, AutoFP: a GUI for highly automated Rietveld refinement using an expert system algorithm based on FullProf, Journal of Applied Crystallography 48 (2015) 1581–1586.
[18] A.S. Anker, E.T. Kjær, M. Juelsholt, K. Jensen, POMFinder: identifying polyoxometallate cluster structures from pair distribution function data using explainable machine learning, Applied Crystallography 57 (2024) 34–43.
[19] W. Sun, M. Lu, Z. Ling, K. Liu, X. Yao, Y. Yang, J. Chen, Scaling Long-Horizon Agent via Context Folding, (n.d.).
[20] M. Kang, W.-N. Chen, D. Han, H.A. Inan, L. Wutschitz, Y. Chen, R. Sim, S. Rajmohan, Acon: Optimizing context compression for long-horizon llm agents, arXiv Preprint arXiv:2510.00615 (2025).
[21] P.A.Jr.I. Pajo, Modular Agent Capabilities in Large Language Models: An Analysis of Anthropic's Claude Skills Framework with Applications in Document Processing and Retrieval-Augmented Generation, (2025). https://doi.org/10.5281/zenodo.17388387.
[22] H. Hu, L. Lei, H. Wang, B. Zhuang, R. Zhang, Q. Luo, X. Sun, F. Qi, DOMAS: a data management software framework for advanced light sources, Synchrotron Radiation 31 (2024) 312–321.
[23] B.H. Toby, R.B. Von Dreele, GSAS-II: the genesis of a modern open-source all purpose crystallography software package, Journal of Applied Crystallography 46 (2013) 544–549.
[24] C.L. Farrow, P. Juhas, J.W. Liu, D. Bryndin, E.S. Božin, J. Bloch, T. Proffen, S.J.L. Billinge, PDFfit2 and PDFgui: computer programs for studying nanostructure in crystals, Journal of Physics: Condensed Matter 19 (2007) 335219. https://doi.org/10.1088/0953-8984/19/33/335219.
[25] P.-F. Liu, X. Li, J. Li, J. Zhu, Z. Tong, M. Kofu, M. Nirei, J. Xu, W. Yin, F. Wang, T. Liang, L. Xie, Y. Zhang, D.J. Singh, J. Ma, H. Lin, J. Zhang, J. He, B.-T. Wang,

Strong low-energy rattling modes enabled liquid-like ultralow thermal conductivity in a well-ordered solid, National Science Review 11 (2024) nwae216. https://doi.org/10.1093/nsr/nwae216.
[26] W. Xie, Y. Wang, R. Du, X. Hu, H. Wang, J. Huang, Y. Xiang, H. Song, Y. Cai, Z. Li, C. Li, Insights into the atomic structure of oxygen vacancy on Bi2MoO6/MXene heterojunction and its role for boosting photocatalytic NO oxidation, Applied Surface Science 638 (2023) 158104. https://doi.org/10.1016/j.apsusc.2023.158104.
[27] J. Li, Z. Ma, H. Wang, L. Li, J. Zhu, H. Chen, Y. Zhang, Z. Ti, J. Zhong, Y. Xia, P.-F. Liu, Y. Zhang, W. Yin, Boosting Thermoelectric Properties of High-Entropy Chalcogenides through Local Structural Distortion and Tailored Chemical Bonding, J. Am. Chem. Soc. 147 (2025) 41629–41638. https://doi.org/10.1021/jacs.5c12728.