XRDBENCH: BENCHMARKING AI AGENTS ON EXPERIMENTAL X-RAY DIFFRACTION ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier multimodal large language models (MLLMs) increasingly demonstrate strong scientific-reasoning and coding capabilities, but whether these capabilities translate into reliable execution of domain-specific experimental workflows remains unclear. We investigate MLLM agents for X-ray diffraction (XRD) analysis, a widely used materials-characterization technique whose interpretation remains labor-intensive and expertise-dependent. We introduce **XRDBench**, a benchmark of 14 tasks organized into four tiers spanning peak-level perception, scientific unit analyses, multi-step characterization, and complex workflows such as phase identification, quantitative phase analysis, and Rietveld refinement. The tasks use predominantly expert-labeled experimental thin-film and powder data and are executed through a sandboxed agentic harness that provides multimodal inputs and access to scientific and crystallographic tools. We evaluate five proprietary frontier models and one open-weight model using task-specific outcome metrics and seven-axis process rubrics. We find that frontier models demonstrate strong XRD knowledge, although no single frontier model consistently outperforms the others, while the evaluated open-weight model remains behind. Removing visual access generally degrades performance, emphasizing the multimodal nature of XRD analysis. Agentic analysis may be particularly valuable for high-throughput and autonomous laboratories, where jointly analyzing related patterns across a compositional gradient reduces token use per pattern without accuracy loss. These findings highlight both the current limitations and the potential of agentic materials characterization.

1 INTRODUCTION

Multimodal large language models (MLLMs) have progressed from powering chatbots to driving agents that can plan, write, and run code, and autonomously act across a wide range of domains. Today's frontier models include proprietary models such as GPT-5.6 (OpenAI, 2026) and Claude Opus 5 (Anthropic, 2026) and open-weight models such as Kimi K2.5 (Kimi Team, 2026) and Qwen3.5 (Qwen Team, 2026). Some frontier models now exceed human-expert baselines on graduate-level science benchmarks such as GPQA Diamond (Rein et al., 2024), and showcase multimodal understanding of diverse figures and diagrams from collegiate textbooks (Yue et al., 2024). Used as agents, these models can resolve real-world software issues (Jimenez et al., 2024) and carry out machine-learning engineering and paper-replication tasks (Chan et al., 2025; Starace et al., 2025). With domain knowledge, quantitative reasoning, multimodal perception, and expert-level coding ability, these models appear ready for complex scientific analysis. Yet there is no evidence that these capabilities combine into domain-expert-level work on experimental data.

X-ray diffraction (XRD) analysis offers a data-rich, theoretically grounded, and scientifically consequential testing ground for this question. XRD is one of the most widely used characterization techniques in materials science as it is how researchers determine which phases a synthesis produced (Cullity and Stock, 2014). The analysis is multimodal: an analyst reads the pattern as an image, computes and analyzes on the raw intensity arrays, and uses context such as the synthesis route and the elements present to inform next steps. The analyses that can be applied span a wide range of difficulties. Some quantities, such as crystallite size via the Scherrer relation (Scherrer, 1918), follow from a straightforward calculation. Other analyses, such as phase analysis, can be long and open-ended processes. Phase analysis consists of identifying which material phases are present in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a sample and in what relative amounts. It necessitates visual perception, quantitative reasoning, and software use, and demands deep knowledge of physics and chemistry. This analysis proceeds through iterations of hypothesis, fitting, and revision, culminating in Rietveld refinement for structure determination (Rietveld, 1969). A rich ecosystem of expert software supports this workflow for experimentalists, such as open-source software like GSAS-II (Toby and Von Dreele, 2013), the BGMN/Profex suite (Doebelin and Kleeberg, 2015), and commercial tools such as HighScore (Degen et al., 2014). Appropriate use of these tools requires domain expertise, parameter tuning, and knowledge of the underlying scientific and experimental context of a system. XRD analysis therefore demands exactly the mix of perception, domain knowledge, and tool use that MLLM agents claim to offer.

No existing benchmark evaluates the complete analysis workflows a scientist would conduct. Existing materials and chemistry benchmarks probe knowledge recall with text-based question answering (Mirza et al., 2025; Zaki et al., 2024), property prediction from composition (Dunn et al., 2020), or visual understanding of published figures (McGrath et al., 2026; Weng et al., 2025). We address this gap with **XRDBench**, a benchmark of fourteen XRD analysis tasks organized into four tiers of increasing reasoning complexity: perception, scientific unit, scientific multi-step, and scientific complex (Figure 1). The tasks are posed mainly on expert-labeled experimental thin-film and powder multi-phase data. The model receives what a scientist would see: the raw intensity array, an image of the XRD pattern, and synthesis and/or instrument context. Our contributions are: (i) a physically grounded benchmark on experimental data spanning the full XRD analysis workflow; (ii) an agentic harness that runs each model in a sandbox with the crystallographic software experts use (GSAS-II, pymatgen (Ong et al., 2013), and SciPy (Virtanen et al., 2020)); and (iii) a multimodal evaluation of frontier and open-weight models that scores both outcome and reasoning trace.

2 Related Work

Existing materials-science benchmarks span several levels of evaluation. Matbench targets conventional supervised machine-learning models for property prediction from material composition or crystal structure (Dunn et al., 2020), whereas ChemBench and MaScQA primarily assess LLM domain knowledge and reasoning through text-based question answering (Mirza et al., 2025; Zaki et al., 2024). CrystalXRD-Bench isolates visual and crystallographic reasoning through a single task: predicting the Miller indices associated with the highest-intensity peak of a simulated pattern (Xu et al., 2026). This provides a precise diagnostic but does not assess broader XRD workflows or transfer to experimental patterns. MATRIX evaluates open-ended interpretations of experimental figures drawn from the literature, including XRD patterns, emphasizing multimodal alignment and image-grounded scientific explanation (McGrath et al., 2026).

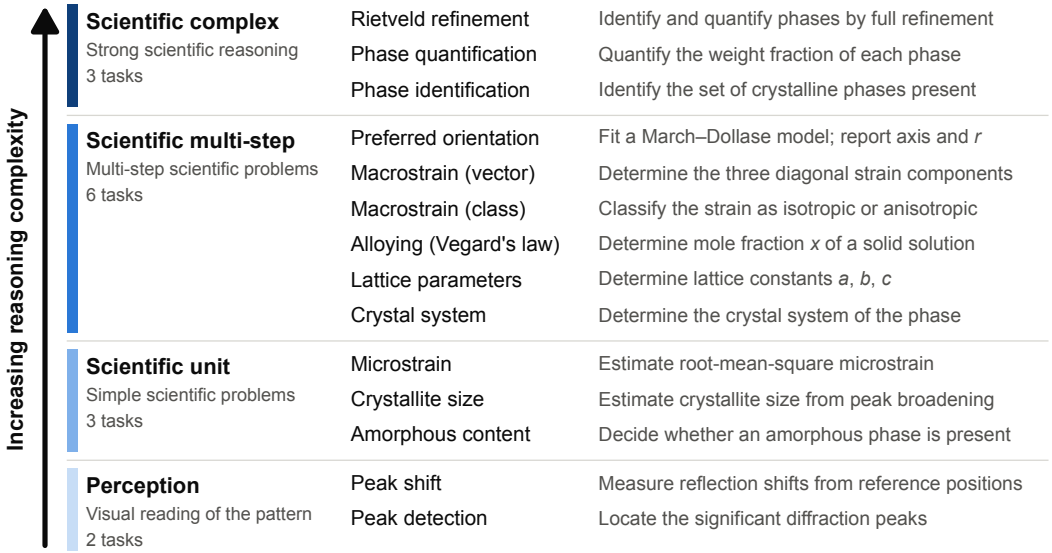
Agentic benchmarks are increasingly being developed to evaluate scientific tasks. These benchmarks include biology research skills (Laurent et al., 2024; Mitchener et al., 2025) and data-driven scientific discovery (Chen et al., 2025). AutoXRD brings such evaluation to XRD analysis through fine-grained diagnostic tasks and executable workflows applied to experimental powder diffraction data (Wu and Sun, 2026). To date, an agentic benchmark specific to XRD and covering diverse experimental samples has not yet been introduced. Here, we provide a comprehensive, fine-grained task suite with expert-reviewed ground truths, a broader experimental coverage spanning powder and thin-film samples, and evaluation of both isolated measurements and composition-gradients.

3 The Benchmark

3.1 Benchmark Tasks

We divide this benchmark into 14 tasks that encompass XRD analyses used to determine the structure and morphology of materials. These analyses span from simple to complex and allow us to evaluate different components of the XRD reasoning process.

Figure 1 gives an overview of the different tasks in the benchmark. We evaluate most tasks on experimental data, which allows us to evaluate if models can analyze and draw insights from real data produced in the lab. We divide the tasks based on their complexity and have four categories as outlined in Figure 1.



Tier	Task	What the agent must do
Scientific complex Strong scientific reasoning 3 tasks	Rietveld refinement	Identify and quantify phases by full refinement
	Phase quantification	Quantify the weight fraction of each phase
	Phase identification	Identify the set of crystalline phases present
Scientific multi-step Multi-step scientific problems 6 tasks	Preferred orientation	Fit a March–Dollase model; report axis and r
	Macrostrain (vector)	Determine the three diagonal strain components
	Macrostrain (class)	Classify the strain as isotropic or anisotropic
	Alloying (Vegard's law)	Determine mole fraction x of a solid solution
	Lattice parameters	Determine lattice constants a, b, c
	Crystal system	Determine the crystal system of the phase
Scientific unit Simple scientific problems 3 tasks	Microstrain	Estimate root-mean-square microstrain
	Crystallite size	Estimate crystallite size from peak broadening
	Amorphous content	Decide whether an amorphous phase is present
Perception Visual reading of the pattern 2 tasks	Peak shift	Measure reflection shifts from reference positions
	Peak detection	Locate the significant diffraction peaks

Figure 1: Breakdown of tasks in the benchmark. We split our tasks into 4 categories of increasing reasoning complexity: perception, scientific unit, scientific multi-step, and scientific complex.

We define the benchmark in this way to develop a fine-grained understanding of the successes and failures of current models at common XRD analyses. Additionally, any of the simpler tasks may be a component of answering the more complex tasks. For example, when trying to identify a possible phase, it is essential for a model to know the difference if a peak shift is due to strain or alloying. Our tasks allow us to identify at what point within a long reasoning trace a model may be failing, and thus what datasets are needed for improving any current failures.

For each task we provide the XRD pattern as an array and as an image, and also share any additional files needed to solve the task such as a reference database, instrument file, or reference stick pattern. The exact prompt for every task is included in the supplementary archive `prompts_and_rubrics.zip`.

3.2 Experimental Data Sources

Our benchmark datasets are drawn from experimental XRD patterns and are labeled by experienced materials scientists. Figure 2 shows an example pattern from each dataset. Additional details on the datasets are given in Appendix A.

Experimental Thin-Film Dataset. The "thin-film" dataset was acquired as part of this work and consists of XRD patterns measured on thin-film mixed-metal-oxide samples, as well as phase labels from a scientist with 6 years experience in XRD analysis. The thin-film samples are synthesized using sputtering to build a compositional gradient across the silicon wafer substrates. The five chemical spaces present in these samples are Pt-Y-O, Eu-Pt-O, Ta-Y-O, Cu-Ta-Y-O, Co-Cu-Pt-O. We use different groups of XRD patterns for different tasks. For all of the perception, scientific unit, and scientific multi-step tasks except for Vegard's law, we use 8 crystalline single-phase patterns and 17 amorphous patterns with expert-derived labels for crystal system, crystallite size, lattice parameter, macrostrain, peak shift, and preferred orientation. For the Vegard's law task we use XRD patterns that come from the same wafer and exhibit a clear composition gradient and solid-solution relationship between peak position and lattice parameter. We look at two phase transitions for the Vegard's law task, with 9 and 6 XRD patterns each. For the scientific complex tasks, we use 35 XRD patterns with phase and space group labels. These 35 patterns come from the 5 chemical spaces mentioned above. Reference XRD patterns for this dataset are obtained from the PDF-5+ database distributed by the International Centre for Diffraction Data (ICDD) (Kabekkodu et al.,

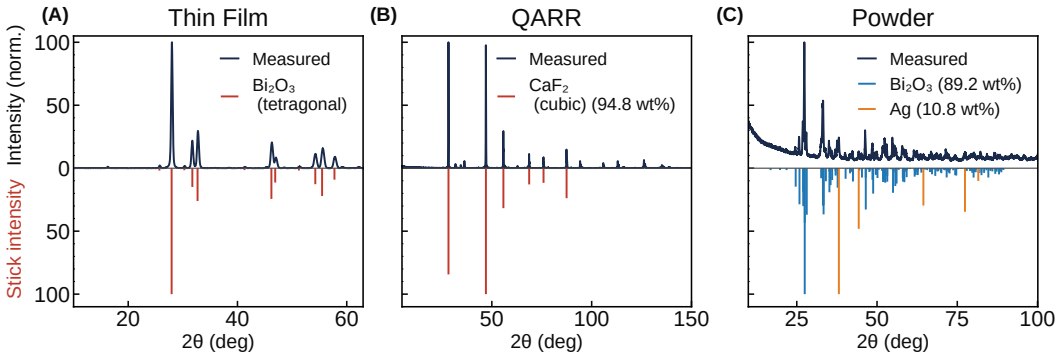


Figure 2: Details of the different data used in this benchmark. (A) An example XRD pattern from the thin-film dataset and corresponding reference stick pattern. (B) An example XRD pattern from the QARR dataset and the reference stick pattern of the dominant phase. (C) An example XRD pattern from the powder dataset with reference stick patterns of two identified phases.

2024), and structures are obtained from matching the ICDD ids with structures in the Inorganic Crystal Structure Database (ICSD) (Zagorac et al., 2019).

Quantitative Analysis Round Robin (QARR). The QARR dataset is built from the IUCr Commission on Powder Diffraction round robin on quantitative phase analysis of powder XRD (Madsen et al., 2001; Scarlett et al., 2002), comprising 12 certified multi-phase powder mixtures with gravimetrically-prepared, X-ray fluorescence (XRF) cross-checked weight fractions. This dataset is used for the QPA and Rietveld refinement tasks. For the Rietveld refinement task, we widen the original 13-phase candidate structures database to 166 phases with chemistry-plausible distractors using the Inorganic Crystal Structure Database (ICSD) as reference (Zagorac et al., 2019). This allows the evaluation of this dataset in an end-to-end Rietveld refinement workflow.

High-throughput Powder Dataset. The "powder" dataset (Walters et al., 2026) consists of 1,351 XRD scans of samples synthesized from pairwise solid-state reactions spanning 46 precursors and 39 elements, which have been labeled with phases using the Dara framework (Fei et al., 2026), with labels further reviewed by an expert and given a quality score. XRDBench uses a subset of the dataset that (i) have reference structures from the ICSD, (ii) are fit with an $R_{\text{wp}} \leq 10\%$, and (iii) have an expert-labeled quality score ≤ 2 (1 = excellent fit, 2 = good fit with minor discrepancies, and 3 = poor fit with missing phases). From the filtered entries, phase identification uses 123 patterns from 103 chemical spaces. Phase quantification and Rietveld refinement use 104 of those, the mixtures whose phases all have a reference space group. The perception, scientific unit, and scientific multi-step tasks use 26 single-phase patterns, 3 of which are already among those 123, plus 23 more. Together that is 146 patterns from 126 chemical spaces.

3.3 Reasoning-trace evaluation

Beyond final-answer correctness, we score the reasoning trace output by each agent. We design a master rubric that outlines the general expectations over different tasks and task-specific rubrics that detail how to apply the master rubric to a given task. The master rubric consists of seven axes: (1) **Grounding** (Does it understand the question being asked?), (2) **Measurement** (Are the numbers extracted rather than asserted?), (3) **Physics** (Is the domain model right for this question?), (4) **Alternatives & Confounds** (Does it consider what else the answer could be, and what else could be producing the signal?), (5) **Code execution** (Does the code run?), (6) **Tool use** (Does it reach for the right tools and the right data for this question?), and (7) **Answer integrity** (Does the reported answer follow from the work?). The master rubric and every task-specific rubric are given in full in the supplementary archive under `prompts_and_rubrics.zip`. An LLM judge, GPT-5.6 Sol, scores each axis on a $0/1/2$ scale (absent/partial/correct) given the full trace, the ground truth, and task context. The judge was chosen by evaluating which LLM gave scores closest to that of a PhD-

level researcher in the field, between GPT-5.6 Sol and Claude Opus 5 Appendix D. The normalized process score is the axis sum over its maximum, $S_{\text{reason}} = \sum_i s_i/14$.

4 THE HARNESS

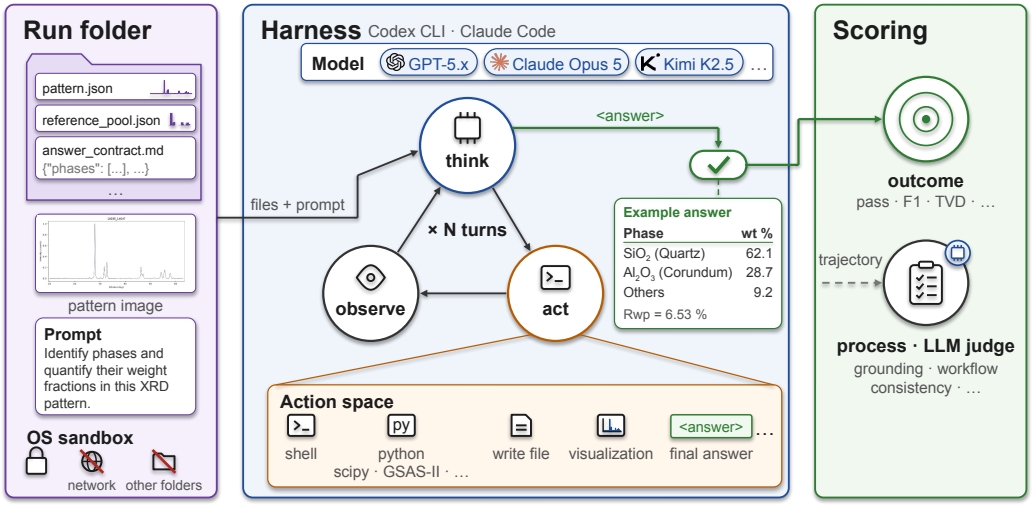


Figure 3: Overview of the harness. Each item runs in its own folder inside an operating-system sandbox. The model loops through think, act and observe until it emits a final answer; the action space is a shell, one scientific Python interpreter, file writes, and plot inspection. The answer is scored against the ground truth and the trajectory is scored by an LLM judge.

Design. Each evaluation item runs in its own working folder. At launch, the folder contains only the task inputs, such as the measured pattern, reference pool, synthesis context, and instrument data where applicable, together with a launcher for a single Python interpreter. The fixed scientific environment provides NumPy, SciPy, pandas, spglib, pybaselines, pymatgen, Matplotlib, Pillow, and GSAS-II for full pattern refinement. The model is driven by an agentic coding harness with shell and file tools. Its prompt specifies the task, environment, relevant instrument facts, and an answer contract requiring a structured JSON object enclosed in `<answer>` tags. An operating-system sandbox denies access to ground truth, other evaluation data, prior predictions, and scores. A run ends when the agent returns a final answer or stops. If it stops without the required tags, the harness may request the answer once more in the same session. The harness fixes the environment, functional action space, and execution protocol, but not the model's strategy. The model may answer immediately or perform many tool assisted analyses. This contract is held fixed within each model comparison. Figure 3 provides an overview.

Vision. When vision is enabled, as it is by default, the agent receives the measured diffraction pattern both as a numerical array and as a plot attached to the initial model context. The agent may also inspect supported image files that it generates during analysis. The no-vision arm omits the attached plot, removes image-specific prompt text, and prevents the agent from inspecting generated images, while retaining the numerical pattern and all other task inputs. This provides a controlled ablation of the image modality with the remaining evaluation protocol unchanged (Section 5).

Harness assignment. Each model family is paired with its native agentic harness, all driving the same sandboxed environment. GPT models run in Codex CLI and Claude models run in Claude Code. Open-weight models such as Kimi K2.5 run in Codex CLI through a chat-completions shim. Pairing each model with the harness it was built for lets every system operate at its best, while the fixed execution environment keeps results comparable. More details on the implementation of the harness are in Appendix C.

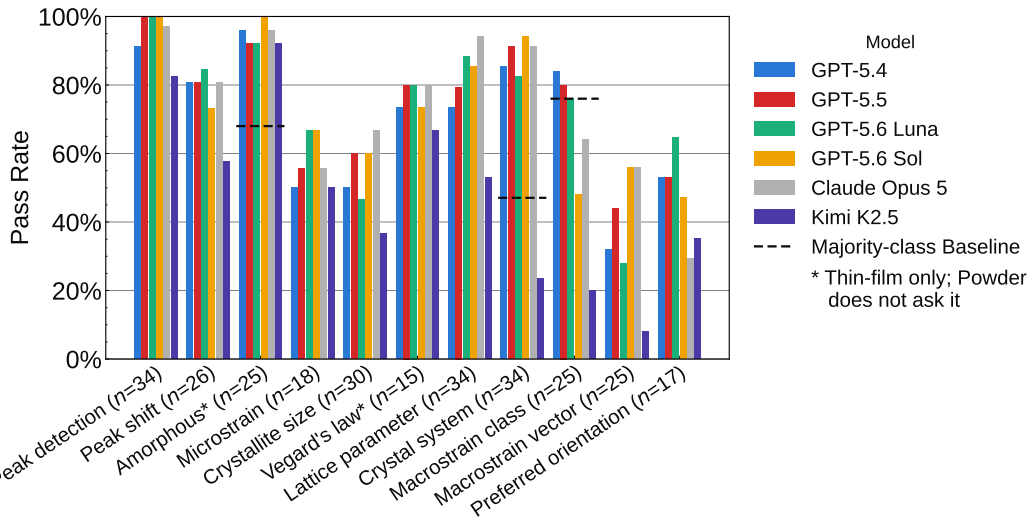


Figure 4: Pooled performance across thin-film and powder datasets for perception, scientific unit, and scientific multi-step tasks.

5 Results and Analysis

5.1 Benchmark on models

We benchmark recent GPT models (up to GPT-5.6 series), Claude Opus 5, and Kimi K2.5 on the XRD analysis tasks. For all tasks, we evaluate performance with GPT-5.4, GPT-5.5, GPT-5.6 Luna, and GPT-5.6 Sol at `xhigh` reasoning effort and Claude Opus 5 at `max`. Kimi K2.5 exposes no effort control. We compare performance of the largest model of the GPT series, GPT-5.6 Sol, across reasoning efforts `low`, `medium`, `xhigh`, and `max`. We evaluate the output traces using GPT-5.6 Sol and report the "process score": the normalized judge score over all axes.

Perception, Scientific Unit, and Scientific Multi-Step Tasks. Figure 4 shows the per-model pass rate for the non-complex tasks on the thin-film and powder diffraction data. In the appendix (Table 10) we compare the pass rate across datasets and different reasoning efforts of GPT-5.6 Sol. Overall, the tasks of macrostrain vector estimation, preferred orientation analysis, crystallite size estimation, and microstrain estimation perform worse across models and efforts. Generally, there is not a model that consistently outperforms the other models across tasks, and performance does not strongly vary with reasoning effort either (Table 10). Kimi K2.5 trails the best proprietary model on 9 of the 11 tasks. The appendix includes the metrics (Table 11) and "pass" criteria (Table 1) for each task, and all of the process scores (Table 12). Figure 9 in the appendix gives an overview of the average process score for these tasks. While pooling washes out between-task variance, performance increases with process score across models and reasoning efforts.

Scientific Complex Tasks. Figure 5 gives a summary of the complex task metrics for the three complex tasks of phase identification (Phase ID), phase quantification (quantitative phase analysis, QPA), and Rietveld refinement (RR). The appendix gives the full per-dataset breakdown for each task and all metrics (Table 5). The phase identification (Phase ID) task for the thin-film dataset is asked in two ways: by inputting all XRD patterns on a wafer and asking the model to label them all (plate), and inputting only one XRD pattern at a time (spot). Claude Opus 5, GPT-5.5, and GPT-5.6 Sol lead performance at the complex tasks. Additionally, reasoning effort does not consistently improve performance across tasks, but produces task-dependent gains, specifically for QPA (Table 4). Figure 6 shows the process scores for the phase ID and Rietveld refinement tasks, compared with a performance metric for each task. For the powder dataset, we find that process scores positively correlate for phase ID and Rietveld refinement – process score increases as exact-set rate increases for the phase ID task, and as R_{wp} decreases for the Rietveld refinement task. For the QARR dataset we find that the process scores do not correlate as strongly with decreasing

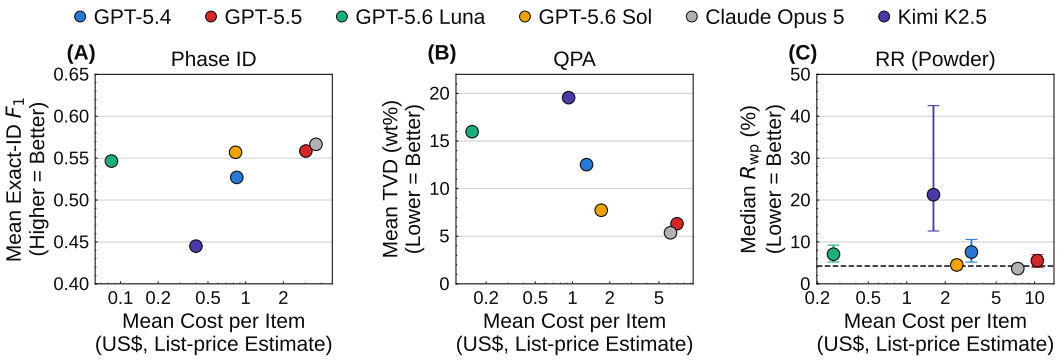


Figure 5: Per-model complex task results compared with their cost per item. (A) Exact-ID F_1 averaged across thin-film and powder datasets for each model for phase ID. (B) Mean total variation difference (TVD) between ground truth and output weight fractions averaged across powder and QARR datasets for QPA. (C) Median and interquartile range of R_{wp} when reported for the powder dataset in the Rietveld refinement (RR) task. R_{wp}: weighted residual profile; measures the weighted difference between the observed data points and the calculated profile during structure refinement. Horizontal dashed line: median R_{wp} (4.25%) from the reference dataset that consists of expert-reviewed high-throughput labels. Details on the cost calculation are in the appendix. Appendix F

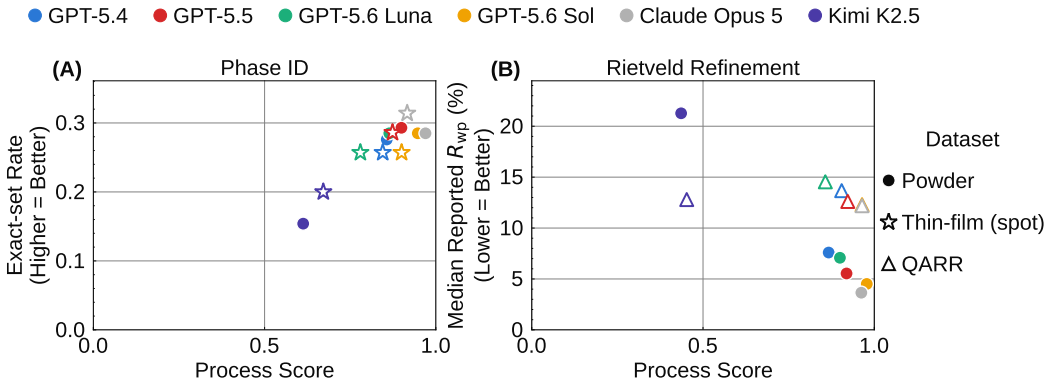


Figure 6: Process scores compared with performance for phase identification and Rietveld refinement tasks. Exact-set rate: exact match between predicted and ground truth phase set. (A) Phase ID, powder and thin-film datasets. (B) Rietveld refinement, powder and QARR datasets.

R_{wp}. This could be due to the performance of the QARR tasks saturating quickly. For the thin-film data, we see a positive correlation between exact-set rate and process scores. These results provide evidence that the rubrics capture task-relevant reasoning and find no significant knowing-doing gap. Detailed process scores are reported in Tables 6 and 7.

Cost and performance are not monotonically related: GPT-5.6 Luna has the lowest mean cost per item but can approach the more expensive model performances at Phase ID and Rietveld refinement. This advantage partly reflects the discounted token rate available for Luna when the evaluation was conducted and should not be interpreted solely as greater computational efficiency.

5.2 Task Analysis

Vision Ablation. The baseline input for all tasks includes an image or images of the XRD pattern, the XRD pattern as an array, and any additional necessary files such as instrument parameter files or reference stick patterns. We probe the multimodal nature of these tasks by removing the input image and preventing the agent from inspecting images generated during tool use.

As shown in Figure 7, removing the image modality degrades performance on 8 of 11 tasks for GPT-5.6 Luna and reduces its mean process score from 0.9 to 0.67. The largest declines occur in preferred orientation analysis, Vegard's law composition recovery, peak shift analysis, and lattice parameter estimation. In contrast, GPT-5.6 Sol sees degradation on 5 of 11 tasks, and the mean process score only reduces from 0.93 to 0.92.

Across different reasoning traces we observe the model generating images to look at the refinement fits, zoom in to particular peaks, and look at trends between different XRD patterns. While we cannot observe the exact reasoning traces of the proprietary models, we hypothesize that these images generated are necessary for the model to make decisions such as choosing parameters for peak fitting or refinement.

Because the ablation removes both the initial image and access to agent-generated plots, it does not isolate the differences between these contributions. Nevertheless, the results demonstrate that visual interaction is an important component of agentic XRD analysis.

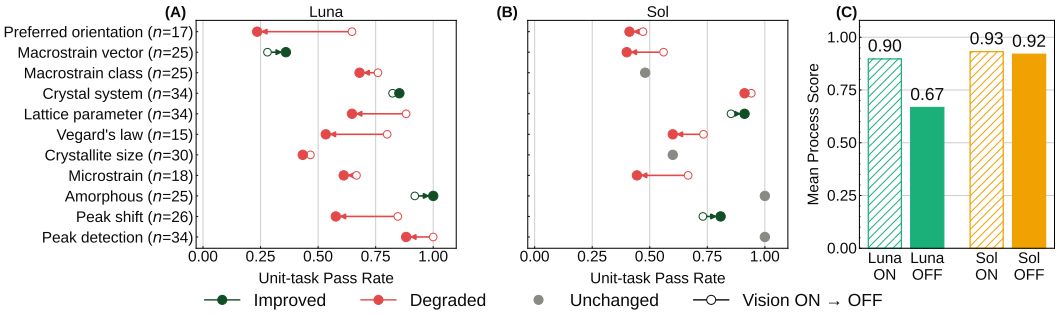


Figure 7: Vision ablation on non-complex tasks. (A) Performance of GPT-5.6 Luna. (B) Performance of GPT-5.6 Sol. (C) Average process scores of GPT-5.6 Luna and GPT-5.6 Sol with the vision modality turned on vs. off. Per-task process scores are in the appendix (Figure 10).

Plate vs. Spot token efficiency. The thin-film dataset allows us to test whether models can exploit physical trends along a composition gradient and reason jointly across related XRD patterns. In the "plate" setting of the phase ID task, the model jointly analyzes a series of 5–7 patterns from the same wafer and identifies the composition and space group for each pattern. In the "spot" setting, each pattern is analyzed independently. Figure 8 compares task performance with the mean cost per spot for the two settings in (A) and (B).

Increasing reasoning effort does not produce consistent improvements in accuracy. By contrast, jointly analyzing multiple patterns reduces the output tokens required per pattern at every reasoning effort. We hypothesize that the plate setting allows the model to reuse analysis steps across patterns from the same chemical system and to treat systematic peak shifts and phase changes along the composition gradient as cross-pattern constraints. This shared context may enable the agent to analyze subsequent patterns more efficiently than when each pattern is presented independently. Figure 8 (C) and (D) give example inputs for each setting.

6 Discussion

XRDBench shows that current agents possess substantial XRD knowledge but do not execute all analysis workflows reliably. Comparison with the expert-reviewed powder refinements provides additional context. Some agent configurations attain median R_{wp} values comparable to the reference median of 4.25%, but performance is not consistent across models (Figure 5 (C)). Moreover, R_{wp} measures profile agreement rather than correctness of the phase model and must therefore be interpreted together with phase-identity and weight-fraction accuracy.

Open-weight models currently trail the proprietary frontier, yet their development remains essential because unpublished compositions and synthesis conditions may require local or private analysis. An important open question is whether post-training methods, including supervised fine-tuning and

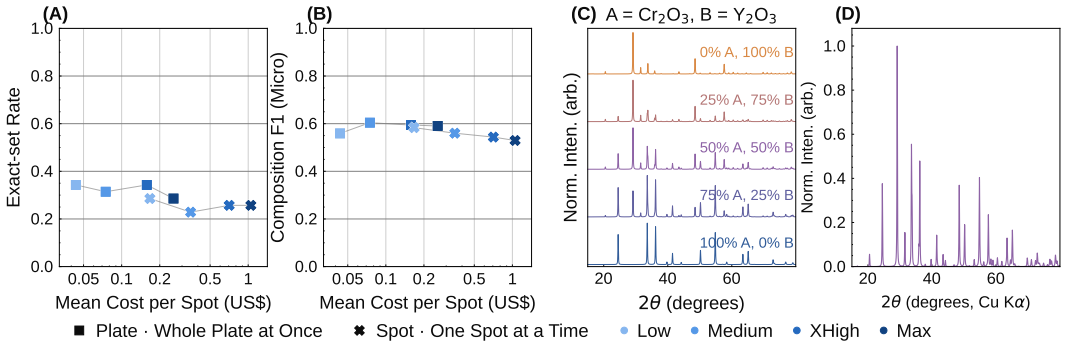


Figure 8: Performance of the thin-film data on the phase ID task as a function of per-spot token efficiency. (A) Exact-set rate. (B) Composition F_1 score. The micro-pooled F_1 score sums all true positives, false positives, and false negatives across the 35 spots before calculating one F_1 score. Per-spot token efficiency is calculated by dividing the total number of output tokens by the number of spots on the plate. (C) Example input for the "plate" setting. Each pattern is labeled with its ground truth phase set. (D) Example input for the "spot" setting.

reinforcement learning, can substantially narrow this gap. Simulated diffraction datasets such as SimXRD-4M (Cao et al., 2025), generated from structure databases such as the Materials Project (Jain et al., 2013), provide large-scale data with high fidelity synthetic labels. Archives of expert Rietveld refinements could likewise provide realistic demonstrations for agents, while the process rubrics developed here could guide reward design.

Agentic Rietveld refinement costs approximately \$0.27–\$10.50 per powder pattern across the evaluated models, with the two most expensive configurations requiring approximately \$5–\$10 per refinement. Multimodal tool use may contribute to token consumption, but our experiments do not isolate the cost of processing images. Cost differences also reflect model-specific pricing, including the discounted GPT-5.6 Luna rate available during evaluation. Although the direct API cost may be modest relative to the expert effort required for manual analysis, specialized pipelines operate at substantially higher throughput: Dara reports a median runtime of 88.9 s per pattern on an Intel Core i9-10920X workstation (Fei et al., 2026). Agentic systems must therefore justify their additional cost through capabilities that fixed pipelines do not provide, such as integrating experimental context, adapting the analysis strategy for complex workflows, and suggesting next steps.

Agentic analysis may be particularly valuable for high-throughput experiments, where related samples share candidate phases and exhibit systematic trends. Joint plate analysis allows agents to reuse analysis steps and exploit cross-pattern similarities, reducing cost per pattern (Figure 8). These capabilities make agentic XRD analysis particularly promising for autonomous laboratories.

7 Conclusion

We introduce XRDBench, a benchmark and agentic harness for evaluating models on experimental XRD analysis from perception to Rietveld refinement. Frontier models demonstrate substantial domain knowledge but remain inconsistent in executing complex workflows, and no model dominates across all tasks. The open-weight model trails the proprietary frontier, while visual access and joint analysis of related patterns provide meaningful benefits. We hope XRDBench will support reproducible measurement of progress and guide the development of more reliable, efficient, and privacy-preserving agents for materials characterization.

AI USE STATEMENT

We used LLM coding agents to assist with implementing the evaluation harness, build scripts, and scorers, and to assist with drafting the text and figures, all under author direction and finished and reviewed by the authors; the experiments were designed and their results interpreted by the authors without generative AI, and we take responsibility for the final content.

ETHICS STATEMENT

This work involves no human subjects; it uses public certified round-robin mixtures, a public autonomous-laboratory dataset, and a closed-source thin-film dataset, and the process scores come from an LLM judge whose biases we address by reporting them separately from outcome metrics.

REPRODUCIBILITY STATEMENT

The harness contract and its implementation details are in Section 4, the tasks in Section 3, the scoring rules in Appendix B, the datasets in Appendix A, the exact task prompts and rubrics in the supplementary archive `prompts_and_rubrics.zip`, and the evaluation repository with per-item run folders is provided as supplementary material so that every reported number can be re-scored from the stored trajectories.

REFERENCES

Anthropic. Introducing Claude Opus 5. https://www.anthropic.com/news/claude-opus-5, 2026.

Bin Cao, Yang Liu, Zinan Zheng, Ruifeng Tan, Jia Li, and Tong-Yi Zhang. SimXRD-4M: Big simulated X-ray diffraction data and crystal symmetry classification benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mkuB677eMM.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Madry. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.07095.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05080.

B. D. Cullity and S. R. Stock. *Elements of X-Ray Diffraction*. Pearson, Harlow, England, third edition, 2014.

Thomas Degen, Mohamed Sadki, Eric Bron, Udo König, and Gwilherm Nénert. The HighScore suite. *Powder Diffraction*, 29(S2):S13–S18, 2014. doi: 10.1017/S0885715614000840.

N. Doebelin and R. Kleeberg. Profex: a graphical user interface for the Rietveld refinement program BGMN. *Journal of Applied Crystallography*, 48(5):1573–1580, 2015. doi: 10.1107/S1600576715014685.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6(1):138, September 2020. doi: 10.1038/s41524-020-00406-3.

Yuxing Fei, Matthew J. McDermott, Christopher L. Rom, Shilong Wang, and Gerbrand Ceder. Dara: Automated multiple-hypothesis phase identification and refinement from powder X-ray diffraction. *Chemistry of Materials*, 38(3):1364–1376, February 2026. doi: 10.1021/acs.chemmater.5c02820.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.06770.

Soorya N. Kabekkodu, Anja Dosen, and Thomas N. Blanton. PDF-5+: A comprehensive Powder Diffraction File for materials characterization. *Powder Diffraction*, 39(2):47–59, 2024. doi: 10.1017/S0885715624000150.

Kimi Team. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. LAB-Bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Ian C. Madsen, Nicola V. Y. Scarlett, Lachlan M. D. Cranswick, and Thaung Lwin. Outcomes of the International Union of Crystallography Commission on Powder Diffraction round robin on quantitative phase analysis: samples 1a to 1h. *Journal of Applied Crystallography*, 34(4):409–426, 2001. doi: 10.1107/S0021889801007476.

Delia McGrath, Curtis Chong, Rohil Kulkarni, Gerbrand Ceder, and Adeesh Kolluru. MATRIX: A multimodal benchmark and post-training framework for materials science. *arXiv preprint arXiv:2602.00376*, 2026.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Mehrdad Asgari, Juliane Eberhardt, Amir Mohammad Elahi, Hani M. Elbeheiry, María Victoria Gil, Christina Glaubitz, Maximilian Greiner, Caroline T. Holick, Tim Hoffmann, Abdelrahman Ibrahim, Lea C. Klepsch, Yannik Köster, Fabian Alexander Kreth, Jakob Meyer, Santiago Miret, Jan Matthias Peschel, Michael Ringleb, Nicole C. Roesner, Johanna Schreiber, Ulrich S. Schubert, Leanne M. Stafast, A. D. Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17(7):1027–1034, July 2025. doi: 10.1038/s41557-025-01815-x.

Ludovico Mitchener, Jon M. Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P. Wellawatte, Andrew White, Lorenzo Sani, and Samuel G. Rodriques. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. doi: 10.1016/j.commatsci.2012.10.028.

OpenAI. GPT-5.6. https://openai.com/index/gpt-5-6/, 2026. Large language model, accessed September 19, 2026.

Qwen Team. Qwen3.5. https://qwen.ai/blog?id=qwen3.5, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

H. M. Rietveld. A profile refinement method for nuclear and magnetic structures. *Journal of Applied Crystallography*, 2(2):65–71, 1969. doi: 10.1107/S0021889869006558.

Nicola V. Y. Scarlett, Ian C. Madsen, Lachlan M. D. Cranswick, Thaung Lwin, Eric Groleau, Gregory Stephenson, Mark Aylmore, and Nada Agron-Olshina. Outcomes of the International Union of Crystallography Commission on Powder Diffraction round robin on quantitative phase analysis: samples 2, 3, 4, synthetic bauxite, natural granodiorite and pharmaceuticals. *Journal of Applied Crystallography*, 35(4):383–400, 2002. doi: 10.1107/S0021889802008798.

Paul Scherrer. Bestimmung der Grösse und der inneren Struktur von Kolloidteilchen mittels Röntgenstrahlen. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, pages 98–100, 1918.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, et al. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025.

Brian H. Toby and Robert B. Von Dreele. GSAS-II: the genesis of a modern open-source all purpose crystallography software package. *Journal of Applied Crystallography*, 46(2):544–549, 2013. doi: 10.1107/S0021889813003531.

Pauli Virtanen et al. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Lauren N. Walters, Matthew J. McDermott, Bernardus Rendy, Yuxing Fei, Kristin A. Persson, and Gerbrand Ceder. The Precursor Genome: A pairwise reaction dataset for solid-state synthesis. *arXiv preprint arXiv:2607.09903*, 2026.

Yonghao Weng, Liqiang Gao, Linwu Zhu, and Jian Huang. MatQnA: A benchmark dataset for multi-modal large language models in materials characterization and analysis. *arXiv preprint arXiv:2509.11335*, 2025.

Yuetong Wu and Maojun Sun. AutoXRD: Autonomous LLM agents and comprehensive evaluation for powder diffraction analysis. *arXiv preprint arXiv:2609.00070*, 2026.

Chengliang Xu, Xiaogang Li, Peiyao Xiao, Beng Wang, Hu Wei, and Bing Zhao. CrystalXRD-Bench: Benchmarking vision-language models for XRD peak indexing across diverse crystalline materials. *arXiv preprint arXiv:2605.29446*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024. arXiv:2311.16502.

Dejan Zagorac, Heiko Müller, Stefan Ruehl, Jelena Zagorac, and Sebastian Rehme. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *Journal of Applied Crystallography*, 52:918–925, 2019. doi: 10.1107/S160057671900997X.

Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. MaScQA: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024. doi: 10.1039/D3DD00188A.

A Data Details

Quantitative Analysis Round Robin (QARR). The Quantitative Analysis Round Robin (QARR) dataset is built from the IUCr Commission on Powder Diffraction (CPD) round robin on quantitative phase analysis (QPA) of powder XRD (Madsen et al., 2001; Scarlett et al., 2002). The original study compared QPA done by different laboratories on the same samples to document sampling errors *within* one laboratory, and differences in the analytical procedures *across* different laboratories. The dataset features multi-phase mixtures of varying proportions. Powder mixtures were prepared *gravimetrically* and cross-checked by X-ray fluorescence (XRF), so every sample carries a certified weight-fraction ground truth rather than a consensus label. We use the 12 mixtures with certified wt% measured using a Cu Kα beam. The measured patterns consist of eight "simple" three-phase corundum–fluorite–zincite-type oxide/fluoride mixtures, and four cases that break naive quantification — preferred orientation, amorphous content, microabsorption, and a natural-like bauxite with seven coexisting phases.

In the original round robin the candidate phases are effectively given (13 references, all true phases among them), so phase identification is near trivial and only quantification would be exercised. We keep the 12 certified samples but widen the candidate database to 166 phases: the 13 originals plus 153 real ICSD distractor phases filtered to the samples' chemical system (polymorphs of the true phases, in-system fluorides, and chemistry-plausible mixed oxides/silicates).

Experimental Thin-Film. Except for the Vegard's law data, the ground truth XRD patterns used for the perception, scientific unit, and scientific multi-step tasks consist of 8 crystalline single-phase patterns and 17 amorphous patterns. The crystalline XRD patterns are curve fit to obtain peak positions, widths, and intensities, and index peaks to their reference powder diffraction cards. Microstrain and crystallite size broadening were deconvolved using Williamson–Hall analysis and an instrument calibration file. After peaks are indexed, the measured peak positions are compared with the reference patterns to calculate peak shifts. The ICDD crystal system is taken as the ground truth, and the indexed peaks are used to solve for the lattice parameter of the sample. The derived lattice parameters are compared with the ICDD lattice parameters and used to calculate the labels for the macrostrain tasks. Preferred orientation is calculated by matching indexed reference reflections, and their integrated intensities, I_{obs} to the corresponding random-powder reference intensities, I_{ref}. For each candidate crystallographic axis, a March–Dollase model is fitted:

$$I_{\mathrm{fit}} = s I_{\mathrm{ref}} \left(r^2 \cos^2 \alpha + \frac{\sin^2 \alpha}{r} \right)^{-3/2},$$

where s is an intensity scale factor, α is the angle between the reflection vector and the candidate preferred-orientation axis, and r is the March coefficient. The axis and value of r that minimize the fitting error are selected. Only strong reference reflections, satisfying

$$I_{\mathrm{ref}} \geq 0.02\, I_{\mathrm{ref,max}},$$

are included, with a maximum of 60 reflections. The fit is accepted when

$$R_p = \frac{\sum_i |I_{\mathrm{obs},i} - I_{\mathrm{fit},i}|}{\sum_i |I_{\mathrm{obs},i}|} \leq 0.30.$$

A value $r \approx 1$ indicates an isotropic sample with no preferred orientation, whereas $r < 1$ and $r > 1$ indicate plate-like and needle-like texture, respectively. The amorphous task is evaluated with the crystalline and amorphous samples as input, while all other tasks only take the crystalline samples.

The Vegard's law task is evaluated on two wafers, from chemical spaces La-Zr-O and Mn-Zr-O and phase transitions from ZrO_2 to La_2O_3 and aZrO_2 to $Zr_0.7Mn_0.3O_2$, respectively (using their ICDD labels as ground truth). Their ground truth lattice parameters are calculated by peak indexing using the matched ICDD reference stick pattern, and the lattice parameter is used to estimate the mole fraction using Vegard's law, $a(x) = (1-x)a_A + xa_B$. The mole fraction is used as the ground truth for this task, for each XRD pattern. In this task the model is given each pattern individually.

Phase Identification Task. Finally, the phase identification task for this dataset has two iterations – plate and spot – where the model is either given a series of XRD patterns from the same compositional space on one wafer, or one XRD pattern, and tasked to identify the phase set from an ICSD reference database. The ground truth ICSD structures are those that match the ICDD labels. For this task, we have corresponding XRF measurements that give the elemental loadings of each XRD pattern in μg/cm^2 for additional context.

High-throughput Powder Dataset. For the powder diffraction dataset, non-complex task labels were extracted from the final fits of the expert-reviewed refinements. We selected 26 single-phase, crystalline patterns with no phase overlaps. These patterns were those with the best expert quality score and lowest R_{wp}, if there were duplicates. Peak locations and counts came from the fitted peak list; crystal system and lattice parameters from the refined structure; crystallite size from the BGMN size parameter using its implied Scherrer constant (K=0.8488); macrostrain class and directional components from deviations of the refined lattice from the reference structure; peak shifts from matched measured and reference reflections; microstrain from Williamson–Hall analysis; and preferred orientation from the converged March–Dollase fit. Labels requiring comparison with a reference pattern were retained only when reference stick positions were available. Because instrumental broadening was not characterized for this dataset, no instrumental-broadening correction was applied to the crystallite-size or microstrain labels.

B METRICS

Table 1 details the metrics and pass rate for each task.

Table 1: Scoring criteria by task.

Task	Metric	Pass threshold
Crystal system	Exact label match over the seven crystal systems, after case folding and alias normalization: $\mathbf{1}[\hat{s} = s]$	Exact match
Amorphous	Binary verdict $\mathbf{1}[\hat{a} = a]$; $\hat{a}$ is read from an explicit `amorphous` flag, or implied by a reported crystal system	Exact match
Macrostrain class	Exact label match, isotropic vs. anisotropic: $\mathbf{1}[\hat{c} = c]$	Exact match
Crystallite size	Relative error $\delta_D = \left\lvert \hat{D} - D \right\rvert / D$. On the thin-film set the label is instead the per-peak Scherrer interval $[D_{\min}, D_{\max}]$ obtained from every fitted reflection, which is the set of defensible answers under method variance. All three sets are graded on single-phase patterns	$\hat{D} \in [D_{\min}, D_{\max}]$ (thin-film set); $\delta_D < 0.20$ (synthetic set, where the generator parameters are exact ground truth); $\delta_D < 0.50$ (powder set, which carries no instrument-broadening characterization, so no β_{inst} can be given or removed)
Microstrain	Relative error on the Williamson–Hall strain term, $\delta_\varepsilon = \lvert \hat{\varepsilon} - \varepsilon \rvert \,/\, \lvert \varepsilon \rvert$	$\delta_\varepsilon < 0.50$; the looser tolerance reflects the size/strain separation being the noisiest step of a Williamson–Hall fit, and labels carrying 20–80% error bars
Lattice parameter	Unit-cell volume relative error, evaluated at the ground-truth cell angles so the metric is invariant to axis permutation: $\delta_V = \left\lvert \hat{V} - V \right\rvert / V$ with $V = abc\sqrt{1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}$	$\delta_V < 0.20$
Macrostrain vector	Per-component grading of the diagonal strain vector, canonicalized by crystal system so symmetry-equivalent components are order-free. With noise floor $\varepsilon_{\text{nf}} = 10^{-3}$ and grading floor $\varepsilon_0 = 5 \times 10^{-4}$ (the measured median per-reflection strain scatter): *resolvable* ($\lvert\varepsilon_i\rvert \geq \varepsilon_{\text{nf}}$), $\lvert\hat{\varepsilon}_i - \varepsilon_i\rvert \leq \max(\varepsilon_0,\ 0.20\,\lvert\varepsilon_i\rvert)$; *sub-noise* ($\lvert\varepsilon_i\rvert < \varepsilon_{\text{nf}}$), the sign is unrecoverable so only the magnitude claim is graded, $\lvert\hat{\varepsilon}_i\rvert \leq \varepsilon_{\text{nf}}$ or $\lvert\hat{\varepsilon}_i - \varepsilon_i\rvert \leq \varepsilon_{\text{nf}}$	All three components within tolerance: $\bigwedge_{i=1}^{3} \text{ok}(i)$
Peak shift	Each ground-truth peak p is matched to the nearest predicted peak $\pi(p)$ with $\lvert 2\theta_{\pi(p)} - 2\theta_p \rvert \leq 0.1^\circ$; unmatched peaks score ∞. Score is the mean absolute shift error $\bar{e} = \frac{1}{\lvert P \rvert} \sum_{p \in P} \left\lvert \hat{\Delta}_{\pi(p)} - \Delta_p \right\rvert$	$\bar{e} \leq 0.05^\circ$
Peak detection	Greedy one-to-one matching of predicted to true peak positions within a window $\tau_{2\theta}$ (0.1° thin-film set, whose peaks have a full width at half maximum (FWHM) of 0.08–0.17°; 0.3° powder set), then $F_1 = 2PR/(P + R)$ with precision $P = \text{TP}/\lvert\hat{P}\rvert$ and recall $R = \text{TP}/\lvert P \rvert$, where TP is the number of true positives (matched peaks). FWHM relative error on matched peaks is reported but not gated; intensities are not graded (labels are normalized)	$F_1 \geq 0.5$
Preferred orientation	Outcome-anchored residual band: a predicted axis $\mathbf{u}$ and March–Dollase coefficient r are correct iff they reproduce the measured intensities about as well as the data-optimal fit. With the sum of squared residuals $\text{SSR}(\mathbf{u}, r) = \min_s \sum_k \left(I_k^{\text{obs}} - s\, P_k I_k^{\text{ref}} \right)^2$, $P_k = \left(r^2 \cos^2\alpha_k + r^{-1} \sin^2\alpha_k \right)^{-3/2}$ and $\cos\alpha_k = \dfrac{\mathbf{h}_k^\top G^* \mathbf{u}}{\sqrt{(\mathbf{h}_k^\top G^* \mathbf{h}_k)\,(\mathbf{u}^\top G^* \mathbf{u})}}$. The labeled axis is not identifiable from the data, so it is metadata, not the grade	$\text{SSR}(\hat{\mathbf{u}}, \hat{r}) \leq \max\left(1.5\,\text{SSR}_{\text{best}},\ 5 \times 10^{-3}\right)$, intensities normalized to unit maximum; ungradable with < 4 usable reflections. Where per-phase intensities are unavailable, graded instead on $\lvert\hat{r} - r\rvert / r < 0.20$ with the axis reported but ungated
Phase ID / Rietveld refinement	Greedy one-to-one match of predicted to reference phases on reduced-formula keys (parentheses expanded, polymorph prefixes and structure qualifiers stripped, subscripts reduced), scored at two nested levels: *composition* (accepted composition, space group ignored; Composition F_1) and *exact ID* (composition and space group, up to data-indistinguishable polymorphs; Exact-ID F_1). With true positives (TP), false positives (FP), and false negatives (FN), $P = \text{TP}/(\text{TP} + \text{FP})$, $R = \text{TP}/(\text{TP} + \text{FN})$, $F_1 = 2PR/(P + R)$; their gap is the polymorph-discrimination cost. Phases outside the reference pool are excluded rather than penalized. Exact-set rate is $\mathbf{1}[\text{FP} = 0 \wedge \text{FN} = 0]$ at the exact-ID level, averaged over spots	Reported as a score; no pass threshold
Quantitative phase analysis (QPA)	Phase-set F_1 as above, plus two wt% distances over the union $U = S \cup \hat{S}$ (amorphous is its own component; missed and phantom phases contribute their full weight): total variation distance $\text{TVD} = \frac{1}{2} \sum_{k \in U} \lvert\hat{w}_k - w_k\rvert$, bounded in $[0, 100]$ and readable as the percentage of mass misassigned; and mean absolute relative error over the crystalline ground-truth phases $\text{MARE} = \frac{1}{\lvert S_c \rvert} \sum_{k \in S_c} \lvert\hat{w}_k - w_k\rvert / w_k$, which weights minor phases equally with major ones (a missed phase scores 1.0)	Reported as a score; no pass threshold. R_{wp} is passed through as a process diagnostic and never enters accuracy
Vegard's law	Mean squared error (MSE) against the instrument-corrected mole fraction, aggregated per plate: $\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{f}_i - f_i \right)^2$	Reported as a score; no pass threshold

C HARNESS IMPLEMENTATION

The reported models are served through Amazon Bedrock. The primary GPT arms use reasoning effort `xhigh`, whereas Claude arms use `max`; Kimi K2.5 exposes no corresponding reasoning effort setting. Up to ten items run concurrently for each task. Filesystem read confinement uses Seatbelt on macOS and bubblewrap on Linux because the coding harnesses do not reliably prevent reads outside the working directory. Codex CLI has no turn limit, whereas Claude Code is capped at 400 turns depending on the task cell; no recorded run reached its cap. If a session stops without an `<answer>` block, it is resumed once with a fixed prompt requesting only the final answer. Successful recovery and recovery without prior tool use are recorded separately in each prediction row.

D CHOOSING THE LLM JUDGE

The process score in the main text is assigned by an LLM judge reading the agent's tool-calling trajectory and reasoning trace where available, the files in the sandbox, the prompt, and scoring its trajectory using the master rubric and task-specific rubric. To choose the judge, we compare the scores of GPT-5.6 Sol and Claude Opus 5 with those of a PhD researcher in the domain, and choose the model that aligns closest with the expert.

Rubrics. The rubrics used for each task consist of a master rubric that is used to outline the general workflow necessary across tasks, and a task-specific rubric to specify how to apply the master rubric to the given prompt and context. The exact rubrics used for reasoning trace evaluation are included, together with the exact task prompts, in the supplementary archive `prompts_and_rubrics.zip`.

Procedure. We drew a ten-trajectory validation set spanning four tasks (phase ID on powder, phase ID on a thin-film plate, QPA, and Rietveld refinement) and two agent configurations, and had a domain expert score each trajectory using the rubrics for each task. Two candidate judges then graded the same set using the same rubrics. We compare distance between expert and LLM judge at the axis level. The comparison covers $n = 70$ scored axis pairs over 10 trajectories.

Table 2 gives every score; Table 3 is the distance.

Table 2: **Per-trajectory process scores: expert versus the two candidate judges.** Each seven-digit string is the axis vector $A_1 \ldots A_7$, scored $0/1/2$; Σ is the total out of 14.

		Expert		GPT-5.6 Sol		Claude Opus 5	
Trajectory	**Model**	$A_{1..7}$	Σ	$A_{1..7}$	Σ	$A_{1..7}$	Σ
PG_1021	GPT-5.6 Luna	2211222	12	2211222	12	2221222	13
PG_0850	GPT-5.6 Sol	2222222	14	2222222	14	2222221	13
PG_0950	GPT-5.6 Luna	2222122	13	2222222	14	2212221	12
PG_2150	GPT-5.6 Luna	2222122	13	2222222	14	2221222	13
PG_1115	GPT-5.6 Luna	2212222	13	2212222	13	2211221	11
plate_10087	GPT-5.6 Sol	2212221	12	2111221	10	2211221	11
PG_0711	GPT-5.6 Luna	1222212	12	2212112	11	1222121	11
PG_1046	GPT-5.6 Sol	2222221	13	2222221	13	2222221	13
PG_1112	GPT-5.6 Luna	2222221	13	2222222	14	2221211	11
qarr_cpd-1b	GPT-5.6 Sol	2222222	14	2222222	14	2222222	14
Mean			**12.9**		**12.9**		**12.2**

Decision. We adopt GPT-5.6 Sol as the LLM judge for all results reported in this work. It is closer to the expert on every statistic.

Table 3: **Judge–expert agreement on the validation set.** Distance is judge minus expert. A negative bias means the judge is stricter than the expert. MAE is the mean absolute error. "Exact vector" counts trajectories where the judge reproduces all seven of the expert's axis scores. "Shared deductions" counts axes the expert scored below maximum that the judge also scored below maximum (recall), and the fraction of the judge's own deductions the expert agrees with (precision). $n = 70$ axis pairs over 10 trajectories.

	Axis level			**Total**	**Exact**	**Shared deductions**
Judge	MAE	Exact	Bias	MAE	vector	recall / precision
GPT-5.6 Sol	**0.11**	**89%**	**+0.00**	**0.60**	**5/10**	64% / **64%**
Claude Opus 5	0.21	79%	−0.10	0.90	2/10	64% / 39%

Table 4: Complex-task metrics for GPT-5.6 Sol across reasoning effort settings. Accuracy metrics (↑) precede error metrics (↓) within each cell, and the arrow on each row gives its direction. n is the number of units each model was scored on: items, except on the thin-film rows, where it is spots. Thin-film rows omit the two single-spot plates.

Task	n	low	medium	xhigh	max
Phase ID, powder					
Composition F_1 ↑	123	0.768	0.782 (122/123)	0.772	0.771
Exact-ID F_1 ↑	123	0.598	0.600 (122/123)	0.596	0.594
Phase ID, thin-film (per plate)					
Exact-set rate ↑	35	0.343	0.314	0.343	0.286
Composition F_1 ↑	35	0.552	0.578	0.609	0.606
Exact-ID F_1 ↑	35	0.552	0.550	0.580	0.543
Phase ID, thin-film (per spot)					
Composition F_1 ↑	35	0.583	0.547	0.547	0.538
Exact-ID F_1 ↑	35	0.573	0.495	0.518	0.498
Exact-set rate ↑	35	0.286	0.229	0.257	0.257
QPA, powder					
TVD (wt%) ↓	104	18.4	12.1	7.2	6.5
MARE ↓	104	0.505	0.374	0.232	0.200
QPA, QARR					
TVD (wt%) ↓	12	11.4	8.0	8.3	4.4
MARE ↓	12	0.243	0.174	0.176	0.108
Rietveld refinement, powder					
Composition F_1 ↑	104	0.780	0.792	0.774 (103/104)	0.780
Exact-ID F_1 ↑	104	0.586	0.610	0.609 (103/104)	0.587
TVD (wt%) ↓	104	35.1	33.1	30.3 (103/104)	31.2
Rietveld refinement, QARR					
Phase-set F_1 ↑	12	0.944	1.000	1.000	1.000
TVD (wt%) ↓	12	15.4	11.1	6.0	5.6
MARE ↓	12	0.337	0.383	0.156	0.112

E Detailed Performance & Process Results

E.1 Complex Tasks

Here we report the task- and dataset-specific breakdown of metrics and process scores for the complex scientific tasks (Tables 4 to 7).

Table 5: Complex-task metrics. Accuracy metrics (↑) precede error metrics (↓) within each cell, and the arrow on each row gives its direction. One column per model, at the matched `xhigh` reasoning effort. n is the number of units each model was scored on: items, except on the thin-film (plate) rows, where it is spots.

Task	n	GPT-5.4	GPT-5.5	GPT-5.6 Luna	GPT-5.6 Sol	Claude Opus 5	Kimi K2.5
Phase ID, powder							
Composition F_1 ↑	123	0.748 (121/123)	**0.784**	0.771	0.772	0.772	0.601
Exact-ID F_1 ↑	123	0.564 (121/123)	0.600	**0.606**	0.596	0.590	0.438
Phase ID, thin-film (per plate)							
Exact-set rate ↑	35	0.086	0.029	0.343	0.343	0.286	0.171
Composition F_1 ↑	35	0.425	0.421	0.572	0.609	0.511	0.452
Exact-ID F_1 ↑	35	0.425	0.392	0.572	0.580	0.483	0.324
Phase ID, thin-film (per spot)							
Composition F_1 ↑	35	0.490	0.540	0.498	0.547	0.571	0.495
Exact-ID F_1 ↑	35	0.490	0.517	0.487	0.518	0.543	0.452
Exact-set rate ↑	35	0.257	0.286	0.257	0.257	0.314	0.200
QPA, powder							
TVD (wt%) ↓	104	8.7	7.7	17.8	7.2	**5.9**	24.1 (96/104)
MARE ↓	104	0.498	0.265	0.531	0.232	**0.159**	1.683 (96/104)
QPA, QARR							
TVD (wt%) ↓	12	16.3 (11/12)	4.9	14.2	8.3	4.9	15.0
MARE ↓	12	0.431 (11/12)	0.109	1.058	0.176	0.108	0.649
Rietveld refinement, powder							
Composition F_1 ↑	104	0.772 (102/104)	**0.799**	0.779	0.774 (103/104)	0.784 (103/104)	0.547 (87/104)
Exact-ID F_1 ↑	104	0.573 (102/104)	**0.628**	0.607	0.609 (103/104)	0.591 (103/104)	0.375 (87/104)
TVD (wt%) ↓	104	34.4 (102/104)	**30.7**	**30.7**	30.3 (103/104)	32.0 (103/104)	47.4 (87/104)
Rietveld refinement, QARR							
Phase-set F_1 ↑	12	0.917	0.988	0.972	1.000	1.000	0.674
TVD (wt%) ↓	12	10.7	6.7	15.0	6.0	5.0	33.4
MARE ↓	12	0.267	0.137	0.461	0.156	0.101	0.948

Table 6: Rubric process scores on the complex cells. The score is the judge's mean over that cell's items, out of the fixed 14 points and normalised to 0–1; it grades *how* the agent worked and is independent of whether the answer was right (Table 5 scores the answer). One column per model, at the matched `xhigh` reasoning effort; the ladder is the effort variant of this table. Thin-film rows omit the two single-spot plates. Every column was graded by GPT-5.6 Sol.

Cell / rubric	GPT-5.4	GPT-5.5	GPT-5.6 Luna	GPT-5.6 Sol	Opus 5	Kimi K2.5
Phase ID, powder						
Phase ID	0.857	0.900	0.864	0.947	**0.970**	0.613
Phase ID, thin-film (per plate)						
Phase ID, series	0.743	0.743	0.671	0.857	0.957	0.614
Phase ID, thin-film (per spot)						
Phase ID	0.845	0.873	0.780	0.900	0.916	0.671
QPA, powder						
QPA	0.938	0.963	0.895	**0.984**	0.961	0.644
QPA, QARR						
QPA	0.911	0.988	0.899	1.000	0.958	0.548
Rietveld refinement, powder						
Rietveld refinement	0.867	0.919	0.900	**0.978**	0.962	0.436
Rietveld refinement, QARR						
Rietveld refinement	0.905	0.923	0.857	0.964	0.964	0.452

Table 7: Process scores on the complex tasks (0–1). The GPT-5.6 Sol columns are that model at four reasoning efforts; the other models are at `xhigh`. Thin-film rows omit the two single-spot plates. Every column was graded by GPT-5.6 Sol.

Task	GPT-5.4	GPT-5.5	GPT-5.6 Luna	GPT-5.6 Sol				Claude Opus 5	Kimi K2.5
	xhigh	xhigh	xhigh	low	medium	xhigh	max	max	n/a
Phase ID powder	0.857	0.900	0.864	0.813	0.904	0.947	0.955	**0.970**	0.613
Phase ID thin-film, plate	0.743	0.743	0.671	0.700	0.786	0.857	0.843	0.957	0.614
Phase ID thin-film, spot	0.845	0.873	0.780	0.710	0.831	0.900	0.941	0.916	0.671
QPA powder	0.938	0.963	0.895	0.913	0.961	**0.984**	0.982	0.961	0.644
QPA QARR	0.911	0.988	0.899	0.821	0.940	1.000	0.917	0.958	0.548
Rietveld refinement powder	0.867	0.919	0.900	0.808	0.943	**0.978**	0.973	0.962	0.436
Rietveld refinement QARR	0.905	0.923	0.857	0.798	0.827	0.964	0.923	0.964	0.452

E.2 Non-Complex Tasks

Here we report the task- and dataset-specific breakdown of metrics and process scores for the non-complex tasks (Figure 9 and Figure 10, Tables 8 to 12).

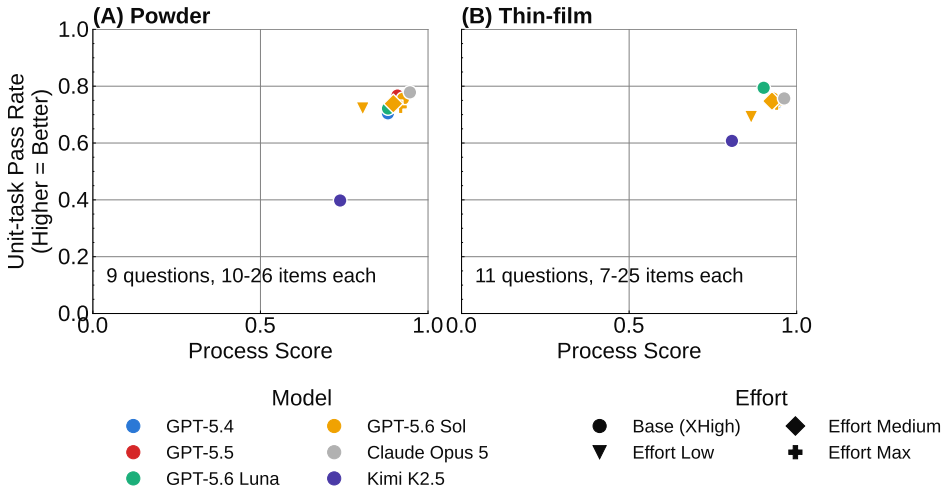


Figure 9: Process scores compared with mean pass rate of perception, unit, and multi-step tasks for each dataset. (A) Powder data. (B) Thin-film data.

Table 8: Non-complex task pass rate (%) by model, per dataset. One column per model, at the matched `xhigh` reasoning effort. Table 10 is the reasoning effort variant of this table. n is the number of gradable items each model was scored on.

Non-complex task	n	GPT-5.4	GPT-5.5	GPT-5.6 Luna	GPT-5.6 Sol	Claude Opus 5	Kimi K2.5
powder							
Peak detection	26	88.5	**100.0**	**100.0**	**100.0**	96.2	80.8
Peak shift	18	**88.9**	**88.9**	**88.9**	**88.9**	**88.9**	55.6
Microstrain	11	27.3	45.5	45.5	54.5	45.5	36.4
Crystallite size	23	47.8	65.2	47.8	**69.6**	65.2	39.1
Lattice parameter	26	73.1	80.8	84.6	88.5	92.3	53.8
Crystal system	26	80.8	88.5	80.8	**96.2**	92.3	23.1
Macrostrain class	18	**94.4**	83.3	77.8	38.9	61.1	11.1
Macrostrain vector	18	44.4	55.6	33.3	55.6	**72.2**	5.6
Preferred orientation	10	**60.0**	40.0	**60.0**	40.0	40.0	30.0
thin-film							
Peak detection	8	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5
Peak shift	8	62.5	62.5	75.0	37.5	62.5	62.5
Amorphous	25	96.0	92.0	92.0	100.0	96.0	92.0
Microstrain	7	85.7	71.4	**100.0**	85.7	71.4	71.4
Crystallite size	7	57.1	42.9	42.9	28.6	71.4	28.6
Vegard's law	15	**86.7**	**86.7**	**86.7**	**86.7**	**86.7**	**86.7**
Lattice parameter	8	75.0	75.0	**100.0**	75.0	**100.0**	50.0
Crystal system	8	**100.0**	**100.0**	87.5	87.5	87.5	25.0
Macrostrain class	7	57.1	**71.4**	**71.4**	**71.4**	**71.4**	42.9
Macrostrain vector	7	0.0	14.3	14.3	**57.1**	14.3	14.3
Preferred orientation	7	42.9	**71.4**	**71.4**	57.1	14.3	42.9

Table 9: Underlying metric per unit task by model (median over answered observations), per dataset. n is the number of observations each model was given. One column per model, at the matched `xhigh` reasoning effort; the ladder is the effort variant of this table, Table 11.

Unit task	Metric	n	GPT-5.4	GPT-5.5	GPT-5.6 Luna	GPT-5.6 Sol	Opus 5	Kimi K2.5
powder								
Peak detection	Peak F_1	26	0.81	**0.83**	**0.83**	0.78	0.76	0.81
Peak shift	$\lvert\Delta 2\theta\rvert$ (deg)	18	0.0084 (17/18)	0.0082 (17/18)	0.0080 (17/18)	0.0124 (17/18)	0.0168 (17/18)	0.0170 (11/18)
Microstrain	MRE (%)	11	71.2	75.0	66.7	**50.0**	62.5	64.7
Crystallite size	MRE (%)	23	51.3	47.4	51.4	45.9	37.5 (22/23)	60.5
Lattice parameter	vol. MRE (%)	26	**0.1**	0.2	0.2	0.2	0.1 (25/26)	5.9 (23/26)
Preferred orientation	MD r (%)	10	16.6	22.3	**11.3**	27.6	23.2 (9/10)	36.4 (8/10)
thin-film								
Peak detection	Peak F_1	8	0.86	0.80	0.86	0.92	**0.96**	0.85
Peak shift	$\lvert\Delta 2\theta\rvert$ (deg)	8	0.0420 (6/8)	0.0475 (7/8)	0.0426 (7/8)	0.0477 (5/8)	0.0470 (7/8)	0.0453 (6/8)
Microstrain	MRE (%)	7	14.9	21.6	17.6	17.5	40.3	33.2 (6/7)
Crystallite size	MRE (%)	7	16.9	24.9	31.4	39.6	**11.2**	24.0 (6/7)
Lattice parameter	vol. MRE (%)	8	0.4	0.2	**0.1**	1.2	**0.1**	0.2 (6/8)
Preferred orientation	residual	7	0.0088	0.0044	**0.0012**	0.0075	0.0361	0.0139 (6/7)

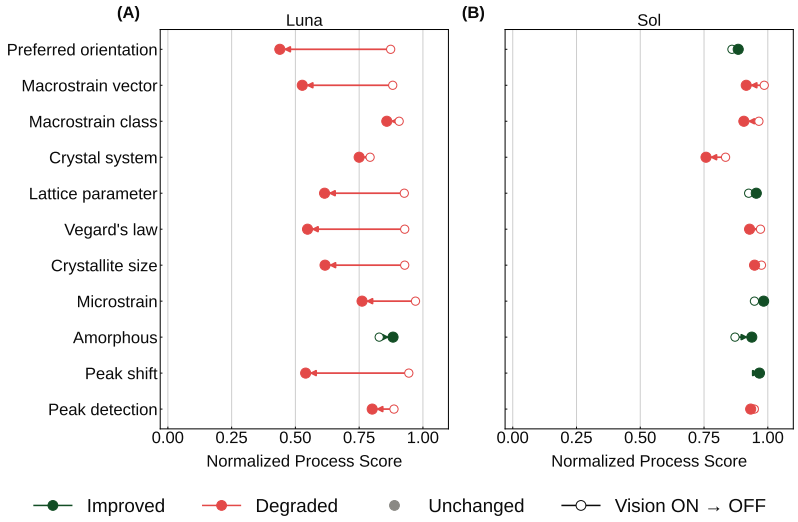


Figure 10: Vision ablation: per-task process scores for (A) GPT-5.6 Luna and (B) GPT-5.6 Sol.

Table 10: Non-complex task pass rate (%). GPT-5.6 Sol is shown at four reasoning efforts; the other models are at `xhigh`. n is the number of gradable items.

Non-complex task	n	GPT-5.4	GPT-5.5	GPT-5.6 Luna	GPT-5.6 Sol				Claude Opus 5	Kimi K2.5
		xhigh	xhigh	xhigh	low	medium	xhigh	max	max	n/a
powder										
Peak detection	26	88.5	**100.0**	**100.0**	100.0	100.0	**100.0**	100.0	96.2	80.8
Peak shift	18	**88.9**	**88.9**	**88.9**	88.9	88.9	**88.9**	88.9	**88.9**	55.6
Microstrain	11	27.3	45.5	45.5	27.3	45.5	54.5	9.1	45.5	36.4
Crystallite size	23	47.8	65.2	47.8	52.2	52.2	**69.6**	73.9	65.2	39.1
Lattice parameter	26	73.1	80.8	84.6	88.5	88.5	88.5	88.5	92.3	53.8
Crystal system	26	80.8	88.5	80.8	92.3	92.3	**96.2**	92.3	92.3	23.1
Macrostrain class	18	**94.4**	83.3	77.8	55.6	72.2	38.9	44.4	61.1	11.1
Macrostrain vector	18	44.4	55.6	33.3	38.9	38.9	55.6	38.9	**72.2**	5.6
Preferred orientation	10	**60.0**	40.0	**60.0**	60.0	40.0	40.0	60.0	40.0	30.0
thin-film										
Peak detection	8	**100.0**	**100.0**	**100.0**	87.5	100.0	**100.0**	100.0	**100.0**	87.5
Peak shift	8	62.5	62.5	75.0	37.5	62.5	37.5	75.0	62.5	62.5
Amorphous	25	96.0	92.0	92.0	100.0	100.0	100.0	100.0	96.0	92.0
Microstrain	7	85.7	71.4	**100.0**	71.4	85.7	85.7	85.7	71.4	71.4
Crystallite size	7	57.1	42.9	42.9	42.9	28.6	28.6	14.3	71.4	28.6
Vegard's law	15	**86.7**	**86.7**	**86.7**	86.7	80.0	**86.7**	86.7	**86.7**	**86.7**
Lattice parameter	8	75.0	75.0	**100.0**	75.0	100.0	75.0	75.0	**100.0**	50.0
Crystal system	8	**100.0**	**100.0**	87.5	87.5	100.0	87.5	75.0	87.5	25.0
Macrostrain class	7	57.1	**71.4**	**71.4**	57.1	71.4	**71.4**	57.1	**71.4**	42.9
Macrostrain vector	7	0.0	14.3	14.3	14.3	0.0	**57.1**	28.6	14.3	14.3
Preferred orientation	7	42.9	**71.4**	**71.4**	14.3	28.6	57.1	42.9	14.3	42.9

Table 11: Non-complex task metric, median over answered items. Lower is better except peak F_1. MRE is the relative error $|\hat{x} - x|/x$ of one observation, as defined in Table 1; MD is March–Dollase. GPT-5.6 Sol is shown at four reasoning efforts; the other models are at `xhigh`. A partial cell (g/n) is not eligible for bold.

Non-complex task	Metric	n	GPT-5.4	GPT-5.5	GPT-5.6 Luna	GPT-5.6 Sol				Claude Opus 5	Kimi K2.5
			xhigh	xhigh	xhigh	low	medium	xhigh	max	max	n/a
powder											
Peak detection	F_1	26	0.81	**0.83**	**0.83**	0.86	0.84	0.78	0.75	0.76	0.81
Peak shift	$\lvert\Delta 2\theta\rvert$	18	0.0084 (17/18)	0.0082 (17/18)	0.0080 (17/18)	0.0108 (17/18)	0.0119 (16/18)	0.0124 (17/18)	0.0100 (16/18)	0.0168 (17/18)	0.0170 (11/18)
Microstrain	MRE (%)	11	71.2	75.0	66.7	75.0	57.1	**50.0**	90.0	62.5	64.7
Crystallite size	MRE (%)	23	51.3	47.4	51.4	49.9	48.8	45.9	43.6	37.5 (22/23)	60.5
Lattice parameter	vol. MRE (%)	26	**0.1**	0.2	0.2	0.3	0.2	0.2	0.2	0.1 (25/26)	5.9 (23/26)
Preferred orientation	MD r (%)	10	16.6	22.3	**11.3**	16.0	20.8	27.6	16.2	23.2 (9/10)	36.4 (8/10)
thin-film											
Peak detection	F_1	8	0.86	0.80	0.86	0.80	0.80	0.92	1.00	**0.96**	0.85
Peak shift	$\lvert\Delta 2\theta\rvert$	8	0.0420 (6/8)	0.0475 (7/8)	0.0426 (7/8)	0.0466 (4/8)	0.0477 (6/8)	0.0477 (5/8)	0.0207 (7/8)	0.0470 (7/8)	0.0453 (6/8)
Microstrain	MRE (%)	7	14.9	21.6	17.6	24.5	17.6	17.5	14.9	40.3	33.2 (6/7)
Crystallite size	MRE (%)	7	16.9	24.9	31.4	26.3	21.1	39.6	43.0	**11.2**	24.0 (6/7)
Lattice parameter	vol. MRE (%)	8	0.4	0.2	**0.1**	0.4	0.1	1.2	0.1	**0.1**	0.2 (6/8)
Preferred orientation	residual	7	0.0088	0.0044	**0.0012**	0.0207	0.0243	0.0075	0.0077	0.0361	0.0139 (6/7)

Table 12: Process scores on the non-complex tasks (0–1). The GPT-5.6 Sol columns are that model at four reasoning efforts; the other models are at `xhigh`. Every column was graded by GPT-5.6 Sol.

Task / rubric	GPT-5.4	GPT-5.5	GPT-5.6 Luna	GPT-5.6 Sol				Claude Opus 5	Kimi K2.5
	xhigh	xhigh	xhigh	low	medium	xhigh	max	max	n/a
Non-complex, powder									
All rubrics	0.880	0.909	0.881	0.805	0.897	0.925	0.918	**0.946**	0.738
Peak detection	0.879	0.940	0.860	0.865	0.912	0.929	0.909	0.942	0.819
Peak shift	0.940	0.964	0.948	0.921	0.976	0.960	0.956	0.937	0.853
Microstrain	0.942	0.974	0.961	0.877	0.981	0.942	0.968	0.955	0.877
Crystallite size	0.904	0.957	0.929	0.941	0.978	0.978	0.963	0.944	0.873
Lattice parameter	0.879	0.948	0.912	0.890	0.953	0.970	0.962	0.951	0.668
Crystal system	0.791	0.654	0.750	0.464	0.665	0.766	0.725	0.948	0.585
Macrostrain class	0.948	0.968	0.901	0.730	0.889	0.968	0.988	0.976	0.726
Macrostrain vector	0.829	0.972	0.873	0.825	0.909	0.984	0.933	0.944	0.675
Preferred orientation	0.857	0.886	0.850	0.814	0.879	0.829	0.957	0.900	0.571
Non-complex, thin-film									
All rubrics	0.936	0.928	0.901	0.864	0.926	0.929	0.940	**0.963**	0.807
Peak detection	0.955	0.955	0.938	0.830	0.946	0.982	0.982	0.964	0.812
Peak shift	0.973	0.982	0.938	0.955	0.973	0.973	1.000	0.991	0.848
Amorphous	0.906	0.869	0.829	0.874	0.846	0.871	0.860	0.991	0.877
Microstrain	0.980	0.990	0.990	0.969	1.000	0.959	1.000	0.969	0.939
Crystallite size	1.000	0.949	0.929	0.980	0.959	0.969	0.959	0.980	0.898
Lattice parameter	0.955	0.929	0.955	0.938	0.991	0.839	0.955	0.991	0.705
Crystal system	0.857	0.893	0.875	0.661	0.929	0.964	0.902	1.000	0.634
Macrostrain class	0.980	0.990	0.918	0.592	0.939	0.959	0.980	1.000	0.786
Macrostrain vector	0.939	0.929	0.898	0.949	0.949	0.990	0.969	0.745	0.735
Preferred orientation	0.898	0.939	0.918	0.878	0.918	0.918	0.990	0.908	0.684
Non-complex, Vegard's law									
All rubrics	0.933	0.957	0.929	0.890	0.895	0.971	0.986	0.981	0.857
Vegard's law	0.933	0.957	0.929	0.890	0.895	0.971	0.986	0.981	0.857

E.3 VARIANCE

To estimate the variance between runs, we run three rollouts of the Rietveld refinement task and the crystal system task with GPT-5.6 Sol driving the harness. Table 13, Table 14, Table 15, Table 16 and Table 17 detail how the metrics change across rollouts.

Table 13: Rietveld refinement, weight-fraction error metrics, models at matched reasoning effort: mean ± s.d. over three independent rollouts.

Model	TVD (%)	MARE
powder		
GPT-5.4	35.29 ± 0.79	0.573 ± 0.023
GPT-5.5	31.45 ± 0.77	0.479 ± 0.019
GPT-5.6 Luna	30.36 ± 0.29	0.520 ± 0.011
GPT-5.6 Sol	30.98 ± 0.72	0.470 ± 0.009
Kimi K2.5	47.67 ± 2.61	0.764 ± 0.035
QARR		
GPT-5.4	10.95 ± 1.39	0.253 ± 0.016
GPT-5.5	6.55 ± 0.11	0.134 ± 0.003
GPT-5.6 Luna	11.13 ± 3.52	0.334 ± 0.113
GPT-5.6 Sol	6.66 ± 0.94	0.143 ± 0.013
Kimi K2.5	37.47 ± 7.82	0.948 ± 0.206

Table 14: Rietveld refinement, weight-fraction error metrics, GPT-5.6 Sol across reasoning efforts: mean ± s.d. over three independent rollouts.

Reasoning effort	TVD (%)	MARE
powder		
low	35.29 ± 0.32	0.583 ± 0.013
medium	32.37 ± 0.66	0.491 ± 0.002
xhigh	30.98 ± 0.72	0.470 ± 0.009
max	31.46 ± 0.40	0.465 ± 0.006
QARR		
low	11.42 ± 3.91	0.301 ± 0.078
medium	8.88 ± 2.30	0.252 ± 0.116
xhigh	6.66 ± 0.94	0.143 ± 0.013
max	7.33 ± 1.64	0.147 ± 0.034

Table 15: Rietveld refinement, phase-identity metrics, GPT-5.6 Sol across reasoning efforts: mean ± s.d. over three independent rollouts.

Reasoning effort	Composition F_1	Exact-ID F_1	Exact-set rate
powder			
low	0.776 ± 0.006	0.582 ± 0.004	0.279 ± 0.000
medium	0.783 ± 0.008	0.595 ± 0.013	**0.304 ± 0.020**
xhigh	0.768 ± 0.008	0.594 ± 0.013	0.269 ± 0.026
max	0.780 ± 0.008	0.588 ± 0.014	0.263 ± 0.015
QARR			
low	0.970 ± 0.028	0.970 ± 0.028	0.889 ± 0.096
medium	0.994 ± 0.010	0.994 ± 0.010	0.972 ± 0.048
xhigh	0.991 ± 0.016	0.991 ± 0.016	0.972 ± 0.048
max	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

Table 16: Crystal-system accuracy (%), GPT-5.6 Sol across reasoning efforts: mean ± s.d. over three independent rollouts.

Reasoning effort	Powder (n=26)	Thin-film (n=8)	ICDD (n=8)
low	93.6 ± 2.2	91.7 ± 7.2	**83.3 ± 7.2**
medium	94.9 ± 2.2	95.8 ± 7.2	**83.3 ± 7.2**
xhigh	96.2 ± 0.0	91.7 ± 7.2	79.2 ± 7.2
max	91.0 ± 2.2	70.8 ± 7.2	79.2 ± 7.2

Table 17: Crystal-system accuracy (%), models at matched reasoning effort: mean ± s.d. over three independent rollouts.

Model	Powder (n=26)	Thin-film (n=8)	ICDD (n=8)
GPT-5.4	61.5 ± 16.8	91.7 ± 7.2	75.0 ± 12.5
GPT-5.5	70.5 ± 15.5	**100.0 ± 0.0**	**79.2 ± 7.2**
GPT-5.6 Luna	85.9 ± 4.4	91.7 ± 7.2	**79.2 ± 7.2**
GPT-5.6 Sol	**96.2 ± 0.0**	91.7 ± 7.2	**79.2 ± 7.2**
Kimi K2.5	25.6 ± 4.4	50.0 ± 21.7	16.7 ± 7.2

F COST

The cost of an item is the list price of the tokens that item's rollout consumed. It is not a Bedrock invoice, and the judge calls that grade the session log are not included.

Each rollout reports four token counts: input that missed the prompt cache, input read from the cache, tokens written into the cache, and output. Output includes reasoning tokens. The cost of one item, in USD, is

$$\left(n_{\text{uncached}}\, p_{\text{in}} + n_{\text{cached}}\, p_{\text{cache}} + n_{\text{write}}\, p_{\text{write}} + n_{\text{out}}\, p_{\text{out}}\right)/10^6,$$

where each p is the model's list price in USD per million tokens, as of 2026-09-25. For the GPT models and Kimi K2.5 the four counts are the cumulative totals in the Codex session log. For Claude Opus 5 they are the usage block on the Claude Code result record. Kimi K2.5 reports no cached tokens and has no Bedrock cache discount, so all of its input is charged at the uncached rate.

Table 18 is the mean of that cost over the items of each cell, on the same runs as the score tables. One item is one rollout: a powder pattern, a plate, a spot, or, on a non-complex task row, one question about one pattern. Table 19 is the same mean for GPT-5.6 Sol at four reasoning efforts. GPT-5.5 reports a cache-hit rate of 4.5% on these items, against 78–97% for the other GPT arms, and its column uses the counts it reported.

Table 18: List-price cost per item, in USD. The last row is the mean over all 722 items, not the mean of the rows. GPT-5.4, GPT-5.5, GPT-5.6 Luna, and GPT-5.6 Sol are at `xhigh` and Claude Opus 5 is at `max`. **Bold** is the lowest cost in the row. Prices are the makers' list rates checked 2026-09-25, with uncached input, cached input, cache writes, and output — reasoning included — priced separately. GPT-5.5 reports a 4.5% cache-hit rate on these items, against 78–97% for the other GPT arms. This column uses that report.

Task	n	GPT-5.4	GPT-5.5	GPT-5.6 Luna	GPT-5.6 Sol	Claude Opus 5	Kimi K2.5
Phase identification							
Powder	123	0.87	3.08	**0.08**	0.87	3.65	0.42
Thin-film, per plate	7	1.59	5.03	**0.11**	1.18	5.58	0.46
Thin-film, per spot	37	0.75	2.82	**0.09**	0.70	3.64	0.35
Quantitative phase analysis							
Powder	104	1.22	7.07	**0.15**	1.71	5.89	0.88
QARR	12	2.00	6.40	**0.20**	1.67	8.89	1.34
Rietveld refinement							
Powder	104	3.24	10.50	**0.27**	2.46	7.55	1.47
QARR	12	2.44	10.59	**0.22**	2.52	9.93	1.49
Non-complex tasks							
Powder	176	1.06	3.90	**0.13**	1.10	7.95	0.39
Thin-film	92	0.86	2.56	**0.07**	0.76	3.22	0.24
ICDD	40	1.50	4.25	**0.11**	1.09	5.90	0.51
Vegard's law	15	1.09	3.02	**0.10**	0.87	3.96	0.30
All tasks	722	1.39	5.10	**0.14**	1.31	5.87	0.64

Table 19: List-price cost per item for GPT-5.6 Sol across reasoning effort, in USD, read the same way as Table 18. The `xhigh` column is the same run as the GPT-5.6 Sol column there. **Bold** is the lowest cost in the row.

		GPT-5.6 Sol reasoning effort			
Task	n	`low`	`medium`	`xhigh`	`max`
Phase identification					
Powder	123	**0.23**	0.41	0.87	1.20
Thin-film, per plate	7	**0.29**	0.52	1.18	1.67
Thin-film, per spot	37	**0.17**	0.35	0.70	1.04
Quantitative phase analysis					
Powder	104	**0.25**	0.53	1.71	2.86
QARR	12	**0.24**	0.62	1.67	2.84
Rietveld refinement					
Powder	104	**0.52**	0.93	2.46	4.06
QARR	12	**0.49**	0.85	2.52	3.63
Non-complex tasks					
Powder	176	**0.18**	0.38	1.10	1.79
Thin-film	92	**0.19**	0.34	0.76	1.16
ICDD	40	**0.17**	0.39	1.09	2.00
Vegard's law	15	**0.20**	0.38	0.87	1.11
All tasks	722	**0.26**	0.49	1.31	2.10