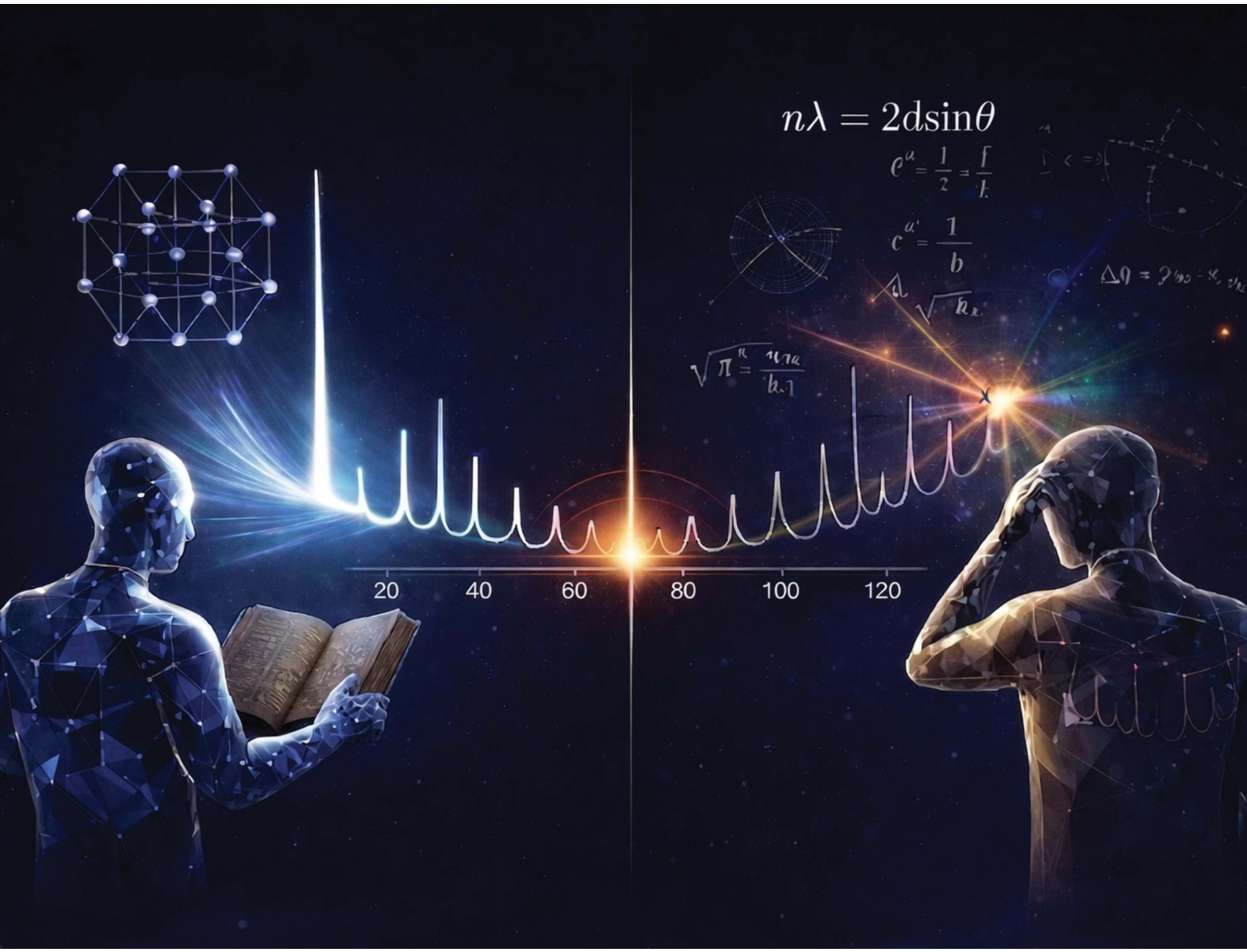
Volume 5
Number 5
May 2026
Pages 1951-2354

Digital Discovery

rsc.li/digitaldiscovery



ISSN 2635-098X

ROYAL SOCIETY OF CHEMISTRY

PAPER
Niaz Abdolrahim *et al.*
OPENXRD: a comprehensive benchmark framework for LLM/MLLM XRD question answering

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 1991

OPENXRD: a comprehensive benchmark framework for LLM/MLLM XRD question answering

Ali Vosoughi,†[a] Ayoub Shahnazari,†[b] Yufeng Xi,[c] Zeliang Zhang,[a] Griffin Hess,[b] Chenliang Xu[a] and Niaz Abdolrahim *[bcd]

We introduce OPENXRD, a comprehensive benchmarking framework for evaluating large language models (LLMs) and multimodal LLMs (MLLMs) in crystallography question answering. The framework measures context assimilation, or how models use fixed, domain-specific supporting information during inference. The framework includes 217 expert-curated X-ray diffraction (XRD) questions covering fundamental to advanced crystallographic concepts, each evaluated under closed-book (without context) and open-book (with context) conditions, where the latter includes concise reference passages generated by GPT-4.5 and refined by crystallography experts. We benchmark 74 state-of-the-art LLMs and MLLMs, including GPT-4, GPT-5, O-series, LLaVA, LLaMA, QWEN, Mistral, and Gemini families, to quantify how different architectures and scales assimilate external knowledge. Results show that mid-sized models (7B–70B parameters) gain the most from contextual materials, while very large models often show saturation or interference and the largest relative gains appear in small and mid-sized models. Expert-reviewed materials provide significantly higher improvements than AI-generated ones even when token counts are matched, confirming that content quality, not quantity, drives performance. OPENXRD offers a reproducible diagnostic benchmark for assessing reasoning, knowledge integration, and guidance sensitivity in scientific domains, and provides a foundation for future multimodal and retrieval-augmented crystallography systems.

Received 21st November 2025
Accepted 9th March 2026

DOI: 10.1039/d5dd00519a

rsc.li/digitaldiscovery

1 Introduction

Crystallography is the scientific discipline concerned with determining the arrangement of atoms and molecules in crystalline solids.[1–3] This structural understanding is crucial for elucidating material properties, such as symmetry, geometry, and physical characteristics.[4,5] This understanding is crucial for progressing materials science, especially in areas like metallurgy, pharmaceuticals, and semiconductor technology.[6–10] Central to crystallography is X-ray diffraction (XRD), a key experimental technique that enables researchers to uncover detailed information about crystal structures by examining how X-rays interact with crystalline lattices.[11–13]

The foundational principle of XRD is Bragg's Law, discovered by Lawrence Bragg in 1912.[14,15] Bragg's law describes how the angle of a diffracted X-ray beam depends on the wavelength of the X-rays and the spacing between atoms and molecules within the material.[16,17] Using XRD, researchers analyze crystal structures to determine important parameters such as atomic arrangements, phase composition, unit cell size, grain size, crystallinity, strain, and lattice defects.[18–25]

Advances in the analysis of large XRD data have demonstrated that conventional deep learning approaches, such as Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs), are powerful tools, particularly for identifying or predicting crystal parameters space group labels,[26–28] lattice constants,[29,30] or phase compositions.[31,32] These models achieve remarkable numerical accuracy, demonstrating their efficacy for crystal classification and structural analysis tasks.

However, despite their success in quantitative predictions, these deep-learning techniques are limited in providing interpretative and explanatory insights into the underlying physics or chemistry of XRD data.[33–35] This lack of interpretability restricts their ability to offer meaningful explanations about the material properties or structural characteristics encoded within the diffraction data. This lack of interpretability restricts the broader application of deep-learning approaches in crystallographic research, where understanding underlying phenomena is essential.[36–38]

In contrast, recent progress in Natural Language Processing (NLP) driven by Large Language Models (LLMs), such as GPT-based architectures, has transformed various computational

[a]*Department of Computer Science, University of Rochester, Rochester, New York 14627, USA*

[b]*Department of Mechanical Engineering, University of Rochester, Rochester, 14627, NY, USA. E-mail: niaz@rochester.edu*

[c]*Material Science Program, University of Rochester, Rochester, 14627, NY, USA*

[d]*Laboratory for Laser Energetics (LLE), University of Rochester, Rochester, 14627, NY, USA*

† These authors contributed equally to this work.

domains.[39–41] LLMs demonstrate exceptional capabilities in open-ended question answering,[42,43] multi-step reasoning,[44,45] domain-specific question answering,[46–48] predicting material properties,[49–51] and extracting information from complex datasets.[52–54] These powerful language models can potentially bridge the interpretability gap faced by traditional deep-learning models when applied to scientific fields, including materials science and crystallography.[55–57]

For instance, Antunes *et al.* introduced CrystaLLM, an autoregressive LLM trained on millions of Crystallographic Information Files (CIFs), which encode detailed crystallographic data, including atomic coordinates, symmetry operations, lattice parameters, and space group information, to generate plausible crystal structures from given chemical compositions, demonstrating the potential of text-based crystal structure generation.[58] In addition, Johansen *et al.* introduced deCIFer, an autoregressive language model on powder XRD data to produce complete crystal structures in CIF format, attaining high match rates between generated structures and experimental diffraction profiles.[59]

Further extending these capabilities, Choudhary introduced DiffractGPT, a generative pretrained transformer that predicts atomic structures directly from XRD patterns, particularly when chemical information is provided.[60] AtomGPT, another transformer-based model by Choudhary, effectively predicts material properties such as formation energies, electronic bandgaps, and superconducting transition temperatures with accuracy comparable to GNNs.[61] Beyond generative tasks, LLMs have also been employed for scientific question-answering and knowledge retrieval in the materials domain. For example, the LLaMP framework combines an LLM with a materials database to fetch and reason over crystallographic data, reducing hallucinations and improving factual accuracy in materials science QA.[53]

In this study, we investigate whether providing domain-specific context generated by a stronger model, such as supporting textual material, can significantly enhance the performance of weaker models or alternative configurations of the same model on specialized XRD-related questions. To rigorously test this hypothesis, we constructed a carefully selected dataset consisting of 217 multiple-choice XRD questions, each reviewed and approved by a domain expert at the PhD level. Each question has only one correct answer among four provided choices, covering a range of topics from fundamental principles to complex scenarios, including issues such as basic structural geometry, unit cell dimensions, and coordination environments, to more complex concepts, including fundamental equations and symmetry analysis. Initially, we evaluate the performance of various large LLMs, including GPT-4.5, GPT-4, O1, O3, and so forth, under closed-book conditions, where models rely solely on their pretrained internal knowledge without external assistance. Our initial results show that GPT-4.5 clearly outperforms other tested models. Based on this outcome, we employ GPT-4.5 to generate approximately one-page textual summaries. These summaries intentionally avoid explicitly stating correct answers yet offer adequate context and guidance to facilitate correct reasoning. We then reassess the performance of the other models under open-book conditions to quantify the improvement when provided with these contextual summaries, aiming to quantify the improvement enabled by this supplementary textual material.

It is important to clarify that OPENXRD is a context-assimilation benchmark rather than a retrieval system, and as such is complementary to Retrieval-Augmented Generation (RAG) rather than competing with it. While RAG is a deployment architecture that couples a retriever with a generator to fetch answer-bearing passages from large corpora at inference time, OPENXRD deliberately removes the retrieval confound by providing fixed, curated, answer-guiding passages that avoid revealing correct answers. This design isolates the language model's ability to integrate external guidance from confounding factors such as retrieval quality, chunking, and ranking. OPENXRD does not perform retrieval and does not aspire to be a deployable QA system; instead, it provides a controlled, reproducible setting to disclose how language models react to guidance—when it helps, when it distracts, and how sensitivity varies with model family, token budget, and content quality. These are properties that end-to-end RAG often obscure because retrieved passages can directly contain the answer. The same evaluation harness can benchmark RAG systems by replacing our oracle helper passages with retrieved chunks from a crystallography corpus, thereby decomposing end-to-end RAG accuracy into retrieval quality and language model assimilation capability. Thus, OPENXRD serves as a diagnostic complement for RAG research, not a competing alternative.

In this sense, our open-book setting can be interpreted as an oracle (Gold-Standard) RAG condition: the model receives a passage that is (i) maximally relevant to the question and (ii) explicitly written to support reasoning without leaking the answer. This design provides an upper bound on what a conventional RAG pipeline could achieve with the same generator if retrieval were perfect. In contrast, standard RAG performance reflects the combined effect of retrieval (corpus coverage, chunking, ranking, and noise) and generation (assimilation). Because OPENXRD fixes the passage, it enables a clean estimate of the generator's assimilation capability that end-to-end RAG accuracy alone cannot disentangle. Practically, for mid-capacity models, improving the quality of the provided evidence can yield substantial gains, whereas high-capacity models can show saturation or even interference when evidence is redundant or stylistically mismatched.

Importantly, OPENXRD does not refine model parameters: all models are evaluated zero-shot and we do not modify or fine-tune any model weights. The only refinement in our pipeline is applied to the supporting passages (AI-generated *versus* expert-reviewed). Accordingly, open-book gains should be interpreted as improved inference-time use of curated guidance (context assimilation), rather than weight-level improvement in the model's underlying reasoning capability (see Section 2.6).

2 Methods

2.1 Dataset selection and curation

To evaluate language model performance in crystallography, we curated a domain-specific multiple-choice question-answering

dataset comprising 217 expert-reviewed items focused exclusively on X-ray diffraction and crystallographic reasoning. To construct this dataset, we partially drew on content from Cullity's introduction to X-ray diffraction.[11] Each question includes a concise prompt, three to four answer options, a single correct answer, a brief explanation providing the scientific rationale, and a associated subtask label that enables fine-grained performance analysis across different crystallographic domains. The dataset encompasses 81 distinct crystallographic subtasks, covering a broad span of domain knowledge including crystal structure fundamentals, diffraction geometry, reciprocal-space concepts, Bragg-law applications, peak indexing, space-group interpretation, unit-cell analysis, scattering physics, microstructure effects, and additional specialized topics (see Fig. 1 and 2).

As illustrated in Fig. 2, the distribution of subtask labels highlights the diversity and complexity of the topics covered in our dataset, ranging from basic definitions to advanced structural phenomena. All questions were reviewed and validated by a domain expert to ensure scientific correctness and comprehensive coverage of essential XRD subtasks.

This annotated dataset serves three primary purposes. First, it provides a benchmarking framework for testing the capability of LLMs to comprehend and apply specialized crystallographic concepts. Second, it enables fine-grained subtask-level analysis, distinguishing model performance across different categories, ranging from fundamental definitions to complex conditions. Third, it supports a comparative study between closed-book and open-book evaluation settings, allowing us to assess the impact of supplementary domain-specific context on model accuracy.

The atomic scattering factor measures how effectively an atom scatters X-rays, and it depends on the atom's electron density and the scattering angle. The atomic scattering factor quantifies how effectively an atom scatters X-rays, depending primarily on its electron density distribution and the scattering angle. At low scattering angles, the atomic scattering factor approaches the total electron count of the atom, but as the angle increases, it decreases due to interference effects within the electron cloud. Consequently, heavier atoms with higher electron counts typically have larger scattering factors at low angles. Additionally, this scattering factor influences the intensity of diffraction peaks, directly affecting how clearly atomic positions can be determined from experimental X-ray diffraction data.

Question

What happens if the path difference between scattered rays is $n\lambda$?

Options

(A) They cancel each other
(B) **They reinforce each other**
(C) They scatter randomly
(D) No interference occurs

Fig. 1 Example question from the OPENXRD dataset in closed-book format. Each question includes a concise prompt, multiple-choice options (3–4 answers), with one correct answer. This illustrates the baseline evaluation condition where models must rely solely on their internal knowledge without external supporting materials.

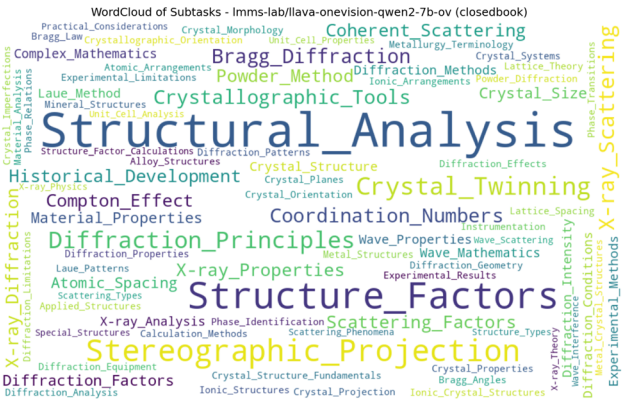


Fig. 2 A word cloud of the subtask labels from our human-curated crystallography dataset. Larger words indicate subtasks with a higher number of questions, illustrating the breadth of topics (*e.g.*, diffraction fundamentals, lattice geometry, advanced structure analysis). This distribution underscores the diversity and domain complexity captured in our QA benchmark, spanning both basic definitions (*e.g.*, counting crystal systems) and intricate reflections (*e.g.*, twin boundaries, space-group anomalies).

2.2 Evaluation framework: closed-book *vs.* open-book

The OPENXRD evaluation framework is designed to assess the reasoning capabilities of language models in crystallography under two distinct conditions: closed-book and open-book modes. In the closed-book mode, the model receives only the question along with its corresponding multiple-choice options. It must rely entirely on its pre-trained internal knowledge to answer, without access to any additional resources or clarifications. This setting evaluates the model's intrinsic understanding of crystallographic principles and its ability to perform domain-specific reasoning based solely on its learned representations. In contrast, the open-book mode supplements the question and answer choices with a brief, domain-relevant textual passage. This supporting material contains essential background information or clarifications pertinent to the question, similar to consulting a short textbook excerpt. The aim is to evaluate the model's ability to integrate external context with its internal knowledge in order to reason more effectively.

Together, these two evaluation settings provide complementary insights. While the closed-book mode tests baseline domain knowledge and internal consistency, the open-book mode examines the model's capacity to synthesize and apply external scientific information. This dual-mode structure is particularly valuable for identifying the limits of a model's independent expertise *versus* its ability to benefit from curated guidance.

We use a multiple-choice format in this foundational release because it provides an unambiguous ground-truth label for each

item, enabling rigorous, automated, and statistically stable comparison across many heterogeneous models under identical closed-book *versus* open-book conditions. In contrast, open-ended outputs typically require human grading or rubric-based evaluation, which introduces evaluator variance and reduces reproducibility at the scale of this study. OPENXRD therefore prioritizes measurement precision for diagnosing context assimilation; open-ended short-answer and multi-step problem-solving evaluations are a valuable complementary direction.

2.3 Supporting textual material generation

2.3.1 AI-generated supporting materials (open-book, without expert review). Due to copyright constraints and the impracticality of systematically scanning large volumes of actual textbooks, GPT-4.5 was used to generate short, supportive textual paragraphs for the open-book evaluations.

The generation process followed three key principles. First, each passage summarized fundamental crystallographic concepts relevant to the specific question. Second, care was taken to avoid directly stating the correct answer, ensuring that models must still reason rather than extract solutions verbatim. Third, the length and content were controlled to maintain clarity and focus, typically resulting in brief paragraphs spanning half to one page.

Prompt template and parameters (reproducible). For each question, we provide GPT-4.5 with the question text and the full set of answer choices without indicating the correct option, and we generate a single answer-guiding passage using fixed decoding settings (temperature = 0.7, top_p = 0.95, max_tokens = 800). The prompt template is: "Given the following question: [QUESTION] and answer choices: [OPTIONS], write a concise explanation of the relevant crystallographic concepts that helps answer the question without revealing the correct answer. Focus on fundamental principles, key definitions, and physical mechanisms; avoid stating or paraphrasing any option as correct. Target length: 500–600 tokens". Passages that (i) explicitly identify the correct option, (ii) meaningfully paraphrase a correct choice in a way that reveals it, or (iii) contain identifiable technical errors are regenerated using the same template with stricter "no answer leakage" instructions. The exact prompt text, parameters, and regeneration checklist are released with the benchmark assets (Data availability).

2.3.2 AI-generated materials with expert review. We engaged three PhD students specializing in crystallography to review and refine the supporting materials. These experts, with four to seven years of research experience in XRD, crystal structure analysis, and materials characterization, were instructed to correct any technical inaccuracies in AI-generated materials, improve explanations with precise domain terminology, ensure comprehensive coverage of relevant concepts, improve clarity while maintaining concision, remove potentially misleading information, and add critical contextual details missing from the original materials. Reviewers followed a standardized quality rubric defined in Section 2.3.3 (accuracy, clarity, completeness, reliability), editing passages to increase these attributes while preserving concision and avoiding answer leakage. For the token-matched ablation in Section 3.5, reviewers also maintained helper length within ±5% of the paired AI-generated passage to eliminate text volume as a confound.

As demonstrated in Fig. 3, the expert-reviewed version provides substantial improvements over the initial AI-generated explanations. Furthermore, Fig. 4 illustrates a representative example from the multiple-choice question-answering dataset, showcasing the original question, corresponding answer choices, and the AI-generated explanation as revised by an expert.

2.3.3 Operational definition of supporting-material quality. To make "information quality" measurable and reproducible, we operationalize helper-text quality using a four-attribute rubric scored on a 0–10 scale: accuracy, clarity, completeness, and reliability. Accuracy reflects correctness of crystallographic terminology, physical mechanisms, and mathematical statements. Clarity reflects whether the explanation is unambiguous and logically structured so that a reader (or model) can follow the reasoning without inference gaps. Completeness reflects whether the passage contains the necessary and sufficient concepts to answer the specific question, including explicit separation of common confounders when relevant (*e.g.*, distinguishing intrinsic angular dependence of the atomic form factor from Debye–Waller effects).

AI-Generated Material

The atomic scattering factor measures how effectively an atom scatters X-rays, and it depends on the atom's electron density and the scattering angle.

Expert-Reviewed Material

The atomic scattering factor quantifies how effectively an atom scatters X-rays, depending primarily on its electron density distribution and the scattering angle. At low scattering angles, the atomic scattering factor approaches the total electron count of the atom, but as the angle increases, it decreases due to interference effects within the electron cloud. Consequently, heavier atoms with higher electron counts typically have larger scattering factors at low angles. Additionally, this scattering factor influences the intensity of diffraction peaks, directly affecting how clearly atomic positions can be determined from experimental X-ray diffraction data.

Fig. 3 Representative example illustrating operational quality attributes (Section 2.3.3). Compared to the AI-generated passage, expert review improves accuracy (mechanistic correctness), clarity (more explicit logical structure), completeness (adds necessary context for the question), and reliability (reduces phrasing that can blur related mechanisms). This qualitative example complements the dataset-level quality scoring reported in Table 6.

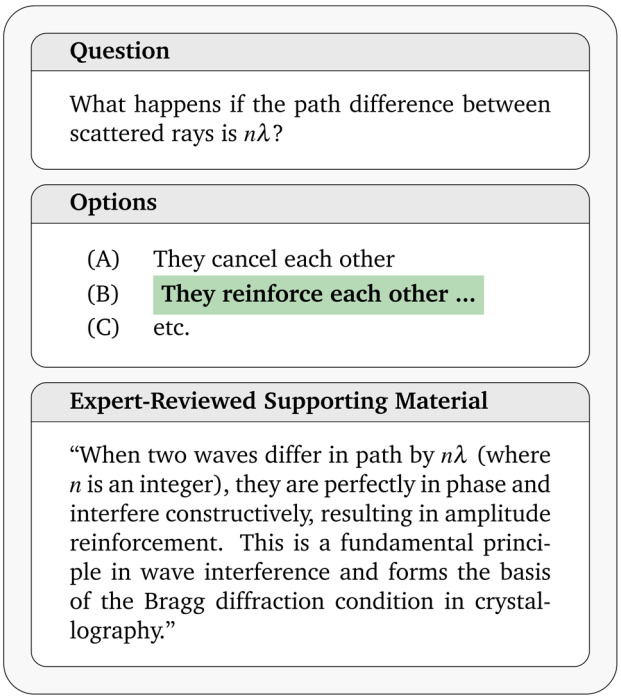


Fig. 4 An open-book mode example with expert-reviewed supporting material that clearly explains how a path difference of $n\lambda$ leads to constructive interference, with additional contextual information about its relevance to crystallography. The correct answer is highlighted in green.

Reliability reflects the extent to which the passage avoids misleading tangents, speculative claims, or mixed mechanisms that can trigger distraction-based errors. During expert review (Section 2.3.2), reviewers edited each passage to improve these four attributes while preserving concision and avoiding answer leakage.

2.4 Input context structuring

We implement a structured input preparation approach *via* a fusion module. For open-book evaluations, we employ a retrieve-then-read style input format (without executing retrieval) where the supporting material precedes the question:

$$x = \mathrm{Format}(x_\mathrm{s}, x_\mathrm{q}, x_\mathrm{o}) \quad (1)$$

where x_s is the supporting textual content, x_q represents the question text, and x_o indicates the provided answer choices. The Format function organizes these components with clear delimiters and instructional guidance. For closed-book evaluations, the supporting material is omitted ($x_\mathrm{s} = \varnothing$), while the remaining formatting is preserved. This structured approach aligns with established retrieval-augmented generation paradigms, wherein external knowledge serves as contextual reference material. Here, the supporting text x_s is oracle-provided (expert-curated and question-specific), so the experiment controls for retrieval variability. In a conventional RAG system, x_s would be produced by a retriever over a large corpus and may include irrelevant or partially relevant chunks. Therefore, OPENXRD should be viewed as benchmarking the generation/assimilation stage under a best-case evidence condition.

Fig. 5 demonstrates the architecture of this open-book QA pipeline. Currently, our implementation is exclusively textual; however, this framework can be extended to multimodal data (such as XRD patterns and crystallographic diagrams) as generative vision models become sufficiently advanced.

2.5 Evaluation metrics

We use several metrics to quantitatively assess model performance across evaluation modes, primarily focusing on accuracy, calculated as:

$$\mathrm{Accuracy} = \frac{|\{\text{correctly answered questions}\}|}{|\{\text{all questions}\}|} \times 100\% \quad (2)$$

For our 217-question benchmark, accuracy is reported both as an aggregated measure across the entire dataset and as a disaggregated measure by specific subtasks. For each model $\mathbf{M}$ and subtask $t \in \mathscr{T}$, accuracy is computed as:

$$\mathrm{Accuracy}_{\mathbf{M},t} = \frac{\left|\{q \in Q_t : \mathbf{M}(q) = y_q\}\right|}{|Q_t|} \times 100\% \quad (3)$$

where Q_t is the set of questions associated with subtask t, y_q is the correct answer, and $\mathbf{M}(q)$ denotes the model's prediction. To assess the efficacy of open-book augmentation, we measure performance improvement Δ as:

$$\Delta_\mathbf{M} = \mathrm{Accuracy}_\mathbf{M}^{\text{Open-Book}} - \mathrm{Accuracy}_\mathbf{M}^{\text{Closed-Book}} \quad (4)$$

Additionally, we quantify the relative improvement gained through expert-reviewed materials compared to AI-generated supporting materials:

$$\Delta_\mathbf{M}^{\text{Expert}} = \mathrm{Accuracy}_\mathbf{M}^{\text{Open-Book,Expert}} - \mathrm{Accuracy}_\mathbf{M}^{\text{Open-Book,AI}} \quad (5)$$

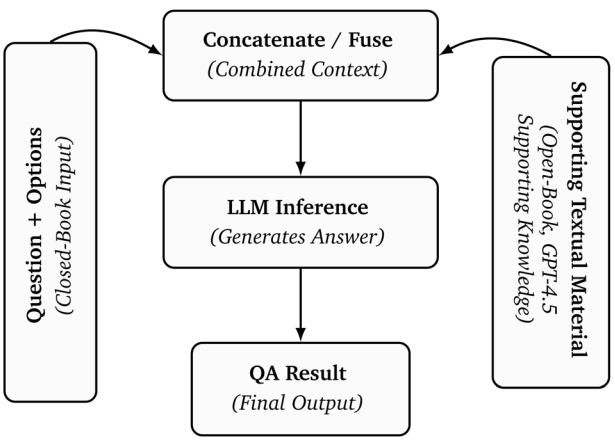


Fig. 5 Illustration of our open-book QA pipeline for crystallography. In closed-book mode, the model sees only the question (left rotated box). In open-book mode, it also receives domain-specific supporting textual material (right rotated box), which is concatenated and fed to the LLM (center pipeline), producing the final QA result.

This comprehensive evaluation framework allows us to explore several aspects. We examine the intrinsic crystallographic knowledge inherent in each model, referred to as closed-book performance. We also assess each model's ability to leverage external knowledge effectively, known as open-book improvement. Additionally, we evaluate the added value provided by expert-reviewed supporting materials, termed expert refinement differential. Furthermore, we analyze model-specific strengths and weaknesses across various crystallographic subtasks.

2.6 Benchmark reproducibility and extension guidelines

OPENXRD's benchmark signal depends on both the questions and the per-question supporting passages used in open-book evaluation. To remove ambiguity and make the benchmark reproducible for future users, we standardize and release the full content pipeline: (i) the helper generation protocol (prompt template, decoding parameters, and regeneration rules that prevent answer leakage and correct technical errors), (ii) the expert review protocol aligned with an explicit quality rubric (Section 2.3.3), and (iii) validation procedures that test whether new supporting materials are answer-guiding without being answer-revealing.

For users evaluating additional models, OPENXRD is run with the same input formatting and evaluation scripts used in this study (Section 2.4–2.5), enabling direct comparison against the baselines reported in Tables 2–6. For users extending the benchmark with new questions, we provide a fixed question schema (prompt, 3–4 options, single correct answer, brief explanation, and subtask label) and require that supporting passages be generated using the released template and then reviewed to improve the four rubric dimensions while avoiding answer leakage. For ablations that compare helper variants, token matching within $\pm 5\%$ is required to control for information quantity (Section 3.5.1).

To validate new or extended supporting materials, we recommend an "answer-guidance" check using held-out models and three conditions: closed-book, explicit-answer materials (upper-bound), and candidate materials. Candidate materials are considered valid if their accuracy lies between the closed-book and explicit-answer conditions and if the relative ordering of model performance is preserved, indicating guidance rather than shortcut leakage. All templates, scripts, and validation utilities are distributed with the benchmark release (Data availability).

2.7 Context assimilation *vs.* parameter fine-tuning

OPENXRD evaluates context assimilation—how models utilize short, answer-guiding passages provided at inference time. We intentionally do not modify model weights through parameter-efficient fine-tuning (PEFT) methods such as LoRA,[62] as this would target a fundamentally different research question: parametric knowledge embedding rather than inference-time guidance utilization.

Several factors motivate this scope decision. First, many of the 74 models in our evaluation pool are API-only or closed-weight systems (GPT-4.5, O3-mini, Claude variants) that cannot be fine-tuned uniformly, making fair PEFT comparisons intractable. Second, our 217-question benchmark is intentionally reserved as a held-out evaluation set to preserve benchmark integrity; fine-tuning on these questions or near-duplicates would introduce data leakage and undermine reproducibility. Third, PEFT optimization conflates parametric adaptation effects with the inference-time assimilation dynamics we aim to isolate.

OPENXRD is complementary to PEFT research rather than competitive with it. Researchers can fine-tune open-weight models on external crystallographic corpora and subsequently use OPENXRD to test whether inference-time guidance remains beneficial, becomes redundant, or introduces interference after domain adaptation. This decomposition of parametric knowledge *versus* contextual reasoning is precisely the diagnostic capability OPENXRD provides. For researchers aiming to perform PEFT studies using our benchmark, we provide the complete dataset, evaluation scripts, and code framework on our project repository. This open-source release enables researchers to design their own PEFT experiments using standard techniques such as LoRA with appropriate data splits to prevent leakage between training and evaluation sets.

3 Results

In this section, we present a detailed evaluation of our OPENXRD benchmark for crystallography, using a curated set of 217 XRD-related questions. Each question can be answered in two ways: closed-book mode, where the model relies solely on its internal knowledge, and open-book mode, where it consults supporting textual material. Our goal is to determine to what extent supporting textual material access bolsters model performance, particularly for advanced domain-specific queries. We conduct our experiments in two phases: first using AI-generated supporting materials provided by GPT-4.5, and then using expert-reviewed versions of these materials to assess whether human domain expertise further enhances performance.

3.1 Setup and baselines

3.1.1 Models compared. We conduct a comprehensive evaluation across 74 state-of-the-art language models and vision-language models spanning multiple architectural families, parameter scales, and specialization domains. This extensive model selection enables systematic analysis of how model architecture, scale, and domain adaptation affect crystallographic reasoning performance under both closed-book and open-book conditions.

OpenAI models: we evaluate 13 OpenAI models across multiple generations and capabilities. The next-generation GPT-5 series includes `gpt-5` and `gpt-5-codex`.[63,64] The reasoning-optimized O-series comprises `o3-mini` (o3-mini-2025-01-31), `o1` (o1-2024-12-17), and `o1-mini`.[65,66] The GPT-4 family includes `gpt-4.5-preview` (gpt-4.5-preview-2025-02-27), `gpt-4o`, `gpt-4o-mini`, `gpt-4-turbo` (gpt-4-turbo-2024-04-

Table 1 Comprehensive "closed-book mode" accuracy results on the 217-item crystallography QA benchmark, including older GPT-4 variants and newly reported models. The rightmost column shows model parameter sizes: B = billions, T = trillions, U = undisclosed. For Mixture-of-Experts (MoE) models, format is total/active parameters. Estimated values marked with ∼, N/A for routing services

Rank	Model (closed-book mode)	Acc. (%)	Correct	Params
1	openai/gpt-5	96.77	210/217	∼300B
2	x-ai/grok-4-fast	96.31	209/217	U
2	openai/gpt-5-codex	96.31	209/217	∼300B
4	google/gemini-2.5-pro	95.39	207/217	U
5	o3-mini	93.55	203/217	U
6	meituan/longcat-flash-chat	93.09	202/217	560B/27B
7	gpt-4.5-preview	92.63	201/217	U
8	anthropic/claude-3.5-sonnet	91.24	198/217	∼175B
9	openai/gpt-4o	90.74	196/216	∼200B
10	dziner-qwen-2.5-72b	90.32	196/217	∼72B
10	perplexity/sonar-pro	90.32	196/217	70B
10	qwen/qwen3-next-80b-a3b-instruct	90.32	196/217	80B/3B
13	deepseek/deepseek-v3.1-terminus	89.86	195/217	671B/37B
13	qwen/qwen-plus	89.86	195/217	U
13	qwen/qwen3-next-80b-a3b-thinking	89.86	195/217	80B/3B
16	deepseek/deepseek-chat	89.40	194/217	671B/37B
16	o1	89.40	194/217	∼200B
16	qwen/qwen-max	89.40	194/217	U
19	anthropic/claude-3-opus	88.94	193/217	U
20	openai/o1-mini	88.02	191/217	∼100B
21	meta-llama/llama-3.1-405b-instruct	87.56	190/217	405B
21	qwen/qwen-2.5-72b-instruct	87.56	190/217	72B
23	mistralai/mistral-large	86.18	187/217	123B
24	amazon/nova-pro-v1	86.11	186/216	∼90B
25	gpt-4-turbo	85.25	185/217	1.8T/280B
26	meta-llama/llama-3-70b-instruct	84.79	184/217	70B
26	meta-llama/llama-3.1-70b-instruct	84.79	184/217	70B
28	dziner-qwen-2.5-coder-32b	83.87	182/217	∼32B
29	openai/gpt-4-0314	83.41	181/217	1.8T/280B
29	gpt-4-turbo-preview	83.41	181/217	1.8T/280B
31	amazon/nova-lite-v1	82.03	178/217	∼20B
32	dziner-qwen-2.5-7b	81.57	177/217	∼7B
32	gpt-4	81.57	177/217	1.8T/280B
34	openai/gpt-4o-mini	81.11	176/217	∼8B
34	openrouter/auto	81.11	176/217	N/A
36	qwen/qwen-2.5-7b-instruct	79.72	173/217	7B
37	google/gemma-2-27b-it	79.26	172/217	27B
38	anthropic/claude-3.5-haiku	77.57	166/214	U
39	mistralai/mixtral-8x22b-instruct	76.81	159/207	141B/39B
40	mistralai/mistral-7b-instruct	75.35	162/215	7B
41	google/gemma-2-9b-it	75.12	163/217	9B
42	amazon/nova-micro-v1	74.65	162/217	∼11B
42	mistralai/mistral-small	74.65	162/217	24B
44	anthropic/claude-3-haiku	74.04	154/208	∼20B
45	openai/gpt-3.5-turbo-16k	72.60	151/208	∼20B
46	mistralai/mixtral-8x7b-instruct	72.33	149/206	47B/13B
47	meta-llama/llama-3-8b-instruct	71.89	156/217	8B
48	mistralai/pixtral-12b	71.83	153/213	12B
49	openai/gpt-3.5-turbo	70.67	147/208	∼20B
50	meta-llama/llama-3.1-8b-instruct	69.12	150/217	8B
51	llava-v1.6-34b	66.82	145/217	34B
52	lmms-lab/llava-onevision-qwen2-7b-si	66.36	144/217	7B
53	lmms-lab/llava-onevision-qwen2-7b-ov-chat	65.90	143/217	7B
54	lmms-lab/llava-onevision-qwen2-7b-ov	65.44	142/217	7B
55	meta-llama/llama-3.2-3b-instruct	64.98	141/217	3B
56	arcee-ai/afm-4.5b	62.21	135/217	4.5B
57	mistralai/mistral-7b-instruct-v0.1	59.50	119/200	7B
58	perplexity/sonar	59.24	125/211	70B
59	undi95/remm-slerp-l2-13b	57.08	121/212	13B
60	llamat-3-chat	57.14	124/217	8B
61	alibaba/tongyi-deepresearch-30b-a3b	55.76	121/217	30B/3B

Table 1 (*Contd.*)

Rank	Model (closed-book mode)	Acc. (%)	Correct	Params
62	gryphe/mythomax-l2-13b	53.77	114/212	13B
63	llava-v1.6-mistral-7b	52.99	115/217	7B
64	llamat-2-chat	50.69	110/217	7B
65	lmms-lab/llava-onevision-qwen2-0.5b-ov	47.47	103/217	0.5B
66	llava-v1.5-13b	46.54	101/217	13B
67	lmms-lab/llava-onevision-qwen2-0.5b-si	46.08	100/217	0.5B
68	qwen/qwen3-coder-flash	43.46	83/191	U
69	qwen/qwen3-coder-plus	22.99	43/187	U
70	honeybee-13b	22.12	48/217	13B
71	qwen/qwen-2.5-coder-32b-instruct	21.03	45/214	32B
72	honeybee-7b	19.35	42/217	7B
73	llava-v1.5-7b	17.97	39/217	7B
74	llamat-2	16.59	36/217	7B

09), `gpt-4-turbo-preview` (gpt-4-preview-0125), `gpt-4-0314` (gpt-4-0613), and the base `gpt-4`.[67–70] We also evaluate earlier-generation models `gpt-3.5-turbo` and `gpt-3.5-turbo-16k`.[71,72] Anthropic Claude models: we assess four Claude variants spanning two generations: `claude-3.5-sonnet`, `claude-3.5-haiku`, `claude-3-opus`, and `claude-3-haiku`.[73–75] Meta LLaMA models: the LLaMA family is represented by six models across three generations and multiple parameter scales: `llama-3.1-405b-instruct` (405B parameters), `llama-3.1-70b-instruct` and `llama-3-70b-instruct` (70B parameters), `llama-3.1-8b-instruct` and `llama-3-8b-instruct` (8B parameters), and `llama-3.2-3b-instruct` (3B parameters).[76,77]

QWEN/Alibaba models: we evaluate 10 models from Alibaba's QWEN ecosystem. General-purpose models include `qwen3-next-80b-a3b-instruct`, `qwen3-next-80b-a3b-thinking` (reasoning-optimized variant), `qwen-2.5-72b-instruct`, `qwen-plus`, `qwen-max`, and `qwen-2.5-7b-instruct`.[78,79] Code-specialized variants include `qwen3-coder-flash`, `qwen3-coder-plus`, and `qwen-2.5-coder-32b-instruct`.[80,81] We also evaluate `tongyi-deepresearch-30b-a3b`, Alibaba's research-focused model.[82] Mistral AI models: seven mistral variants are evaluated: `mistral-large` and `mistral-small` (dense models),[83,84] `mistral-7b-instruct` and `mistral-7b-instruct-v0` (7B base models),[85] the mixture-of-experts models `mixtral-8x22b-instruct` and `mixtral-8x7b-instruct`,[86,87] and the multimodal `pixtral-12b`.[88] DeepSeek models: we evaluate two DeepSeek variants: `deepseek-v3.1-terminus` and `deepseek-chat`.[89] Google models: three google models are assessed: the flagship `gemini-2.5-pro` and the open-weight models `gemma-2-27b-it` and `gemma-2-9b-it`.[90,91] Amazon Nova models: we evaluate Amazon's Nova series across three capability tiers: `nova-pro-v1`, `nova-lite-v1`, and `nova-micro-v1`.[92] X.AI models: we include X.AI's `grok-4-fast` in our evaluation.[93] LLaVA vision-language models: ten LLaVA variants are evaluated, spanning multiple generations and backbone architectures. These include `llava-v1.6-34b` and `llava-v1.6-mistral-7b` from the 1.6 series,[94] `llava-v1.5-13b` and `llava-v1.5-7b` from the 1.5 series,[95] and six LLaVA-OneVision models with QWEN2 backbones: `llava-onevision-qwen2-7b-si`, `llava-onevision-qwen2-7b-ov-chat`, `llava-onevision-qwen2-7b-ov` (7B variants), and `llava-onevision-qwen2-0.5b-ov`, `llava-onevision-qwen2-0.5b-si` (0.5B variants).[96] These models combine vision encoders (ViT) with LLM backbones (Mistral, LLaMA, QWEN) for multimodal understanding. In our text-only evaluation setting, these models process supporting textual materials that may include image captions but not the

Table 2 Model categorization by size for systematic performance analysis. The 74 evaluated models are grouped into three categories to examine scale-dependent effects on external knowledge assimilation in crystallography

Category	Size range	Representative models
Small models	<10B parameters	Mistral-7B, Phi-3-Mini (3.8B), Llama-3-8B, LLaVA-v1.5-7B, QWEN2-0.5B, HoneyBee-7B, LLaMAT-2 (7B), AFM-4.5B
Mid-sized models	10B–70B parameters	Llama-3-70B, LLaVA-v1.6-34B, QWEN-2.5-32B, Mixtral-8 × 7B (∼56B active), Gemma-2-27B, Mistral-Small (24B)
Large models	>70B parameters (or closed-weight API models)	Llama-3.1-405B, QWEN-2.5-72B, dziner-qwen-2.5-72B, GPT-4, GPT-5, GPT-4.5, O3-mini, Claude-3-Opus, Gemini-2.5-Pro, DeepSeek-V3.1

actual images themselves. LLaMAT crystallography-specialized models: we evaluate three domain-adapted LLaMAT variants that have been pre-trained on crystallographic information files (CIF): `llamat-3-chat`, `llamat-2-chat`, and `llamat-2`.[97] These models represent explicit attempts to enhance materials science reasoning through domain-specific pre-training.

Dziner models: we assess three Dziner-QWEN variants spanning different parameter scales: `dziner-qwen-2.5-72b`, `dziner-qwen-2.5-coder-32b`, and `dziner-qwen-2.5-7b`.[98]

Perplexity models: two perplexity variants are evaluated: `sonar-pro` and `sonar`.[99,100] Additional specialized models: our evaluation includes several community-developed and specialized models: Meituan's `longcat-flash-chat`,[101] Arcee-AI's `afm-4.5b`,[102] the community-tuned models `undi95/remm-slerp-l2-13b` and `gryphe/mythomax-l2-13b`,[103,104] the vision-language models `honeybee-13b` and `honeybee-7b`,[105] and the routing system `openrouter/auto`.[106]

This comprehensive model selection, spanning from 0.5B to 405B+ parameters and covering generalist, reasoning-optimized, code-specialized, domain-adapted, and vision-language architectures, enables robust analysis of how different model capabilities interact with domain-specific supporting materials in crystallographic question answering.

Several models were considered but ultimately excluded from our evaluation due to architectural incompatibilities or technical limitations. The `Ether-0` model,[107] designed specifically as a reward model for scoring the quality of scientific reasoning in chemistry rather than as a generative question-answering system, was deemed inappropriate for our benchmark. Reward models are trained using preference learning objectives to rank or score candidate responses, fundamentally differing from generative models that produce answers directly. This architectural mismatch makes `Ether-0` incompatible with our multiple-choice QA evaluation framework, which requires models to generate or select answers rather than evaluate them.

We also encountered technical challenges with several LLaMAT variants, specialized adaptations of LLaMA models pre-trained on crystallographic information files (CIF). While three LLaMAT models (`llamat-3-chat`, `llamat-2-chat`, and `llamat-2`) were successfully evaluated after configuring `max_model_len=2048` to accommodate their tokenization schemes, three additional variants (`llamat-3-base`, `llamat-3-cif`, and `llamat-2-cif`) exhibited incomplete HuggingFace configurations: the base `llamat-3` model lacked proper architecture specifications in `config.json`, while both CIF-specialized variants had malformed model cards with missing tokenizer bindings that prevented inference initialization. Given that the successfully evaluated chat-optimized LLaMAT variants achieved moderate baseline performance (50.69–57.14% accuracy in closed-book mode; see Table 1), the exclusion of these broken variants does not materially impact our conclusions about crystallography-specialized models.

To facilitate systematic performance analysis across different capability levels, we categorize the 74 evaluated models into three size groups based on parameter count and architectural complexity, as summarized in Table 2. This categorization enables structured comparison of how model scale affects the ability to assimilate external domain knowledge in crystallography tasks.

3.1.2 Inference setup. We employ a zero-shot inference paradigm for all models without parameter updating or fine-tuning on our crystallography evaluation set. For each query, we construct a prompt consisting of the question and enumerated options, formatted consistently across all models. No in-context learning examples or demonstrations are provided, requiring models to rely entirely on their parametric knowledge. All foundation models are evaluated using their published weights and architectures; O-family models have documented parameter-efficient training on scientific corpora, while vision-language models in the LLaVA family were primarily trained on general domain text–image pairs without domain-specific adaptation to crystallography.

3.1.3 Experimental progression. Our experimental approach follows two distinct phases: In the first phase, we use GPT-4.5 to generate supporting textual materials for each question, carefully prompting it to provide relevant information without revealing the answer. Following this, in the second phase, we engaged three PhD students specializing in crystallography, with four to seven years of research experience, to review and refine the supporting materials for improved accuracy and pedagogical clarity. This two-phase approach allows us to quantify the value added by human domain expertise in the construction of supporting materials relative to state-of-the-art AI generation.

3.2 Closed-book mode observations

Table 1 presents the closed-book mode accuracy achieved on our 217-question crystallography benchmark, and Fig. 6 illustrates the categorized closed-book performance of the models.

O3-mini performs extremely well, achieving about 93.55% accuracy, even surpassing some GPT-4 variants, possibly reflecting training or optimization details not disclosed; we report the observed performance. However, smaller LLaVA-based models face challenges with advanced reflection extinctions, often incorrectly identifying which plane indices vanish

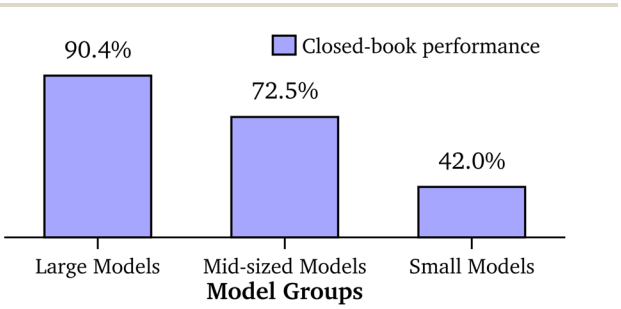


Fig. 6 Comparison of closed-book performance by model size group. Large models (>70B parameters or advanced architectures like GPT-4, GPT-5, O1, O3-mini, $n = 27$); mid-sized models (7B–70B parameters including LLaVA-34B, QWEN2-7B, Mistral-7B, various 8B–13B models, $n = 22$); small models (≤7B parameters including LLaMA-3.2-3B, Honeybee-7B, LLaMAT-2, $n = 25$). Analysis excludes unreliable llava-v1.6-vicuna variants.

Table 3 Comparison of model accuracy in closed-book mode *vs.* open-book mode with AI-generated supporting materials. Δ = (open-book mode) − (closed-book mode)

Model	Closed-book mode (%)	Open-book mode (%)	Δ
openai/gpt-5	96.77	94.93	−1.84
x-ai/grok-4-fast	96.31	95.39	−0.92
openai/gpt-5-codex	96.31	94.47	−1.84
google/gemini-2.5-pro	95.39	88.48	−6.91
o3-mini	93.55	89.40	−4.15
meituan/longcat-flash-chat	93.09	90.32	−2.77
gpt-4.5-preview	92.63	90.32	−2.31
anthropic/claude-3.5-sonnet	91.24	89.40	−1.84
openai/gpt-4o	90.74	87.56	−3.18
dziner-qwen-2.5-72b	90.32	87.10	−3.22
perplexity/sonar-pro	90.32	90.32	+0.00
qwen/qwen3-next-80b-a3b-instruct	90.32	89.40	−0.92
deepseek/deepseek-v3.1-terminus	89.86	90.78	+0.92
qwen/qwen-plus	89.86	90.78	+0.92
qwen/qwen3-next-80b-a3b-thinking	89.86	87.10	−2.76
deepseek/deepseek-chat	89.40	91.24	+1.84
o1	89.40	88.02	−1.38
qwen/qwen-max	89.40	88.02	−1.38
anthropic/claude-3-opus	88.94	88.02	−0.92
openai/o1-mini	88.02	88.48	+0.46
meta-llama/llama-3.1-405b-instruct	87.56	82.95	−4.61
qwen/qwen-2.5-72b-instruct	87.56	86.18	−1.38
mistralai/mistral-large	86.18	83.87	−2.31
amazon/nova-pro-v1	86.11	86.64	+0.53
gpt-4-turbo	85.25	85.71	+0.46
meta-llama/llama-3-70b-instruct	84.79	80.65	−4.14
meta-llama/llama-3.1-70b-instruct	84.79	83.41	−1.38
dziner-qwen-2.5-coder-32b	83.87	84.79	+0.92
openai/gpt-4-0314	83.41	77.42	−5.99
gpt-4-turbo-preview	83.41	82.49	−0.92
amazon/nova-lite-v1	82.03	84.79	+2.76
dziner-qwen-2.5-7b	81.57	79.26	−2.31
gpt-4	81.57	82.03	+0.46
openai/gpt-4o-mini	81.11	82.03	+0.92
openrouter/auto	81.11	81.57	+0.46
qwen/qwen-2.5-7b-instruct	79.72	75.12	−4.60
google/gemma-2-27b-it	79.26	78.80	−0.46
anthropic/claude-3.5-haiku	77.57	88.02	+10.45
mistralai/mixtral-8x22b-instruct	76.81	80.09	+3.28
mistralai/mistral-7b-instruct	75.35	77.42	+2.07
google/gemma-2-9b-it	75.12	77.42	+2.30
amazon/nova-micro-v1	74.65	79.26	+4.61
mistralai/mistral-small	74.65	79.26	+4.61
anthropic/claude-3-haiku	74.04	82.49	+8.45
openai/gpt-3.5-turbo-16k	72.60	71.01	−1.59
mistralai/mixtral-8x7b-instruct	72.33	75.96	+3.63
meta-llama/llama-3-8b-instruct	71.89	73.27	+1.38
mistralai/pixtral-12b	71.83	75.58	+3.75
openai/gpt-3.5-turbo	70.67	71.50	+0.83
meta-llama/llama-3.1-8b-instruct	69.12	75.12	+6.00
llava-v1.6-34b	66.82	72.81	+5.99
lmms-lab/llava-onevision-qwen2-7b-si	66.36	71.89	+5.53
lmms-lab/llava-onevision-qwen2-7b-ov-chat	65.90	72.35	+6.45
lmms-lab/llava-onevision-qwen2-7b-ov	65.44	71.43	+5.99
meta-llama/llama-3.2-3b-instruct	64.98	62.67	−2.31
arcee-ai/afm-4.5b	62.21	64.52	+2.31
mistralai/mistral-7b-instruct-v0.1	59.50	54.03	−5.47
perplexity/sonar	59.24	68.84	+9.60
undi95/remm-slerp-l2-13b	57.08	61.50	+4.42
llamat-3-chat	57.14	30.41	−26.73
alibaba/tongyi-deepresearch-30b-a3b	55.76	80.18	+24.42

Table 3 (*Contd.*)

Model	Closed-book mode (%)	Open-book mode (%)	Δ
gryphe/mythomax-l2-13b	53.77	61.68	+7.91
llava-v1.6-mistral-7b	52.99	58.53	+5.54
llamat-2-chat	50.69	21.13	−29.56
lmms-lab/llava-onevision-qwen2-0.5b-ov	47.47	51.15	+3.68
llava-v1.5-13b	46.54	44.24	−2.30
lmms-lab/llava-onevision-qwen2-0.5b-si	46.08	51.15	+5.07
qwen/qwen3-coder-flash	43.46	84.33	+40.87
qwen/qwen3-coder-plus	22.99	88.94	+65.95
honeybee-13b	22.12	23.04	+0.92
qwen/qwen-2.5-coder-32b-instruct	21.03	27.01	+5.98
honeybee-7b	19.35	23.04	+3.69
llava-v1.5-7b	17.97	23.96	+5.99
llamat-2	16.59	19.82	+3.23

for BCC or confusing zone-axis notation. Despite these issues, nearly all models correctly answer basic factual questions like the number of crystal systems and who discovered X-ray diffraction.

3.3 Open-book mode observations using AI-generated materials (no expert review)

Table 3 lists the closed-book mode *vs.* open-book mode accuracy on our 217-question crystallography benchmark. Here and in subsequent tables, Δ represents the performance improvement defined in eqn (4), with model subscripts omitted as each row corresponds to a different model. We first evaluate open-book mode accuracy by giving each model relevant supporting textual material generated by GPT-4.5.

We observe distinct performance patterns across model scales, architectures, and domain specializations. Remarkably, specialized coder models demonstrate the most dramatic improvements: qwen3-coder-plus gains +65.95% and qwen3-coder-flash gains +40.87% from AI-generated supporting materials, suggesting these models have significant reasoning capacity but minimal crystallographic knowledge. Among general-purpose models, mid-sized architectures (7B–70B parameters) show substantial benefits, with models like llama-3.1-8b-instruct (+6.00%), llava-onevision-qwen2-7b-ov-chat

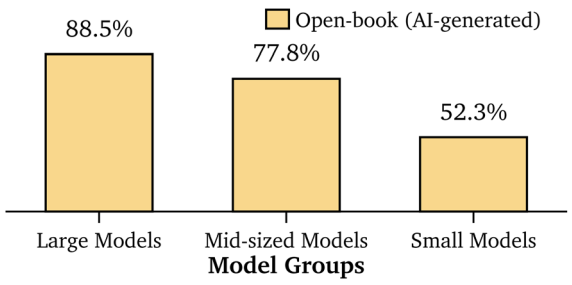


Fig. 7 Open-book performance with AI-generated supporting materials by model size group. Large models (>70B parameters or advanced architectures, $n = 27$); mid-sized models (7B–70B parameters, $n = 22$); small models (≤7B parameters, $n = 25$). Compared to closed-book performance (Fig. 6), large models show slight degradation (−1.9%), while mid-sized models gain +5.3% and small models gain +10.3%, demonstrating that AI-generated context benefits models with knowledge gaps but can interfere with already-capable models.

Question

Which scattering mechanism causes damping of the atomic scattering factor at higher angles?

Options

(A) Compton scattering
(B) **Electron cloud interference**
(C) Thermal vibration
(D) Core-level absorption

Model Responses

LLaVA-v1.6-34B (Closed-book): B
LLaVA-v1.6-34B (Open-book): C

Supporting Material Excerpt

"...The atomic scattering factor decreases with increasing scattering angle due to interference effects within the electron cloud. Additionally, thermal vibrations (Debye-Waller factor) can further dampen the intensity at higher angles, though this is a separate effect from the intrinsic angular dependence..."

Fig. 8 Example of confusion in LLaVA-v1.6-34B due to conflicting supporting material. The model had correct knowledge in closed-book mode but was misled by additional content that mentioned both the correct answer and a distractor as possible factors affecting scattering, albeit in different ways.

Table 4 Comparison of model accuracy in closed-book mode *vs.* open-book mode with expert-reviewed supporting materials. Δ = (open-book mode) − (closed-book mode)

Model	Closed-book mode (%)	Open-book mode (%)	Δ
openai/gpt-5	96.77	93.09	−3.68
x-ai/grok-4-fast	96.31	90.53	−5.78
openai/gpt-5-codex	96.31	92.17	−4.14
google/gemini-2.5-pro	95.39	92.63	−2.76
o3-mini	93.55	89.86	−3.69
meituan/longcat-flash-chat	93.09	89.86	−3.23
gpt-4.5-preview	92.63	89.40	−3.23
anthropic/claude-3.5-sonnet	91.24	89.40	−1.84
openai/gpt-4o	90.74	87.10	−3.64
dziner-qwen-2.5-72b	90.32	86.18	−4.14
qwen/qwen3-next-80b-a3b-instruct	90.32	89.40	−0.92
perplexity/sonar-pro	90.32	88.02	−2.30
deepseek/deepseek-v3.1-terminus	89.86	90.78	+0.92
qwen/qwen-plus	89.86	89.40	−0.46
qwen/qwen3-next-80b-a3b-thinking	89.86	87.10	−2.76
deepseek/deepseek-chat	89.40	90.32	+0.92
qwen/qwen-max	89.40	89.40	+0.00
o1	89.40	88.94	−0.46
anthropic/claude-3-opus	88.94	88.94	+0.00
openai/o1-mini	88.02	88.02	+0.00
meta-llama/llama-3.1-405b-instruct	87.56	84.33	−3.23
qwen/qwen-2.5-72b-instruct	87.56	85.71	−1.85
mistralai/mistral-large	86.18	86.18	+0.00
amazon/nova-pro-v1	86.11	86.64	+0.53
gpt-4-turbo	85.25	85.71	+0.46
meta-llama/llama-3-70b-instruct	84.79	82.95	−1.84
meta-llama/llama-3.1-70b-instruct	84.79	85.19	+0.40
dziner-qwen-2.5-coder-32b	83.87	77.42	−6.45
openai/gpt-4-0314	83.41	82.49	−0.92
gpt-4-turbo-preview	83.41	84.79	+1.38
amazon/nova-lite-v1	82.03	85.71	+3.68
dziner-qwen-2.5-7b	81.57	79.26	−2.31
gpt-4	81.57	82.49	+0.92
openai/gpt-4o-mini	81.11	81.11	+0.00
openrouter/auto	81.11	81.57	+0.46
qwen/qwen-2.5-7b-instruct	79.72	78.34	−1.38
google/gemma-2-27b-it	79.26	79.72	+0.46
anthropic/claude-3.5-haiku	77.57	85.92	+8.35
mistralai/mixtral-8x22b-instruct	76.81	83.89	+7.08
mistralai/mistral-7b-instruct	75.35	80.65	+5.30
google/gemma-2-9b-it	75.12	81.57	+6.45
amazon/nova-micro-v1	74.65	80.65	+6.00
mistralai/mistral-small	74.65	83.41	+8.76
anthropic/claude-3-haiku	74.04	83.33	+9.29
openai/gpt-3.5-turbo-16k	72.60	75.85	+3.25
mistralai/mixtral-8x7b-instruct	72.33	77.14	+4.81
meta-llama/llama-3-8b-instruct	71.89	75.58	+3.69
mistralai/pixtral-12b	71.83	77.88	+6.05
openai/gpt-3.5-turbo	70.67	76.33	+5.66
meta-llama/llama-3.1-8b-instruct	69.12	78.34	+9.22
llava-v1.6-34b	66.82	78.34	+11.52
lmms-lab/llava-onevision-qwen2-7b-si	66.36	75.12	+8.76
lmms-lab/llava-onevision-qwen2-7b-ov-chat	65.90	76.04	+10.14
lmms-lab/llava-onevision-qwen2-7b-ov	65.44	74.19	+8.75
meta-llama/llama-3.2-3b-instruct	64.98	65.44	+0.46
arcee-ai/afm-4.5b	62.21	67.28	+5.07
mistralai/mistral-7b-instruct-v0.1	59.50	62.44	+2.94
perplexity/sonar	59.24	67.14	+7.90
undi95/remm-slerp-l2-13b	57.08	65.12	+8.04
llamat-3-chat	57.14	31.80	−25.34
alibaba/tongyi-deepresearch-30b-a3b	55.76	86.64	+30.88

Table 4 (*Contd.*)

Model	Closed-book mode (%)	Open-book mode (%)	Δ
gryphe/mythomax-l2-13b	53.77	66.51	+12.74
llava-v1.6-mistral-7b	52.99	64.06	+11.07
llamat-2-chat	50.69	16.13	−34.56
lmms-lab/llava-onevision-qwen2-0.5b-ov	47.47	54.38	+6.91
llava-v1.5-13b	46.54	48.39	+1.85
lmms-lab/llava-onevision-qwen2-0.5b-si	46.08	53.91	+7.83
qwen/qwen3-coder-flash	43.46	85.71	+42.25
qwen/qwen3-coder-plus	22.99	88.94	+65.95
honeybee-13b	22.12	21.20	−0.92
qwen/qwen-2.5-coder-32b-instruct	21.03	30.77	+9.74
honeybee-7b	19.35	17.51	−1.84
llava-v1.5-7b	17.97	28.57	+10.60
llamat-2	16.59	19.82	+3.23

(+6.45%), anthropic/claude-3-haiku (+8.45%), and llava-v1.6-34b (+5.99%) gaining meaningfully from domain-specific context. Larger models (70B+) such as llama-3.1-405b-instruct (−4.61%), llama-3-70b-instruct (−4.14%), and qwen-2.5-72b-instruct (−1.38%) show minimal improvement or slight degradation, while very large frontier models like o3-mini see only minor gains (−4.15%), suggesting they already possess substantial internal crystallographic knowledge.

Notably, multiple frontier models show performance degradation with AI-generated supporting materials: GPT-4.5-preview (−2.31%), Claude-3.5-Sonnet (−1.84%), dziner-qwen-2.5-72b (−3.22%), and mistral-large (−2.31%). This counterintuitive pattern reveals an important failure mode: for models with comprehensive pre-trained knowledge, additional context, even when domain-relevant, can introduce interference rather than assistance.

Fig. 8 shows a concrete example where supporting materials have caused confusion, directly illustrating this interference mechanism. In this case, LLaVA-v1.6-34B answered correctly in closed-book mode (Option B: Electron cloud interference) but was misled in open-book mode to select Option C (Thermal vibration). The supporting material discussed both phenomena in close proximity: "The atomic scattering factor decreases with increasing scattering angle due to interference effects within the electron cloud. Additionally, thermal vibrations (Debye–Waller factor) can further dampen the intensity…" This juxtaposition caused the model to conflate two distinct mechanisms, the intrinsic angular dependence (correct answer) *versus* temperature-dependent effects (distractor), demonstrating that the failure mode is primarily conceptual interference and distraction from related concepts rather than simple information overload or redundancy.

Our detailed analysis of the AI-generated supporting materials revealed several limitations: occasional technical inaccuracies in specialized crystallographic terminology, inadequate depth of explanation for particularly complex phenomena, information presented at times being too general to help with specific question nuances, and in some cases, the supporting materials were tangential to the precise knowledge needed for the question. Fig. 7 summarizes these patterns by model size group, showing that mid-sized and small models benefit substantially from AI-generated materials (+5.3% and +10.3% respectively), while large models experience slight degradation (−1.9%).

These observations led us to hypothesize that expert-reviewed materials might yield greater performance improvements, particularly for mid-sized models with significant capacity but incomplete domain knowledge.

3.4 Open-book mode observations using AI-generated materials with expert review

Building on the findings from the initial open-book evaluations, we introduced expert review to improve the quality and effectiveness of the supporting materials. Three PhD-level crystallography experts were engaged to refine the AI-generated materials.

Table 4 compares model accuracy in closed-book and open-book modes using expert-reviewed supporting materials across our 217-question crystallography benchmark. The results show substantially improved performance, particularly among LLaVA-based models.

Remarkably, specialized coder models show the most dramatic gains, with qwen3-coder-plus (+65.95%) and qwen3-coder-flash (+42.25%) achieving exceptional improvements from very low baselines. Among more general-purpose vision-language models, the largest accuracy gains are observed in LLaVA-v1.6-34B (+11.52%) and LLaVA-v1.6-mistral-7B (+11.07%), highlighting the effectiveness of expert-curated guidance in boosting model reasoning.

Several other mid-sized models, such as LLaVA-onevision-QWEN2-7B-ov-chat (+10.14%) and LLaVA-onevision-QWEN2-7B-ov (+8.75%), also demonstrate notable improvements. Fig. 9 illustrates these gains aggregated by model size group, showing that expert curation provides additional improvements over AI-generated materials, particularly for mid-sized (+1.7%) and small models (+2.5%), while large models maintain stable performance.

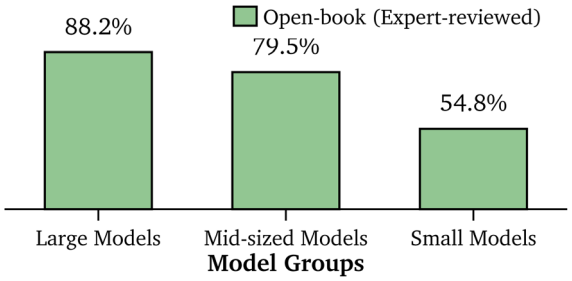


Fig. 9 Open-book performance with expert-reviewed supporting materials by model size group. Large models (>70B parameters or advanced architectures, $n = 27$); mid-sized models (7B–70B parameters, $n = 22$); small models (≤7B parameters, $n = 25$). Compared to AI-generated materials (Fig. 7), expert curation provides additional gains for mid-sized (+1.7%) and small models (+2.5%), while large models remain stable, demonstrating that content quality improvements benefit models with reasoning capacity but incomplete domain knowledge. Relative to the mid-sized closed-book baseline in Fig. 6 (72.5%), the expert-reviewed open-book condition reaches 79.5%.

From an end-user perspective, the mid-sized group's mean accuracy increases from 72.5% in closed-book mode (Fig. 6) to 79.5% when paired with expert-reviewed passages (Fig. 9). This 7.0 percentage-point gain reduces the error rate from 27.5% to 20.5% (≈25% fewer wrong answers), which corresponds to roughly 15 additional correct responses on our 217-question benchmark. In interactive use, this is approximately a shift from ∼3 incorrect answers per 10 questions to ∼2, which can reduce how often a user encounters a disruptive failure and needs to re-check or re-ask. However, because ∼1 in 5 responses remain incorrect, and because mathematically intensive subtasks show little improvement even with high-quality context (Section 3.6), we interpret this as a meaningful usability improvement for decision-support and explanation, not sufficient reliability for fully automated crystallographic decisions.

A notable failure mode identified in our analysis is catastrophic degradation in domain-specialized models, which experience substantial performance drops despite receiving expert-reviewed supporting materials. As summarized in Tables 3 and 4, the LLaMAT family, pretrained on materials-science literature and domain-focused instruction datasets, shows the most severe declines. LLaMAT-3-chat falls from 57.14% (closed-book) to 31.80% (open-book with expert review), a −25.34% drop, while LLaMAT-2-chat decreases from 50.69% to 16.13%, a −34.56% degradation that renders the model nearly unusable in open-book mode. Only the base llamat-2 model shows a modest improvement (+3.23%), though its low absolute accuracy indicates fundamental capacity limitations regardless of external support. These results demonstrate that domain specialization alone does not guarantee effective external-knowledge integration.

Similar patterns emerge in other domain-focused families. The HoneyBee models, designed primarily for scientific figure interpretation, show limited transferability to text-only crystallographic reasoning. HoneyBee-13B drops from 22.12% to 21.20% (−0.92%), and HoneyBee-7B from 19.35% to 17.51% (−1.84%), suggesting that pretraining on visual-scientific corpora does not enhance performance on purely textual tasks, and that their small parameter scale (7B–13B) restricts knowledge retention. The dZiner family, consisting of QWEN variants fine-tuned on materials-science corpora, exhibits a similar sensitivity to interference. While dZiner-QWEN-2.5-72B attains strong closed-book accuracy (90.32%), it drops to 86.18% (−4.14%) with expert-reviewed materials. The medium-sized dZiner-QWEN-2.5-coder-32B declines by −6.45%, and the smaller dZiner-QWEN-2.5-7B by −2.31%. These negative open-book effects mirror degradation trends observed in other high-capacity models, indicating that domain-specific pretraining raises baseline performance but simultaneously increases vulnerability to information interference.

Across all three families, the underlying mechanism is representational rigidity introduced during domain adaptation. These models internalize crystallographic and materials-science concepts in narrowly defined textual, visual, or instructional formats. When exposed to expert-reviewed passages that express these concepts using more pedagogical, descriptive, or stylistically diverse language, they encounter distributional mismatch between their internal priors and the external context. Even when the supplemental information is factually correct, this mismatch triggers interference rather than assistance, leading to performance degradation. This reveals a fundamental trade-off in domain specialization: while narrow, domain-focused pretraining can strengthen closed-book accuracy, it reduces flexibility in assimilating new or differently framed knowledge sources.

Interestingly, the knowledge-interference failure pattern extends systematically across top-tier models, as explained in Section 3.3. As summarized in Table 4, in the transition from the closed-book setting to the open-book mode with expert-reviewed supporting materials, several frontier systems show measurable degradation despite their high baseline accuracy. GPT-4.5-preview declines from 92.63% to 89.40% (−3.23 percentage points), while other state-of-the-art models exhibit comparable or greater decreases: openai/gpt-5 (−3.68%), x-ai/grok-4-fast (−5.78%), openai/gpt-5-codex (−4.14%), and O3-mini (−3.69%). Importantly, all of these systems exceed 90% closed-book accuracy, confirming that they already possess extensive crystallographic knowledge. The consistency of this degradation across diverse architectures (spanning distinct training paradigms, model families, and organizations) indicates that knowledge interference is an intrinsic characteristic of high-capacity language models rather than a model-specific artifact. These models appear highly sensitive to redundant or overlapping external information: when provided with additional context that reformulates or reiterates knowledge already internalized, they can misallocate attention, resulting in subtle yet systematic declines in performance even when the supplemental material is accurate and expert-curated.

3.5 Ablation study: content quality *vs.* information quantity

To rigorously test whether performance gains stem from information quantity or expert curation quality, we conducted a comprehensive token-matched comparison between AI-

Table 5 Token count statistics for questions and helper texts. Expert-reviewed materials maintain virtually identical length to AI-generated materials (mean difference: 2.9 tokens, 0.51%), enabling quality-controlled comparison

Metric	Questions	AI-generated	Expert-reviewed
Mean ± SD	50.1 ± 15.2	566.0 ± 79.2	568.9 ± 82.7
Median [IQR]	49.0 [38.0–60.0]	558.0 [508.0–605.0]	558.0 [513.0–610.0]
Range	[21–98]	[378–919]	[378–919]
Total	10 882	122 825	123 457
Difference	—	—	**+2.9 tokens (+0.51%)**
Correlation	—	—	$r = 0.948$

Table 6 LLM-based helper quality scores (0–10) across 217 AI/expert passage pairs. Scores quantify the operational definition of quality in Section 2.3.3 (accuracy, clarity, completeness, reliability). Expert-reviewed helpers achieve higher scores across all attributes and in the overall mean

Metric	AI helper	Human helper	Δ
Accuracy	7.43	9.25	+1.82
Clarity	7.53	8.91	+1.38
Completeness	7.40	9.20	+1.81
Reliability	7.30	9.17	+1.87
Overall mean	7.41	9.13	+1.72

generated and expert-reviewed supporting materials. Tokens are the fundamental units of text that language models process—roughly corresponding to words or sub-word pieces (*e.g.*, "crystallography" might be split into "crystal" + "lography"). By controlling token count, we ensure that both material types contain equivalent amounts of text, allowing us to isolate content quality effects from simple volume differences. This ablation directly addresses the question: do expert improvements reflect better content or simply more text?

3.5.1 Token-matched experimental design. To isolate the effects of content quality from those of information quantity, we implemented a token-matched ablation study in which every AI-generated supporting passage was paired with its expert-reviewed counterpart of nearly identical length. This design ensured that both sets of materials provided the same textual volume, allowing any observed performance differences to be directly attributed to qualitative factors such as accuracy, relevance, and pedagogical clarity.

As summarized in Table 5, the mean token count was 566.0 ± 79.2 for AI-generated materials and 568.9 ± 82.7 for expert-reviewed materials, representing a negligible difference of only 2.9 tokens (0.51% change). Moreover, the high correlation ($r = 0.948$) between AI- and expert-reviewed helper lengths across all 217 questions, along with their nearly identical median and interquartile ranges (558.0 [508.0–605.0] *vs.* 558.0 [513.0–610.0]), confirms that the two distributions are tightly matched.

This strict token control eliminates text volume as a potential confounding factor and establishes a high-fidelity basis for quality-driven comparison. In other words, any measurable performance improvements can be confidently attributed to the qualitative enhancements introduced through expert curation, rather than to variations in the amount of information provided. This methodological precision is critical for disentangling the true impact of expert refinement from superficial gains that might otherwise arise from longer or more verbose passages.

3.5.2 LLM-based helper quality scoring. LLM-based quality scoring of helper pairs. To quantitatively validate that expert review improves helper-text quality independently of length, we scored all 217 AI/expert helper pairs using the operational rubric in Section 2.3.3 (accuracy, clarity, completeness, reliability; 0–10). We used an external LLM judge (GPT-4o-mini *via* OpenRouter) to assign rubric scores to each passage, then averaged across the dataset. As shown in Table 6, expert-reviewed helpers score higher on every attribute, with the largest gains in reliability and accuracy, yielding a +1.72 improvement in the overall mean score.

3.5.3 Differential performance under token-matched conditions. Table 7 presents a direct comparison of performance gains (Δ) between token-matched AI-generated and expert-reviewed supporting materials across all 74 evaluated models. Despite identical token counts, substantial performance differences emerge, revealing the effect of content quality independent of text volume.

Most models show higher accuracy with expert-reviewed materials, confirming that qualitative improvements, such as enhanced conceptual clarity, precise terminology, and improved contextual alignment, produce measurable performance gains even under identical token counts. The strongest relative improvements occur in small models (<7B parameters), which benefit most from the additional clarity and precision introduced by expert curation. For instance, LLaVA-v1.6-mistral-7B improves from +5.54% to +11.07% (+5.53% differential), and mistralai/mistral-7b-instruct-v0.1 exhibits the largest quality-driven transformation (+8.41%), shifting from performance degradation (−5.47%) with AI-generated text to a measurable gain (+2.94%) with expert-reviewed material, under identical token budgets.

Mid-sized models (10–70B) also show clear but smaller improvements, exemplified by LLaVA-v1.6-34B (+5.99% to +11.52%, +5.53% differential). These models possess enough reasoning capacity to leverage higher-quality information but already contain partial crystallographic knowledge, so their absolute gains are moderate.

Conversely, large models (>70B), such as GPT-4.5-preview, O3-mini, and O1, exhibit minimal or slightly negative differences, reflecting a saturation of internal knowledge representations.

Table 7 Direct comparison of performance gains between AI-generated supporting materials with and without expert review

Model	AI-generated Δ (%)	Expert-reviewed Δ (%)	Difference
openai/gpt-5	−1.84	−3.68	−1.84
x-ai/grok-4-fast	−0.92	−5.78	−4.86
openai/gpt-5-codex	−1.84	−4.14	−2.30
google/gemini-2.5-pro	−6.91	−2.76	+4.15
o3-mini	−4.15	−3.69	+0.46
meituan/longcat-flash-chat	−2.77	−3.23	−0.46
gpt-4.5-preview	−2.31	−3.23	−0.92
anthropic/claude-3.5-sonnet	−1.84	−1.84	+0.00
openai/gpt-4o	−3.18	−3.64	−0.46
dziner-qwen-2.5-72b	−3.22	−4.14	−0.92
qwen/qwen3-next-80b-a3b-instruct	−0.92	−0.92	+0.00
perplexity/sonar-pro	+0.00	−2.30	−2.30
deepseek/deepseek-v3.1-terminus	+0.92	+0.92	+0.00
qwen/qwen-plus	+0.92	−0.46	−1.38
qwen/qwen3-next-80b-a3b-thinking	−2.76	−2.76	+0.00
deepseek/deepseek-chat	+1.84	+0.92	−0.92
qwen/qwen-max	−1.38	+0.00	+1.38
o1	−1.38	−0.46	+0.92
anthropic/claude-3-opus	−0.92	+0.00	+0.92
openai/o1-mini	+0.46	+0.00	−0.46
meta-llama/llama-3.1-405b-instruct	−4.61	−3.23	+1.38
qwen/qwen-2.5-72b-instruct	−1.38	−1.85	−0.47
mistralai/mistral-large	−2.31	+0.00	+2.31
amazon/nova-pro-v1	+0.53	+0.53	+0.00
gpt-4-turbo	+0.46	+0.46	+0.00
meta-llama/llama-3-70b-instruct	−4.14	−1.84	+2.30
meta-llama/llama-3.1-70b-instruct	−1.38	+0.40	+1.78
dziner-qwen-2.5-coder-32b	+0.92	−6.45	−7.37
openai/gpt-4-0314	−5.99	−0.92	+5.07
gpt-4-turbo-preview	−0.92	+1.38	+2.30
amazon/nova-lite-v1	+2.76	+3.68	+0.92
dziner-qwen-2.5-7b	−2.31	−2.31	+0.00
gpt-4	+0.46	+0.92	+0.46
openai/gpt-4o-mini	+0.92	+0.00	−0.92
openrouter/auto	+0.46	+0.46	+0.00
qwen/qwen-2.5-7b-instruct	−4.60	−1.38	+3.22
google/gemma-2-27b-it	−0.46	+0.46	+0.92
anthropic/claude-3.5-haiku	+10.45	+8.35	−2.10
mistralai/mixtral-8x22b-instruct	+3.28	+7.08	+3.80
mistralai/mistral-7b-instruct	+2.07	+5.30	+3.23
google/gemma-2-9b-it	+2.30	+6.45	+4.15
amazon/nova-micro-v1	+4.61	+6.00	+1.39
mistralai/mistral-small	+4.61	+8.76	+4.15
anthropic/claude-3-haiku	+8.45	+9.29	+0.84
openai/gpt-3.5-turbo-16k	−1.59	+3.25	+4.84
mistralai/mixtral-8x7b-instruct	+3.63	+4.81	+1.18
meta-llama/llama-3-8b-instruct	+1.38	+3.69	+2.31
mistralai/pixtral-12b	+3.75	+6.05	+2.30
openai/gpt-3.5-turbo	+0.83	+5.66	+4.83
meta-llama/llama-3.1-8b-instruct	+6.00	+9.22	+3.22
llava-v1.6-34b	+5.99	+11.52	+5.53
lmms-lab/llava-onevision-qwen2-7b-si	+5.53	+8.76	+3.23
lmms-lab/llava-onevision-qwen2-7b-ov-chat	+6.45	+10.14	+3.69
lmms-lab/llava-onevision-qwen2-7b-ov	+5.99	+8.75	+2.76
meta-llama/llama-3.2-3b-instruct	−2.31	+0.46	+2.77
arcee-ai/afm-4.5b	+2.31	+5.07	+2.76
mistralai/mistral-7b-instruct-v0.1	−5.47	+2.94	+8.41
perplexity/sonar	+9.60	+7.90	−1.70
undi95/remm-slerp-l2-13b	+4.42	+8.04	+3.62
llamat-3-chat	−26.73	−25.34	+1.39
alibaba/tongyi-deepresearch-30b-a3b	+24.42	+30.88	+6.46
gryphe/mythomax-l2-13b	+7.91	+12.74	+4.83

Table 7 (*Contd.*)

Model	AI-generated Δ (%)	Expert-reviewed Δ (%)	Difference
llava-v1.6-mistral-7b	+5.54	+11.07	+5.53
llamat-2-chat	−29.56	−34.56	−5.00
lmms-lab/llava-onevision-qwen2-0.5b-ov	+3.68	+6.91	+3.23
llava-v1.5-13b	−2.30	+1.85	+4.15
lmms-lab/llava-onevision-qwen2-0.5b-si	+5.07	+7.83	+2.76
qwen/qwen3-coder-flash	+40.87	+42.25	+1.38
qwen/qwen3-coder-plus	+65.95	+65.95	+0.00
honeybee-13b	+0.92	−0.92	−1.84
qwen/qwen-2.5-coder-32b-instruct	+5.98	+9.74	+3.76
honeybee-7b	+3.69	−1.84	−5.53
llava-v1.5-7b	+5.99	+10.60	+4.61
llamat-2	+3.23	+3.23	+0.00

When most relevant crystallographic concepts are already encoded within the model parameters, additional context contributes little new information and can occasionally disrupt attention alignment or introduce representational interference.

Overall, these results confirm that content quality, not text length, governs open-book performance. Expert review yields the greatest relative benefits for small and mid-sized models, while large models approach a ceiling where additional context produces diminishing or even adverse effects.

Fig. 10 summarizes the overall impact of supporting material quality across model size categories. Small models (<7B parameters) achieve the largest relative improvement from expert-reviewed materials, increasing their accuracy by +8.52 percentage points compared to +6.18% with AI-generated content. Mid-sized models (7B–70B) also benefit noticeably, showing a +4.44% gain *versus* +2.71% with AI-generated materials. In contrast, large models (>70B) display minimal or slightly negative changes with either material type, indicating that their extensive internal knowledge already limits the marginal benefit of added context.

This trend can be explained by differences in knowledge representation and context utilization capacity across model scales. Smaller models lack sufficient internal crystallographic knowledge, so expert-reviewed passages provide missing conceptual structure and precise terminology that directly enhance reasoning accuracy. Mid-sized models already encode partial domain knowledge but still rely on external information to fill gaps, making them responsive to high-quality supporting text. In contrast, very large models possess extensive parametric knowledge, and adding redundant or overlapping context can introduce interference effects, where new information competes with their established internal representations. As a result, they gain little additional benefit and may even experience slight degradation when the external input does not perfectly align with their internal reasoning pathways.

Overall, these findings indicate that while expert curation improves performance across all model scales, the relative gains are most pronounced for small and mid-sized models that can still meaningfully integrate external guidance, whereas large models operate near their knowledge saturation limit.

3.5.4 Quantity-driven effects: evidence from token budget constraints. To further disentangle quality from quantity effects, Section 3.7 presents complementary evidence that information quantity alone can degrade performance. Our token budget analysis (Fig. 11) reveals that llamat-3-chat's open-book accuracy dropped from 36.87% at 256 tokens to 26.73% at 2048+ tokens, a 10 percentage point degradation, when provided with longer supporting materials. This inverse relationship demonstrates severe token starvation: the model exhausts its computational budget on input processing, leaving insufficient capacity for answer generation. The supporting materials, rather than aiding reasoning, become a computational burden that crowds out response formulation.

This finding establishes a critical control, if performance gains stemmed primarily from text volume rather than content quality, we would not observe degradation with increased token budgets in resource-constrained models. The token starvation phenomenon proves that "more text" does not guarantee better

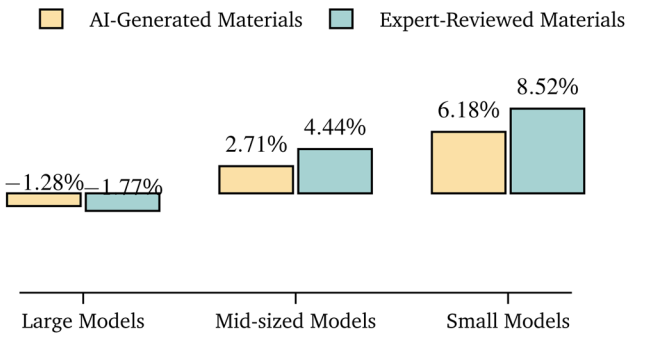


Fig. 10 Average performance gains from AI-generated *vs.* expert-reviewed supporting materials across model size categories. Large models (>70B parameters or advanced architectures like GPT-4, GPT-5, O1, O3-mini, $n = 27$); mid-sized models (7B–70B parameters including LLaVA-34B, QWEN2-7B, mistral variants, $n = 27$); small models (≤7B parameters including QWEN2-0.5B, LLaMA-3.2-3B, LLaVA-7B, LLaMAT variants, Honeybee-7B, $n = 20$). Note that large models show performance degradation with supporting materials, while mid-sized and small models benefit substantially from expert-reviewed materials. Overall, these results show that as models become larger and contain more built-in knowledge, the additional benefit from external information gradually diminishes.

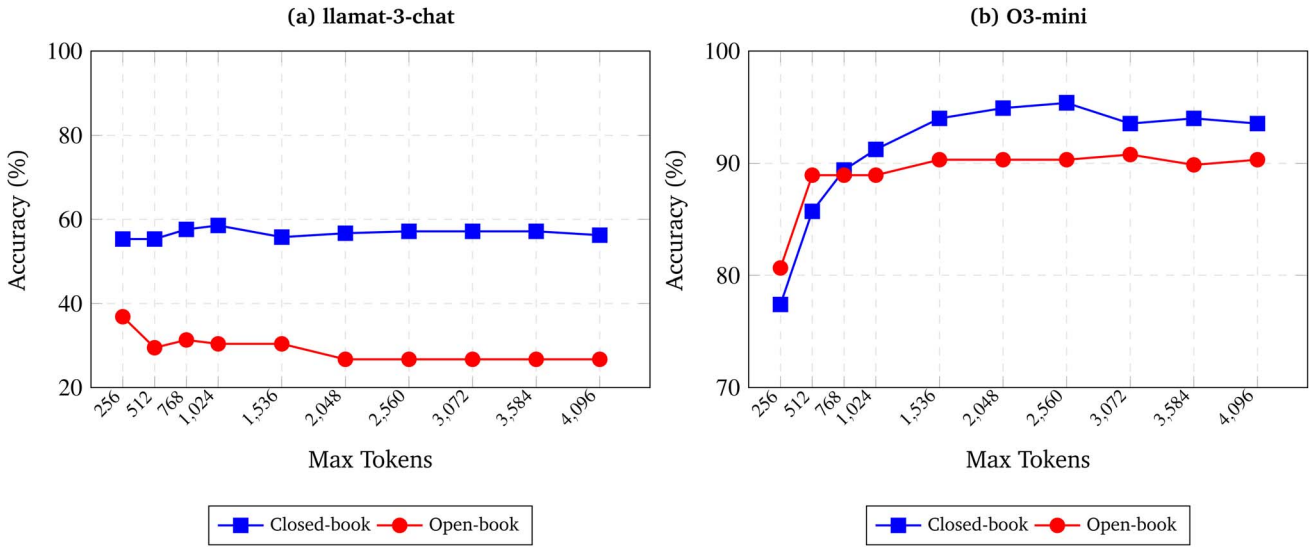


Fig. 11 Impact of token budget constraints on closed-book and open-book performance for (a) llamat-3-chat and (b) O3-mini. The *x*-axis shows maximum output tokens allowed, while the *y*-axis shows accuracy on the 217-question benchmark. Note the different *y*-axis scales: llamat-3-chat ranges 20–100% while O3-mini ranges 70–100% due to their different baseline capabilities.

outcomes and, in fact, can be counterproductive when models lack the capacity to efficiently process additional context.

3.5.5 Interpretation: quality, not quantity, drives improvements. The combination of token-matched comparisons (identical length, differential outcomes) and token budget analysis (increased length, degraded outcomes) provides convergent evidence that content quality, specifically, accuracy, relevance, and pedagogical clarity, drives the observed performance improvements, not mere text volume. Expert-reviewed materials successfully addressed the limitations identified in AI-generated content: technical inaccuracies, insufficient depth, tangential information, and suboptimal pedagogical framing. These qualitative enhancements enable mid-capacity models to better integrate external knowledge, bridging the gap between their processing capability and incomplete domain expertise.

Critically, the token-matched design ensures that expert improvements cannot be attributed to providing "more information" in terms of token count. Instead, expert review enhances information density, precision, and alignment with question-specific reasoning needs, dimensions of quality orthogonal to text length. This ablation establishes that the OPENXRD framework's performance gains reflect genuine knowledge enhancement rather than computational artifacts of longer context windows.

3.6 Detailed subtask-level performance with expert-reviewed materials

To gain deeper insights, we examined subtask-level performance with expert-reviewed materials. Table 8 shows how LLaVA-onevision-QWEN2-7B-ov behaves on key XRD subtasks. The rationale for selecting this model is its substantial enhancement from expert-reviewed materials, which offers valuable insights into the limits of performance improvement. For several subtasks such as Crystal Structure, Laue Patterns, Metal Structures, and Powder Diffraction, expert-reviewed materials transformed performance from complete failure (0%) to perfect accuracy (100%). Expert materials particularly improved performance on structural tasks, enhancing Structure Factors by 25% and Coordination Numbers by 40%. However, complex mathematical derivations like Bragg's Law and Calculation Methods remained challenging even with expert materials, suggesting that some concepts require more than textual explanation. An illustrative example is the question "What is the structure factor F for a base-centered unit cell when h and k are mixed (one even, one odd)?" Correct answer is being $F = f$, while a model consistently selecting $F = 2f$, reveals that it has not internalized how lattice symmetry produces systematic absences, an error that echoes its 0% score on Bragg's Law questions. Additionally, for a few subtasks where the model had strong initial performance, such as Complex Mathematics and Wave Scattering, supporting materials appeared to interfere with the model's existing knowledge.

Mathematically intensive subtasks including Bragg's Law derivations, structure factor calculations, and reflection condition analysis showed universal 0% improvement with supporting materials across all 74 models. For instance, when asked about structure factors for base-centered lattices with mixed h, k indices, models consistently failed to execute the symbolic computation $F = f[1 + e^{i\pi(h+k)}]$ despite correct textual explanations, revealing that current LLMs cannot perform the formal algebraic operations required for crystallographic problem-solving.

3.7 Impact of token budget constraints on open-book performance

While the previous analyses examined model performance under standard inference conditions, we now investigate how computational constraints—specifically, output token budget

Table 8 Selected subtask-level accuracy (%) for LLaVA-onevision-QWEN2-7B-ov, comparing closed-book mode *vs.* open-book mode with expert-reviewed materials. Some subtasks have fewer questions, so a single item shifts accuracy considerably

Subtask	Closed-book mode	Open-book mode	Improvement
Large gains			
Atomic spacing	66.7	100.0	+33.3
Coordination numbers	20.0	60.0	+40.0
Crystal structure	0.0	100.0	+100.0
Laue patterns	0.0	100.0	+100.0
Metal structures	0.0	100.0	+100.0
Powder diffraction	0.0	100.0	+100.0
Structure factors	60.0	85.0	+25.0
Little/no improvement			
Bragg's law	0.0	0.0	+0.0
Calculation methods	0.0	0.0	+0.0
Diffraction limitations	0.0	0.0	+0.0
Performance drops			
Complex mathematics	100.0	50.0	−50.0
Wave scattering	100.0	0.0	−100.0

limitations, affect models' ability to leverage external knowledge. Token budget restrictions are particularly relevant in deployment scenarios where computational resources or API costs impose practical limits on generation length. Understanding how these constraints interact with open-book augmentation is crucial for real-world applications.

We systematically evaluated two models across varying maximum token budgets (256 to 4096 tokens): llamat-3-chat, a 7B-parameter specialized model, and O3-mini, a reasoning-optimized model designed for scientific tasks. Fig. 11 presents the accuracy trends as token budgets increase.

The results reveal markedly different patterns between the two models. For llamat-3-chat (Fig. 11a), closed-book performance remains relatively stable across token budgets (55–58%), showing minimal sensitivity to generation length constraints. However, open-book performance exhibits a counterintuitive decline: accuracy drops from 36.87% at 256 tokens to 26.73% at 2048+ tokens, a degradation of approximately 10 percentage points. This inverse relationship suggests severe token starvation: as the model attempts to process longer supporting materials, it exhausts its token budget on input comprehension and internal reasoning, leaving insufficient capacity for coherent answer generation. The supporting materials, rather than aiding reasoning, become a computational burden that crowds out the model's ability to formulate responses.

In contrast, O3-mini (Fig. 11b) demonstrates more favorable scaling properties. Closed-book accuracy improves substantially from 77.42% at 256 tokens to a peak of 95.39% at 2560 tokens, indicating the model benefits from additional reasoning space for complex crystallographic problems. Open-book performance similarly improves from 80.65% to approximately 90%, stabilizing around 1024 tokens. The gap between closed-book and open-book performance narrows at higher token budgets, suggesting that O3-mini can more effectively integrate external knowledge when given adequate computational headroom. However, even O3-mini shows slight performance degradation beyond 2560 tokens in closed-book mode, hinting at potential overfitting or unnecessary elaboration when unconstrained.

These findings have important implications for deployment strategies. First, token budget selection requires careful calibration: smaller models like llamat-3-chat exhibit pathological behavior under open-book conditions when token budgets exceed their processing capacity, while more capable models like O3-mini require minimum token budgets (1024 tokens) to effectively leverage external materials. Second, the phenomenon of "open-book degradation" in resource-constrained models suggests that naïvely providing supporting materials without ensuring sufficient token allocation for answer synthesis can be counterproductive. Third, the results underscore fundamental architectural differences: reasoning-optimized models (O3-mini) appear better equipped to manage the token allocation trade-off between comprehending external context and generating accurate responses.

From a practical standpoint, these observations suggest that effective open-book augmentation requires not only high-quality supporting materials but also appropriate computational provisioning. For deployment scenarios with strict token limits, practitioners should either: (1) use models specifically designed for reasoning tasks that can efficiently allocate tokens between context processing and response generation, (2) employ adaptive token budgets that scale with input complexity, or (3) pre-process supporting materials to reduce their token footprint while preserving essential information. Future work should investigate token allocation strategies that dynamically balance between context comprehension and response generation, potentially through attention mechanism modifications or explicit token budgeting protocols.

4 Discussion

Our results show that using domain-specific prompts, or open-book mode, significantly enhances question-answering accuracy in crystallography. This is especially true for smaller or more general models, where the additional context helps fill knowledge gaps. However, the quality and relevance of the supporting textual material are crucial. In some instances, the material generated by GPT-4.5 did not align well with the question's needs, causing confusion or contradictory information. This underscores the importance of relevance filtering or validation of supporting material before it is used.

Although OPENXRD uses a multiple-choice format, we can still comment on reliability-relevant behavior using analyses already included in the manuscript. The token-matched study (Section 3.5) shows that expert-reviewed passages outperform AI-generated passages at nearly identical token counts, supporting that gains arise from higher-quality guidance rather than additional text. In addition, our error analyses illustrate interference when context mixes confounders or partially misaligns with the question intent (*e.g.*, Fig. 8), indicating that external context can either support or hinder performance depending on its alignment and quality. For this reason,

OPENXRD should be interpreted as a controlled diagnostic of context assimilation and answer reliability under fixed guidance, not a direct evaluation of open-ended solution quality or reasoning style.

While OPENXRD focuses on inference-time context assimilation, it is explicitly designed to complement parameter-efficient fine-tuning (PEFT) approaches. PEFT methods can encode crystallographic knowledge directly into model weights through techniques like LoRA,[62] whereas OPENXRD diagnoses whether models, adapted or not, can effectively utilize answer-guiding context that does not directly reveal solutions. A comprehensive evaluation strategy might compare: (i) closed-book performance after PEFT on external XRD corpora, (ii) open-book performance after PEFT using our oracle passages, and (iii) our baseline open-book performance without PEFT. This three-way comparison decomposes gains attributable to parametric adaptation *versus* gains from inference-time guidance, a separation difficult to obtain in end-to-end training studies. For instance, a PEFT-adapted model that shows minimal open-book improvement might indicate successful knowledge internalization, while a model showing large open-book gains suggests incomplete adaptation where external guidance remains valuable. OPENXRD standardizes this diagnostic capability across architectures and scales.

Advanced mathematical reasoning remains a challenge. Our analysis indicates that open-book mode rarely resolves complex mathematical problems, such as structure factor calculations or multi-step interference proofs. Future enhancements could include integrating symbolic math modules or domain-specific solvers to better handle these tasks, rather than relying solely on textual references.

The complexity of visual data in crystallography, such as tabulated results, diffraction patterns, and annotated structural diagrams, presents additional challenges. Accurate extraction of this information requires improvements in optical character recognition (OCR), figure parsing, and multi-modal alignment. Developing tailored visual backbones or domain-tuned OCR engines will be essential for applying our methods to real-world scenarios, like analyzing lab notebooks or older textbook scans.

Text-based supporting materials prove insufficient for mathematically intensive subtasks requiring symbolic manipulation. Consider a diagnostic case from our benchmark: "For a base-centered unit cell, what is the structure factor F when h and k are mixed (one even, one odd)?" Despite expert-reviewed materials correctly explaining that atoms at (0,0,0) and (1/2,1/2,0) produce destructive interference when h and k have mixed parity (yielding $F = 0$), multiple frontier models incorrectly answered $F = 2f$. Analysis reveals they cannot maintain intermediate symbolic states during phase calculations $\phi = 2\pi(hx + ky + lz)$ or correctly apply $F = f[1 + e^{i\pi(h+k)}] = 0$ when $(h + k)$ is odd. This failure extends to Bragg's Law derivations, structure factor algebra, and powder diffraction indexing, all showing 0% improvement across our 74-model evaluation regardless of size or architecture.

To overcome this architectural limitation, future work should integrate LLMs with symbolic computation engines such as SymPy[108] or Wolfram Alpha[109] for exact algebraic manipulations, crystallographic knowledge graphs encoding systematic absence rules and space group constraints as structured logical representations rather than text, and domain-specific software including GSAS-II[110] for powder diffraction refinement and Mercury[111] for structure visualization. Such hybrid architectures would enable the LLM to parse problems linguistically, delegate mathematical operations to specialized modules, and synthesize results, combining linguistic flexibility with computational rigor. This approach extends beyond crystallography to other physics-heavy domains requiring both conceptual reasoning and mathematical precision.

In parallel, a notable tension exists between generalist and specialist models. While larger LLMs like GPT-4 and GPT-4.5 have broad knowledge, they often miss the nuances of specific domains that specialized models capture more effectively. A hybrid approach, where a specialist model generates domain-specific references for a generalist model, could offer a powerful solution. Additionally, ethical and copyright considerations must be managed carefully, especially as we look to expand these systems to other technical fields with extensive proprietary literature.

Our research identifies two primary methods for enhancing model performance with domain expertise: directly embedding knowledge into large models, such as frontier models like GPT-5, GPT-4.5, and O3-mini, and providing models with access to expert-reviewed references during inference. This latter strategy proves especially useful for mid-capacity models, enabling them to nearly match the performance of their larger counterparts by leveraging a "knowledge bridge".

For example, a medium-sized model like LLaVA-v1.6-34B, when paired with expert materials, achieves 78.34% accuracy, approaching the performance of much larger models like Llama-3.1-405B (84.33%) and even surpassing some frontier models in cost-effectiveness but with far fewer resources required. The integration of expert-reviewed domain knowledge is not only beneficial in textual analysis but also in multi-modal contexts such as medical imaging or geospatial analysis, where expert-reviewed context can similarly boost specialized question-answering performance and unlock new application areas.

However, the addition of external information is not always beneficial. For instance, multiple frontier models show performance degradation with expert-reviewed supporting materials: GPT-4.5-preview (−3.23%), GPT-5 (−3.68%), Grok-4-fast (−5.78%), Claude-3.5-Sonnet (−1.84%), and dziner-qwen-2.5-72b (−4.14%), possibly due to conflicts with their pre-trained internal knowledge. This suggests that more information is not necessarily better, particularly for top-tier models that have been extensively trained, as the interference from external context can disrupt their already-comprehensive internal reasoning processes.

Our results show that both small and mid-sized models benefit from external knowledge, with small models exhibiting the largest relative gains, whereas large models experience interference. This reveals a fundamental insight: external knowledge augmentation is most effective within a specific capability range where models have sufficient processing power but incomplete domain coverage. This finding has important

implications for cost-effective deployment strategies in specialized scientific domains.

Future research should focus on automated methods to determine when expert review is most beneficial, multi-stage approaches for targeted expert review, and the potential for fine-tuning models to better utilize external references, as well as mechanisms to detect when models have sufficient internal knowledge that external context becomes counterproductive.

Beyond incremental improvements, addressing the universal mathematical reasoning failures documented in our benchmark requires architectural innovation. Hybrid systems coupling LLMs with symbolic algebra engines, structured crystallographic knowledge graphs encoding lattice symmetries and systematic absence rules, and specialized software for diffraction simulation and structure refinement would enable models to delegate exact computations while maintaining linguistic problem-solving capabilities. Multi-modal extensions incorporating actual XRD patterns and crystal structure visualizations would test whether vision-language models can integrate visual and textual information. Dynamic retrieval replacing oracle passages with corpus-retrieved content would decompose RAG accuracy into retrieval *versus* assimilation components.

Moreover, our results suggest a concrete way to integrate our contribution into RAG: treat expert review as a knowledge-base quality intervention (and/or post-retrieval refinement) rather than as a replacement for retrieval. The token-matched ablation shows that accuracy gains persist when passage length is controlled, implying that improvements arise from precision, relevance, and pedagogical structure, not from providing more text. In RAG deployments for XRD, the analogous lever is to improve the retrieved evidence *via* domain-aware filtering, reranking, or expert-curated reference notes, thereby reducing distraction and mechanism-mixing that can otherwise degrade generator performance. Fine-tuning studies using OPENXRD before and after domain adaptation would reveal whether parametric knowledge embedding complements or supersedes inference-time guidance. Cross-domain validation in related scattering techniques and computational chemistry would test whether our findings about model capacity and external knowledge effectiveness generalize beyond crystallography.

Despite these advancements, certain limitations remain. Complex mathematical concepts and advanced tasks like reflection or symmetry analysis still pose challenges, as models do not always fully grasp these concepts through text-based materials alone. Moreover, there are instances where supporting materials can detract from a model's performance if they do not align perfectly with the model's pre-existing knowledge base.

Given these limitations, the reliance of open-book question answering on human-curated content may be difficult to sustain across all fields, and the uneven gains observed across model sizes underscore the need for further work on how best to select, refine, and integrate external knowledge.

5 Conclusions

The OPENXRD benchmark provides a controlled and extensible framework for evaluating how large language models (LLMs) and multimodal LLMs integrate external scientific knowledge. Through token-matched experiments and token-budget analyses, the study isolates the effect of content quality from text quantity and reveals that expert-reviewed supporting materials consistently enhance model accuracy, even when passage length is identical. Small and mid-sized models both exhibit clear accuracy improvements when augmented with expert-reviewed materials, with the relative gain being most pronounced for small models, whereas large models show minimal or slightly negative changes due to knowledge saturation. These findings demonstrate that as model capacity increases, the relative advantage of external context diminishes, even though overall accuracy continues to scale with size.

Beyond these controlled experiments, OPENXRD demonstrates strong scalability and generalizability across 74 diverse model architectures, including general-purpose systems (GPT, Claude), reasoning-optimized models (O-series), code-specialized variants (QWEN-coder), domain-adapted frameworks (LLaMAT, dZiner), and vision-language models (LLaVA). This diversity confirms that OPENXRD is model-agnostic, functioning as a diagnostic layer for both closed-book and retrieval-augmented reasoning systems. For instance, the code-specialized model QWEN3-coder-plus improved dramatically, from 22.99% in the closed-book setting (Table 1) to 88.94% in the open-book setting (Table 4) with expert-reviewed materials, demonstrating that OPENXRD's insights extend beyond language-only systems and generalize effectively across diverse architectural paradigms. Moreover, because OPENXRD isolates inference-time reasoning while remaining compatible with dynamic retrieval pipelines, it can be directly extended into a deployable RAG-evaluation suite that separates retrieval quality from assimilation capability.

Finally, the framework's practical deployment insights show how these findings can guide real-world applications. OPENXRD identifies mid-sized models (7B–70B parameters) augmented with expert-reviewed passages as an optimal balance between cost and performance, approaching or even matching large-model accuracy while requiring significantly fewer computational resources. This provides a cost-effective deployment pathway for organizations with limited budgets. The framework also enables the design of adaptive token-budget strategies that scale with task complexity, along with pre-processing approaches that reduce token usage while preserving essential information. Such model-selection and resource-allocation guidelines directly address the economic and computational realities of deploying AI systems at scale.

In conclusion, OPENXRD is not an artificial or static experiment, but a reproducible, extensible, and practically informative framework. It quantifies how model capacity, resource budgets, and content quality interact under real-world constraints, provides scalable and legally compliant data generation methods, and delivers empirical evidence to support efficient design of open-book and retrieval-augmented reasoning systems. The framework's diagnostic capabilities empower organizations to make informed decisions about when to scale model size *versus* when to enhance smaller models with curated external knowledge, thereby advancing

both the scientific understanding and the practical deployment of domain-specialized language models.

Author contributions

A. V. and A. S. contributed equally to this work as co-first authors. A. V.: conceptualization, methodology, software, formal analysis, investigation, data curation, writing – original draft, visualization. A. S.: conceptualization, methodology, data curation, investigation, validation, writing – review & editing. A. S. designed the crystallography question set and led the expert review process for GPT-generated questions. Z. Z., Y. X., and G. H.: methodology, software, investigation, validation, writing – review & editing. Z. Z., Y. X., and G. H. provided substantial contributions to model development, experimental design improvements, and participated in the expert review and refinement of GPT-generated questions to ensure scientific accuracy. C. X. and N. A.: conceptualization, supervision, project administration, resources, writing – review & editing, funding acquisition. C. X. and N. A. jointly led and supervised the project, providing strategic direction and ensuring research quality throughout all phases of the work.

Conflicts of interest

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

Data availability

The complete dataset of 217 expert-curated crystallography questions, supporting materials (both AI-generated and expert-reviewed versions), evaluation code, and detailed documentation are publicly available at GitHub: https://github.com/niaz60/OpenXRD and archived on Zenodo at https://doi.org/10.5281/zenodo.18980534.

Acknowledgements

We gratefully acknowledge financial support from multiple sources. A. V., G. H., C. X., and N. A. received support from the National Nuclear Security Administration (NNSA) under grant NA0004078. A. S., Z. Z., Y. X., C. X., and N. A. were supported by the National Science Foundation (NSF) under grant 2202124. Additional funding for N. A. was provided by the Department of Energy (DOE) under award DE-SC0020340. We thank the crystallography experts who participated in the review and validation of our question dataset. We also acknowledge the computational resources and infrastructure that enabled the extensive model evaluations presented in this work.

Notes and references

1 B. B. He, *Two-dimensional X-ray Diffraction*, John Wiley & Sons, 2018.
2 I. Rathore, V. Mishra and P. Bhaumik, *Emerging Top. Life Sci.*, 2021, **5**, 127–149.
3 R. Ubic, *Crystallography and Crystal Chemistry*, Springer, 2024.
4 B. K. Vainshtein, *Fundamentals of Crystals: Symmetry, and Methods of Structural Crystallography*, Springer Science & Business Media, 2013, vol. 1.
5 U. Müller and G. De La Flor, *Symmetry Relationships Between Crystal Structures: Applications of Crystallographic Group Theory in Crystal Chemistry*, Oxford University Press, 2024, vol. 24.
6 D. Raabe, *Chem. Rev.*, 2023, **123**, 2436–2608.
7 G. Altuntaş, O. Altuntaş, M. K. Öztürk and B. Bostan, *Int. J. Metalcast.*, 2023, **17**, 1340–1349.
8 V. Bijak, M. Szczygiel, J. Lenkiewicz, M. Gucwa, D. R. Cooper, K. Murzyn and W. Minor, *Expert Opin. Drug Discovery*, 2023, **18**, 1221–1230.
9 S. N. Afraj, C.-C. Lin, A. Velusamy, C.-H. Cho, H.-Y. Liu, J. Chen, G.-H. Lee, J.-C. Fu, J.-S. Ni, S.-H. Tung, *et al.*, *Adv. Funct. Mater.*, 2022, **32**, 2200880.
10 S. Inoue, K. Nikaido, T. Higashino, S. Arai, M. Tanaka, R. Kumai, S. Tsuzuki, S. Horiuchi, H. Sugiyama, Y. Segawa, *et al.*, *Chem. Mater.*, 2021, **34**, 72–83.
11 B. D. Cullity and R. Smoluchowski, *Phys. Today*, 1957, **10**, 50.
12 A. Ali, Y. W. Chiang and R. M. Santos, *Minerals*, 2022, **12**, 205.
13 A. A. Bunaciu, E. G. UdriŞTioiu and H. Y. Aboul-Enein, *Crit. Rev. Anal. Chem.*, 2015, **45**, 289–299.
14 B. Cantor, *The Equations of Materials*, Oxford University Press, 2020.
15 L. Bragg, *Sci. Am.*, 1968, **219**, 58–74.
16 C. G. Pope, *J. Chem. Educ.*, 1997, **74**, 129.
17 J. Stöhr, *The Nature of X-rays and Their Interactions with Matter*, Springer, 2023.
18 S. Fatimah, R. Ragadhita, D. F. Al Husaeni and A. B. D. Nandiyanto, *ASEAN J. Sci. Eng.*, 2022, **2**, 65–76.
19 S. Shah, P. Mohammadi, B. G. Singidas, K. A. Smith, Y. Gu, L. J. Langston, B. Taylor, R. P. Siebenaller, M. A. Susner, S.-W. Cheong, *et al.*, *npj 2D Mater. Appl.*, 2026, **10**, 17.
20 J.-W. Lee, W. B. Park, M. Kim, S. P. Singh, M. Pyo and K.-S. Sohn, *Inorg. Chem. Front.*, 2021, **8**, 2492–2504.
21 V. Uvarov, *Appl. Crystallogr.*, 2019, **52**, 252–261.
22 K. He, N. Chen, C. Wang, L. Wei and J. Chen, *Cryst. Res. Technol.*, 2018, **53**, 1700157.
23 S. S. S. Qadr, *Academia Open*, 2023, **8**, 10–21070.
24 S. Dolabella, A. Borzì, A. Dommann and A. Neels, *Small Methods*, 2022, **6**, 2100932.
25 P. Mohammadi and S. Singh, *Phys. Rev. B*, 2025, **112**, 125205.
26 J. E. Salgado, S. Lerman, Z. Du, C. Xu and N. Abdolrahim, *npj Comput. Mater.*, 2023, **9**, 214.
27 A. Ziletti, D. Kumar, M. Scheffler and L. M. Ghiringhelli, *Nat. Commun.*, 2018, **9**, 2775.
28 W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin and K.-S. Sohn, *IUCrJ*, 2017, **4**, 486–494.

29 J. I. Gómez-Peralta, X. Bokhimi and P. Quintana, *J. Phys. Chem. A*, 2023, **127**, 7655–7664.
30 A. Chakraborty and R. Sharma, *Vis. Comput.*, 2022, **38**, 1275–1282.
31 J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh and K.-S. Sohn, *Nat. Commun.*, 2020, **11**, 86.
32 N. J. Szymanski, C. J. Bartel, Y. Zeng, M. Diallo, H. Kim and G. Ceder, *npj Comput. Mater.*, 2023, **9**, 31.
33 F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. Von Lilienfeld, *J. Chem. Theory Comput.*, 2017, **13**, 5255–5264.
34 V.-A. Surdu and R. Győrgy, *Appl. Sci.*, 2023, **13**, 9992.
35 B. D. Lee, J.-W. Lee, J. Ahn, S. Kim, W. B. Park and K.-S. Sohn, *Adv. Intell. Syst.*, 2023, **5**, 2300140.
36 L. Wu, S. Yoo, A. F. Suzana, T. A. Assefa, J. Diao, R. J. Harder, W. Cha and I. K. Robinson, *npj Comput. Mater.*, 2021, **7**, 175.
37 N. E. Omori, A. D. Bobitan, A. Vamvakeros, A. M. Beale and S. D. Jacques, *Philos. Trans. R. Soc., A*, 2023, **381**, 20220350.
38 E. Chávez-Angel, M. B. Eriksen, A. Castro-Alvarez, J. H. Garcia, M. Botifoll, O. Avalos-Ovando, J. Arbiol and A. Mugarza, *Adv. Intell. Syst.*, 2025, 2400986.
39 Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, *et al.*, *Meta-Radiol.*, 2023, **1**, 100017.
40 M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu and S. Mirjalili, *A survey on large language models: applications, challenges, limitations, and practical usage*, Authorea Preprints, 2023, vol. 1, preprint, pp. 1–26.
41 M. Dahl, V. Magesh, M. Suzgun and D. E. Ho, *J. Leg. Anal.*, 2024, **16**, 64–93.
42 Y. Zhuang, Y. Yu, K. Wang, H. Sun and C. Zhang, *Adv. Neural Inf. Process. Syst.*, 2023, **36**, 50117–50143.
43 E. Kamalloo, N. Dziri, C. Clarke and D. Rafiei, Evaluating Open-Domain Question Answering in the Era of Large Language Models, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 5591–5606.
44 Z. Chen, Y. Li and K. Wang, *Optimizing Reasoning Abilities in Large Language Models: A Step-by-Step Approach*, Authorea Preprints, 2024.
45 N. Patel, M. Kulkarni, M. Parmar, A. Budhiraja, M. Nakamura, N. Varshney and C. Baral, *arXiv*, 2024, preprint, arXiv:2406.17169, DOI: 10.48550/2406.17169.
46 M. Zaki, N. A. Krishnan, *et al.*, *Digital Discovery*, 2024, **3**, 313–327.
47 L. S. Balhorn, J. M. Weber, S. Buijsman, J. R. Hildebrandt, M. Ziefle and A. M. Schweidtmann, *arXiv*, 2023, preprint, arXiv:2309.10048, DOI: 10.48550/arXiv.2309.10048.
48 A. Mirza, N. Alampara, S. Kunchapu, M. Ríos-García, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh *et al.*, *arXiv*, 2024, preprint, arXiv:2404.01475, DOI: 10.48550/2404.01475.
49 K. Choudhary, *J. Phys. Chem. Lett.*, 2025, **16**, 7028–7035.
50 A. N. Rubungo, C. Arnold, B. P. Rand and A. B. Dieng, *arXiv*, 2023, preprint, arXiv:2310.14029, DOI: 10.48550/2310.14029.
51 C. Qian, H. Tang, Z. Yang, H. Liang and Y. Liu, *arXiv*, 2023, preprint, arXiv:2307.07443, DOI: 10.48550/2307.07443.
52 Y. Li, V. Gupta, M. N. T. Kilic, K. Choudhary, D. Wines, W.-k. Liao, A. Choudhary and A. Agrawal, *Digital Discovery*, 2025, **4**, 376–383.
53 Y. Chiang, C. Chou and J. Riebesell, *arXiv*, 2024, preprint, arXiv:2401.17244, DOI: 10.48550/2401.17244.
54 T. Gupta, M. Zaki, N. A. Krishnan and M. Mausam, *npj Comput. Mater.*, 2022, **8**, 102.
55 S. Yu, N. Ran and J. Liu, *Artif. Intell. Chem.*, 2024, **2**, 100076.
56 Y. Shi, N. Rampal, C. Zhao, D. J. Fu, C. Borgs, J. T. Chayes and O. M. Yaghi, *Digital Discovery*, 2025, **4**, 2676–2683.
57 H. Wang, K. Li, S. Ramsay, Y. Fehlis, E. Kim and J. Hattrick-Simpers, *Digital Discovery*, 2025, **4**, 1612–1624.
58 L. M. Antunes, K. T. Butler and R. Grau-Crespo, *Nat. Commun.*, 2024, **15**, 10570.
59 F. L. Johansen, U. Friis-Jensen, E. B. Dam, K. M. Ø. Jensen, R. Mercado and R. Selvan, *arXiv*, 2025, preprint, arXiv:2502.02189, DOI: 10.48550/2502.02189.
60 K. Choudhary, *J. Phys. Chem. Lett.*, 2025, **16**, 2110–2119.
61 K. Choudhary, *J. Phys. Chem. Lett.*, 2024, **15**, 6909–6917.
62 E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, *International Conference on Learning Representations*, 2022.
63 OpenAI, *Introducing GPT-5*, OpenAI Blog, 2025, https://openai.com/index/introducing-gpt-5/, Released August 7, 2025.
64 OpenAI, *Introducing Upgrades to Codex: GPT-5-Codex*, OpenAI Blog, 2025, https://openai.com/index/introducing-upgrades-to-codex/, Released September 23, 2025; optimized version of GPT-5 for agentic coding.
65 OpenAI, *OpenAI o3-mini System Card, OpenAI Technical Report*, 2025, https://cdn.openai.com/o3-mini-system-card-feb10.pdf, Released January 31, 2025.
66 OpenAI, OpenAI o1 System Card, *arXiv*, 2024, preprint, arXiv:2412.16720, DOI: 10.48550/arXiv.2412.16720, https://arxiv.org/abs/2412.16720, System card for o1 and o1-mini models.
67 OpenAI, *Introducing GPT-4.5*, OpenAI Blog, 2025, https://openai.com/index/introducing-gpt-4-5/, Code-named Orion, released February 27, 2025; OpenAI's largest pre-trained model.
68 OpenAI, GPT-4o System Card, *arXiv*, 2024, preprint, arXiv:2410.21276, DOI: 10.48550/arXiv.2410.21276, https://arxiv.org/abs/2410.21276, Omnimodal model accepting text, audio, image, and video inputs; includes GPT-4o-mini.
69 OpenAI, *New Models and Developer Products Announced at DevDay*, OpenAI Blog, 2023, https://openai.com/index/new-models-and-developer-products-announced-at-devday/, GPT-4 Turbo with 128K context window; gpt-4-1106-preview and subsequent versions.
70 OpenAI, GPT-4 Technical Report, *arXiv*, 2023, preprint, arXiv:2303.08774, DOI: 10.48550/arXiv.2303.08774, https://arxiv.org/abs/2303.08774, Multimodal model accepting image and text inputs.
71 OpenAI, *Introducing ChatGPT and Whisper APIs*, OpenAI Blog, 2023, https://openai.com/index/introducing-chatgpt-

and-whisper-apis/, GPT-3.5-turbo (gpt-3.5-turbo-0301), released March 1, 2023; powers ChatGPT.
72 OpenAI, *GPT-3.5 Turbo Fine-Tuning and API Updates*, OpenAI Blog, 2023, https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/, GPT-3.5-turbo-16k with extended 16K context window announced, June 13, 2023.
73 Anthropic, *The Claude 3 Model Family: Opus, Sonnet, Haiku, Anthropic technical report*, 2024.
74 Anthropic, *Claude 3.5 Sonnet Model Card Addendum, Anthropic technical report*, 2024.
75 Anthropic, *Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet, Anthropic technical report*, 2024.
76 Llama Team, *arXiv*, 2024, preprint, arXiv:2407.21783, DOI: 10.48550/arXiv.2407.21783.
77 Meta AI, *Llama 3.2: Revolutionizing edge AI and vision with open, customizable models*, Meta AI Blog, 2024, https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, Announced at Meta Connect 2024. Covers LLaMA-3.2-1B/3B-instruct.
78 A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou and Z. Qiu, *arXiv*, 2025, preprint, arXiv:2505.09388, DOI: 10.48550/arXiv.2505.09388.
79 Qwen Team, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang and Z. Qiu, *arXiv*, 2024, preprint, arXiv:2412.15115, DOI: 10.48550/arXiv.2412.15115.
80 Q. Team, Qwen3-Coder: The Code Version of Qwen3, *arXiv*, 2025, preprint, arXiv:2505.09388, DOI: 10.48550/arXiv.2505.09388, Based on Qwen3 Technical Report, https://github.com/QwenLM/Qwen3-Coder.
81 B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Dang, A. Yang, R. Rong, Y. Xue, J. Fu, J. Ma and J. Lin, *arXiv*, 2024, preprint, arXiv:2409.12186, DOI: 10.48550/arXiv.2409.12186.
82 T. D. Team and T. Lab, *Tongyi-DeepResearch: An Agentic Large Language Model for Deep Information-Seeking Tasks*, 2025, https://github.com/Alibaba-NLP/DeepResearch, 30B total parameters, 3B activated per token.
83 Mistral AI Team, *Au Large: Mistral Large*, Mistral AI Blog, 2024, https://mistral.ai/news/mistral-large, 123B parameters, 32K context window, available on Azure and La Plateforme.
84 Mistral AI Team, *Mistral Small: Optimized Model for Low Latency Workloads*, Mistral AI Blog, 2024, https://mistral.ai/news/mistral-large, Released alongside Mistral Large, optimized for latency and cost.
85 A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix and W. El Sayed, *arXiv*, 2023, preprint, arXiv:2310.06825, DOI: 10.48550/arXiv.2310.06825.
86 A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. Bou Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. Le Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix and W. El Sayed, *arXiv*, 2024 preprint, arXiv:2401.04088, DOI: 10.48550/arXiv.2401.04088.
87 Mistral AI Team, *Cheaper, Better, Faster, Stronger: Mixtral 8x22B*, Mistral AI Blog, 2024, https://mistral.ai/news/mixtral-8x22b, 141B parameters with 39B active, Apache 2.0 license.
88 P. Agrawal, S. Antoniak, E. Bou Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. De Monicault, S. Garg, T. Gervet, S. Ghosh, A. Héliou, P. Jacob, A. Q. Jiang, K. Khandelwal, T. Lacroix, G. Lample, D. Las Casas, T. Lavril, T. Le Scao, A. Lo, W. Marshall, L. Martin, A. Mensch, P. Muddireddy, V. Nemychnikova, M. Pellat, P. Von Platen, N. Raghuraman, B. Rozière, A. Sablayrolles, L. Saulnier, R. Sauvestre, W. Shang, R. Soletskyi, L. Stewart, P. Stock, J. Studnia, S. Subramanian, S. Vaze, T. Wang and S. Yang, *arXiv*, 2024, preprint, arXiv:2410.07073, DOI: 10.48550/arXiv.2410.07073.
89 DeepSeek-AI, DeepSeek-V3 Technical Report, *arXiv*, 2024, preprint, arXiv:2412.19437, DOI: 10.48550/arXiv.2412.19437, https://arxiv.org/abs/2412.19437.
90 Gemini Team, *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*, Google deepmind technical report, 2025.
91 Gemma Team, M. Riviere, S. Pathak, P. Giuseppe Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, *et al.*, Gemma 2: Improving Open Language Models at a Practical Size, *arXiv*, 2024, preprint, arXiv:2408.00118, DOI: 10.48550/arXiv.2408.00118, https://arxiv.org/abs/2408.00118.
92 Amazon Artificial General Intelligence, *The Amazon Nova Family of Models: Technical Report and Model Card, Amazon technical report*, 2024.
93 xAI, *Grok 4 Fast Model Card, xai technical report*, 2025.
94 H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen and Y. J. Lee, *LLaVA-NeXT: Improved reasoning, OCR, and world knowledge*, 2024, https://llava-vl.github.io/blog/2024-01-30-llava-next/.
95 H. Liu, C. Li, Y. Li and Y. J. Lee, Improved Baselines with Visual Instruction Tuning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26296–26306, https://ieeexplore.ieee.org/document/10655294.

96 B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu and C. Li, LLaVA-OneVision: Easy Visual Task Transfer, *arXiv*, 2024, preprint, arXiv:2408.03326, DOI: 10.48550/arXiv.2408.03326, https://arxiv.org/abs/2408.03326.

97 V. Mishra, S. Singh, D. Ahlawat, M. Zaki, V. Bihani, H. S. Grover, B. Mishra, S. Miret, Mausam and N. M. A. Krishnan, *arXiv*, 2024, preprint, arXiv:2412.09560, DOI: 10.48550/arXiv.2412.09560.

98 M. Ansari, J. Watchorn, C. E. Brown and J. S. Brown, *AI for Accelerated Materials Design-NeurIPS 2024*, 2024.

99 Perplexity AI, *Sonar Pro: Advanced Search Model with Enhanced Results*, 2025, https://docs.perplexity.ai/getting-started/models/models/sonar-pro, Context window: 200,000 tokens. Enhanced search capabilities with 2x more citations.

100 A. I. Perplexity, *Meet New Sonar: A Blazing Fast Model Optimized for Perplexity Search*, https://www.perplexity.ai/hub/blog/meet-new-sonar, 2025, https://www.perplexity.ai/hub/blog/meet-new-sonar, Built on Llama 3.3 70B. Context window: 127,000 tokens.

101 Meituan LongCat Team, Bayan, B. Li, B. Lei *et al.*, *arXiv*, 2025, preprint, arXiv:2509.01322, DOI: 10.48550/arXiv.2509.01322.

102 Arcee.ai, *AFM-4.5B: A 4.5 Billion Parameter Instruction-Tuned Foundation Model*, https://huggingface.co/arcee-ai/AFM-4.5B, 2025, HuggingFace model card. Trained on 8 trillion tokens (6.5T pretraining + 1.5T midtraining). Apache-2.0 license.

103 Undi95, *ReMM-SLERP-L2-13B: A Recreation of MythoMax with Updated Models*, https://huggingface.co/Undi95/ReMM-SLERP-L2-13B, 2023, HuggingFace model card. Re:MythoMax (ReMM) is a recreation trial of the original MythoMax-L2-B13 with updated models using SLERP merging technique.

104 Gryphe, *MythoMax-L2-13b: An Improved Variant of MythoMix*, 2023, https://huggingface.co/Gryphe/MythoMax-L2-13b, HuggingFace model card. A merge of MythoLogic-L2 and Huginn using a highly experimental tensor type merge technique.

105 J. Cha, W. Kang, J. Mun and B. Roh, Honeybee: Locality-enhanced Projector for Multimodal LLM, *arXiv*, 2023, preprint, arXiv:2312.06742, DOI: 10.48550/arXiv.2312.06742, https://arxiv.org/abs/2312.06742.

106 OpenRouter, *OpenRouter Auto Router: Dynamic Model Selection Powered by Not Diamond*, 2023, https://openrouter.ai/openrouter/auto, Model routing service created November 8, 2023. Routes prompts to optimal models using Not Diamond meta-model.

107 S. M. Narayanan, J. D. Braza, R. R. Griffiths, A. Bou, G. Wellawatte, M. C. Ramos, L. Mitchener, S. G. Rodriques and A. D. White, *arXiv*, 2025, preprint, arXiv:2506.17238, DOI: 10.48550/arXiv.2506.17238.

108 A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, *et al.*, *PeerJ Comput. Sci.*, 2017, **3**, e103.

109 Wolfram Research, *Wolfram Alpha*, 2009, https://www.wolframalpha.com.

110 B. H. Toby and R. B. Von Dreele, *J. Appl. Crystallogr.*, 2013, **46**, 544–549.

111 C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler, *et al.*, *J. Appl. Crystallogr.*, 2020, **53**, 226–235.