Seeing Beyond Words: MatVQA for Challenging Visual-Scientific Reasoning in Materials Science

Sifan Wu[1,2] **Huan Zhang**[1,2] **Yizhan Li**[1,2] **Farshid Effaty**[1,2]
Hongyuan Mei[3] **Amirreza Ataei**[4] **Bang Liu**[1,2,5]

[1] Institute Courtois & DIRO, Université de Montréal; [2] Mila - Quebec AI Institute;
[3] Toyota Technological Institute of Chicago; [4] Chemia Discovery Inc.;
[5] Canada CIFAR AI Chair

Abstract

The emergence of Multimodal Large Language Models (MLLMs) that integrate vision and language modalities has unlocked new potentials for scientific reasoning, outperforming prior benchmarks in both natural language and coding domains. Current materials science evaluation datasets such as MaScQA and SciQA remain largely text-based and fail to capture the visual and research-level analytic complexity required in materials discovery and design. We introduce MatVQA, a scalable benchmark specifically designed to address this gap. Generated via an automated pipeline, MArxivAgent, from recent materials literature, MatVQA features 1672 questions across four critical structure-property-performance (SPP) reasoning tasks. Uniquely, MatVQA employs an iterative process to eliminate textual shortcuts, compelling MLLMs to perform fine-grained, low-level visual analysis of material imagery (e.g., microscopy, diffraction patterns) integrated with multi-step scientific reasoning. Benchmarking 19 open- and closed-source MLLMs on MatVQA reveals substantial gaps in current multimodal reasoning capabilities. The MatVQA benchmark is publicly available[1] to facilitate further research on applying MLLMs to complex materials science problems.

1 Introduction

Multimodal Large Language Models (MLLMs) have recently demonstrated remarkable success in a variety of applications, including natural language understanding (Bedi et al., 2024; Brown et al., 2020; Hollmann et al., 2025), code generation (Jain et al., 2024; Nam et al., 2024), and diverse scientific domains (Latif et al., 2024; Song et al., 2023; Zhang et al., 2024b,c). Their ability to interpret and reason over complex, multimodal data while integrating specialist domain knowledge is pivotal for tackling challenging problems in fields such as mathematics, medicine, finance, and geospatial sciences. While MLLMs have shown promise in facilitating scientific discovery in areas like biology (Jung et al., 2024; Luu & Buehler, 2024), chemistry (Jablonka et al., 2023; Zhang et al., 2024a; Li et al., 2026), software engineering (Belzner et al., 2023), and healthcare (Peng et al., 2023; Nazi & Peng, 2024), their application to research-level reasoning in highly specialized, multimodal scientific tasks, particularly within materials science, remains less explored.

Materials science, an interdisciplinary field at the intersection of physics, chemistry, and often biology, demands extensive domain knowledge and sophisticated reasoning capabilities to understand and design material. Initial efforts have applied Large Language Models (LLMs) to augment materials science research, such as the HoneyComb agent for material-related question answering (Zhang et al., 2024b). However, these approaches, along with existing materials science datasets like MaScQA (Zaki et al., 2024), SciQA (Auer et al., 2023), and MSE-MCQs (Wang et al., 2025), predominantly focus on text-based question answering. Consequently, they often overlook the critical aspect of visual perception and lack questions that require deep, research-level reasoning. The recent emergence of benchmarks like MicroVQA (Burgess et al., 2025) for biological microscopy underscores the necessity of evaluating MLLMs on research-grade multimodal tasks. To fill this void, we introduce MatVQA. MatVQA is distinguished by its design for easy scalability and its specific focus on assessing two capabilities largely underaddressed by existing benchmarks: the complex, multi-step reasoning constructed from fundamental scientific principles, and the demanding, low-level visual perception required to interpret nuanced experimental data.

[1] https://huggingface.co/datasets/trqcbf/matvqa_v2

Findings of the Association for Computational Linguistics: ACL 2026, pages 23599–23614
July 2-7, 2026 ©2026 Association for Computational Linguistics

Table 1: Overview of multimodal science benchmarks and detailed MatVQA attributes (in (b), SPP represents **structure–property–performance**).

(a) Comparison with current multimodal benchmarks for Science

Benchmark	Level	Domain	Source	Size
ScienceQA	High-school	Science	Exams	16.8k
MicroBench	Graduate	Microscopy	Datasets	17.2k
MMSci	PhD	Science	Figures	7,132
MacBench	Research	Chem&Mat	Lab	628
LabBench	Research	Biology	WebQA	181
MicroVQA	Research	Microscopy	Expert	1,042
MatVQA (Ours)	Research	Materials	Paper	1,672

(b) MatVQA benchmark attributes

MatVQA feature	**Value**
Total questions	1,672
Causal SPP questions	950
Comp SPP questions	112
Hypo SPP questions	256
Quan SPP questions	354
Unique images	731
Unique papers	120
Research areas	32

MatVQA is built by **MArxivAgent**, an automated and verifiable pipeline engineered for the efficient generation of challenging multiple-choice questions (MCQs) directly from arXiv materials-science papers. This automated construction underpins MatVQA's inherent scalability, facilitating ongoing expansion and adaptation. After generation by advanced LLMs, 50% of the MCQs are vetted by domain experts to ensure quality. Grounded in real-world literature, the benchmark targets four research-critical structure–property–performance (SPP) tasks that map how a material's structure governs its measurable properties and ultimate performance. These tasks—Quantitative SPP, Comparative SPP, Causal SPP, and Hypothetical Variation—are derived from core components of scientific inquiry and by their very nature necessitate the complex reasoning MatVQA aims to evaluate. Solving them requires a tight integration of visual and textual evidence, collectively probing the core cognitive operations of materials research.

These tasks mirror the questions materials scientists pose when characterising, designing, and optimising materials, and they rigorously test an MLLM's ability to perform fine-grained visual perception, numerical reasoning, comparative analysis, causal inference, and forward prediction. As shown in Table 1, MatVQA contains 1672 multiple-choice questions automatically generated by advanced language models and randomly verified by domain experts carefully.

Motivated by the forensic study of multimodal biases in (Burgess et al., 2025), we identify two textual artifacts that can subvert genuine vision–language evaluation: **(i) Language shortcuts** appear when the prompt provides verbose image descriptions, weak distractors, or stylistic cues that allow the answer to be inferred from the text alone. **(ii) Caption shortcuts** happen when the information embedded in its stem or options—typically paraphrased from the figure caption—suffices to answer the question without inspecting the figure itself. Caption shortcuts are especially prevalent in materials-science corpora because captions explicitly condense the key morphological, crystallographic, or spectroscopic observations. Their inadvertent reuse limits MCQs to high-level descriptors (e.g., phase label, mean grain size) and excludes the low-level visual cues—diffraction peaks, defect textures, subtle contrast variations—crucial for authentic reasoning. Additional leakage stems from distractors generated with caption vocabulary, reliance on caption-embedded numerical values, and minimal emphasis on spatial or pixel-level patterns.

Eliminating both shortcut classes is therefore imperative: it compels models to ground their answers in fine-grained visual evidence and provides a benchmark that more faithfully measures genuine vision–language competence.

To excise these artifacts, MArxivAgent executes an *iterative shortcut-elimination loop*. After initial question synthesis, an evaluator agent answers the MCQ using (a) only the stem and options and (b) the stem, options plus caption but *without* the image. Success in either mode triggers a rewriter that removes or rephrases the incriminating text while a consistency checker enforces fidelity to the original scientific claim. By progressively eliminating both language and caption shortcuts, we elevate the benchmark from coarse- to fine-grained difficulty: solving the final questions requires precise, low-level visual scrutiny (e.g., counting diffraction spots, discerning lattice fringes) coupled with multi-hop scientific reasoning. This refinement is essential for measuring the true multimodal competence that front-line materials research demands.

We evaluated a suite of 19 open- and closed-source MLLMs on MatVQA and compared the performance of a select subset against human experts and vision-language model baselines. In summary, our key contributions are:

- We release **MatVQA**, the first benchmark designed to evaluate research-level multimodal reasoning for materials science.
- We propose four Structure-Property-Performance (SPP) tasks that encapsulate core scientific inquiries regarding material structure, properties, and performance.
- We design **MArxivAgent**, a fully automated, three-stage pipeline that (i) extracts reasoning paths from scientific literature, (ii) iteratively eliminates language shortcuts, and (iii) subsequently removes caption shortcuts, producing high-difficulty, visually grounded MCQs.

2 Related Work

2.1 MLLM Reasoning Benchmarks

Recent work has introduced a range of benchmarks that probe how well multimodal large language models (MLLMs) integrate visual and textual reasoning. MATHVISTA (Lu et al., 2024) tests fine-grained visual understanding and compositional math reasoning with thousands of expert-designed problems that expose the gap between today's models and human mathematicians. Similarly, benchmarks focusing on code reasoning, such as CRUXEval (Gu et al., 2024) evaluates input–output prediction for Python functions, CodeMMLU (Nguyen et al., 2025) measures code comprehension across multiple languages and domains, and CRQBench (Dinella et al., 2024) derives reasoning questions from real-world code reviews. There are also more MLLM benchmarks contributed to various domains(Talmor et al., 2021; Zou et al., 2024; Chow et al., 2025; Guo et al., 2024; Lozano et al., 2024). Furthermore, EMMA BENCH (Hao et al., 2025) (Enhanced MultiModal reAsoning) targets organic multimodal reasoning across mathematics, physics, chemistry, and coding. EMMA tasks require advanced cross-modal reasoning that cannot be solved by considering each modality independently, thus providing a challenging test suite for MLLMs.

2.2 Material Science Benchmarks

LLM4Mat-Bench (Rubungo et al., 2024) is presented as the largest benchmark to date for evaluating the performance of large language models (LLMs) in predicting the properties of crystalline

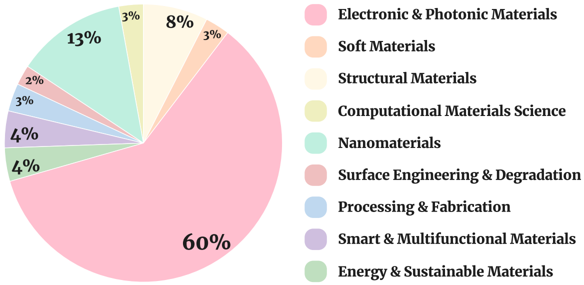


Figure 1: Domain attribution for **MatVQA**

materials. The benchmark includes a vast dataset of approximately 1.9 million crystal structures from ten public databases, covering 45 distinct material properties. ALDbench (Yanguas-Gil et al., 2024), a new benchmark specifically designed to assess the capabilities of large language models (LLMs) in the domain of materials synthesis, consists of 70 open-ended questions, ranging from graduate to expert level, covering various aspects of ALD. MicroVQA (Burgess et al., 2025) introduce a visual question answering benchmark in the field of biological microscopy. The benchmark is designed to evaluate three key reasoning capabilities crucial for scientific research: expert image understanding, hypothesis generation, and experiment proposal. While prior efforts such as MicroVQA relied entirely on expert-crafted items, MatVQA achieves comparable scientific rigor through automated generation followed by large-scale human validation, ensuring both scalability and reliability.

3 The MatVQA benchmark

3.1 Overview of MatVQA Benchmark

MatVQA is a fully synthetic dataset of 1672 VQA triplets. The questions cover a broad spectrum in material science, ranging from metals, ceramics, electronic materials to coating. Fig 1 shows the distribution of MatVQA samples. Especially, MatVQA closes three critical gaps in MLLMs in material science research:

(1) **Domain coverage.** Existing MLLM benchmarks omit the complex research-grade tasks characteristic of materials science. MatVQA formalizes four *structure–property–performance* (SPP) tasks spanning microstructural analysis, mechanistic interpretation, and forward design-core capabilities for material experimentation and modeling.

(2) **Scientific updated.** Unlike datasets based on legacy exams, MatVQA draws its questions from the latest arXiv manuscripts, ensuring

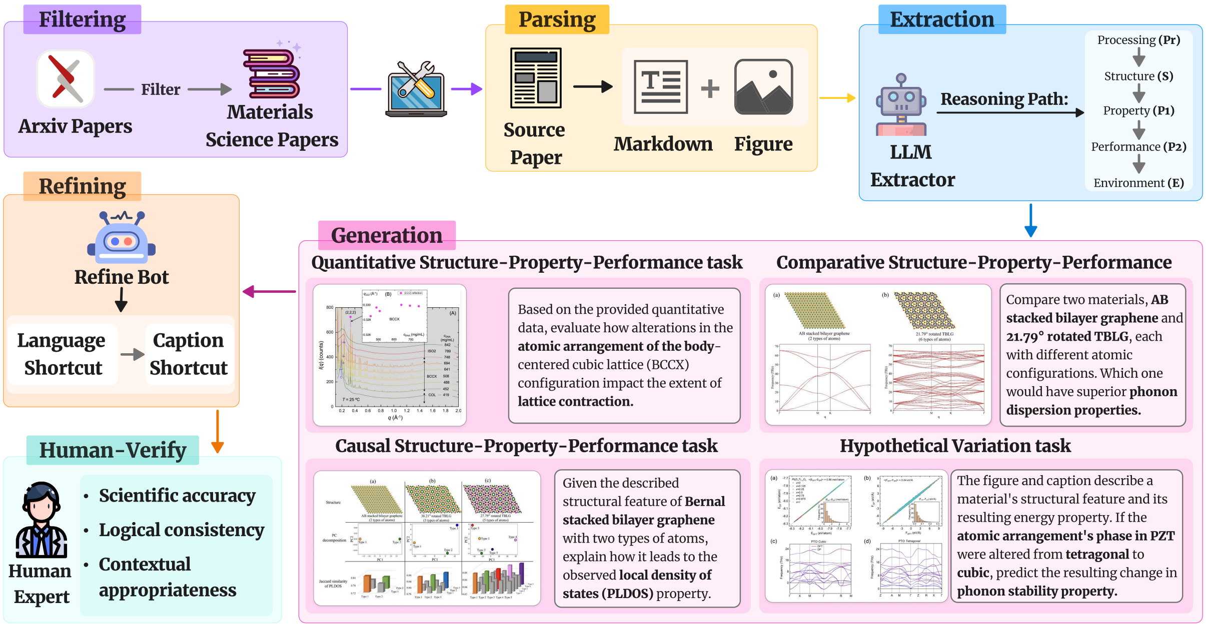


Figure 2: Construction Pipeline of MatVQA

topical relevance and cutting-edge difficulty.

(3) **Scalability with quality.** The **MArxivAgent** pipeline automatically extracts reasoning chains, generates high-fidelity MCQs, and removes language and caption shortcuts. 50% of the dataset was human-verified by two domain experts with field-specific assignments.

These innovations create the first benchmark capable of measuring an MLLM's aptitude for precise, fine-grained visual reasoning across the full breadth of materials-science inquiry.

3.2 Scientific Reasoning Tasks in MatVQA

As depicted in Figure 2, we design four high-impact tasks around five core components in material science—Structure (S), Property (P), Performance (Pe), Processing (Pr), and Environment (E):

Quantitative SPP. Quantifies how structural parameters (e.g., grain size, defect density) affect performance metrics such as strength or ductility.

Comparative SPP. Analyzes materials with different structural architectures (e.g., framework and layered forms) to reveal how variations influence the balance among material properties.

Causal SPP. Traces causality between material components, connecting a specific step in the manufacturing process (like adding a certain chemical) to the resulting structure and overall performance.

Hypothetical Variation. Conducts "what-if" analyses of unrealised structural variants, guiding exploration of topological insulators, superconductors, and metamaterials.

Covering structural, energy, electronic, photonic, separation, catalytic, and environmental applications, these tasks integrate data science, high-throughput computation, and advanced characterization. Developed through expert interviews, representative examples are provided in the Appendix.

4 MArxivAgent: MatVQA Data-Generation Process

This section details the pipeline used to create multiple-choice VQA items for MatVQA. As shown in Figure 3, the workflow has two main phases: (i) automated generation of raw visual question–answer triplets with verifiable reasoning chains and (ii) iterative shortcut removal followed by expert validation. Creating VQA datasets is traditionally labour-intensive, for example, MicroVQA required 12 specialists and more than 600h. To eliminate this bottleneck we developed **MArxivAgent**, a fully automated system that yields research-grade, reasoning-centric MCQs for materials science.

4.1 Phase 1: Generation of Raw VQA

Data Source. MatVQA builds on the arXiv open-access platform[2], leveraging granted remark and republication rights. We retrieved 500 open-access materials-science papers from arXiv (2024) to ensure transparency and verifiability. We selected arXiv because of its openness, early dissem-

[2] https://arxiv.org/

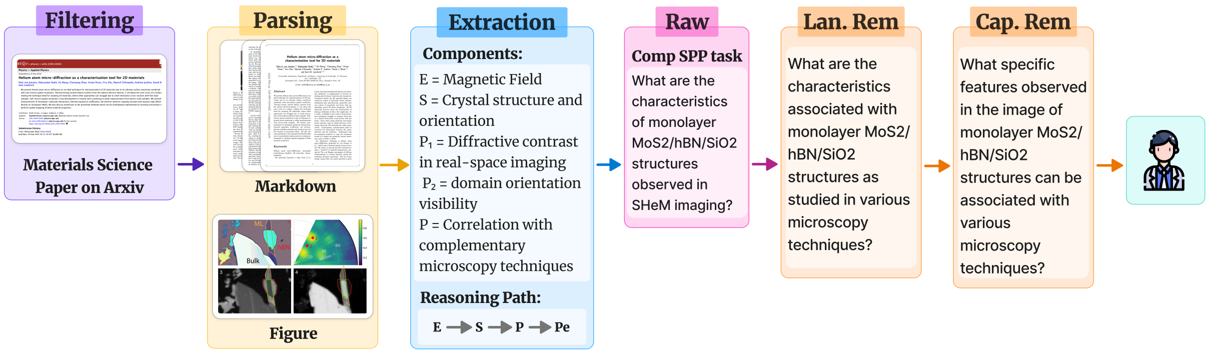


Figure 3: MArxivAgent Pipeline for MCQ automatically Generation. "Lan. Rem" represents the question after langauge shortcut removal. "Cap. Rem" represents the question after removing caption.

ination of state-of-the-art research, and compatibility with reproducible benchmarking. Next, we filtered this corpus using task-related keywords (e.g., "property", "structure"). Because most manuscripts lack accessible source files, we employed Marker[3] to extract figures and captions directly from the PDF collection. Beyond these visual elements, we also harvested the surrounding textual context for each figure—capturing extended explanations and analyses critical for complex reasoning.

Verifiable Reasoning Path Extractor. To generate questions that probe complex reasoning over scientific figures, MArxivAgent first derives the most comprehensive reasoning chain from the provided caption and its surrounding context. This chain is structured around five core components—Structure (S), Property (P), Performance (Pe), Processing (Pr), and Environment (E). To keep consistencey identification of these elements, we incorporate the MatOnto[4] in the components identification step. Each reasoning step is then cross-verified against the original paper text to confirm evidence alignment, ensuring the resulting chain is both accurate and verifiable.

Multiple Chioces Question Construction. Given a figure, its caption, and the verified reasoning path, the agent generates task-specific questions—causal, hypothetical, comparative, or quantitative—and rewrites them in MCQ format.

Overall, this stage produced 1672 multiple-choice items: 950 causal, 256 hypothetical, 112 comparative, and 354 quantitative.

4.2 Phase 2: Two-Stage Refinement

The generated raw MCQs is inadequate for testing MLLM's capabilities. On one hand, the distractors are easily eliminated based on general material knowledge or too vague compared to correct option. Language shortcuts-information in the MCQ that allows answering the question without access to the image as proved in (Burgess et al., 2025). On the other hand, we found that some questions can be directly answered by the figure caption without access to the figure. Complex reasoning required perception on the given image and further do multi-hop reasoning on the provided information. While the figure caption can only describe the main idea of this figure rather than demonstrate every details of the figure. We defined this kind of shortcut as caption shortcut. Therefore, we aim to construct questions removing these two shortcut to have high-quality MCQ distractors and questions required both fine-grained perception and multi-hop reasoning.

While our refinement process draws inspiration from prior work on iterative shortcut elimination, such as in MicroVQA [13], MArxivAgent pioneers a second-stage 'caption shortcut removal' module. This novel component is explicitly designed to compel low-level visual perception rather than reliance on high-level summaries in figure captions, ensuring our benchmark robustly evaluates visual reasoning instead of serving as a proxy for text comprehension. The complete two-stage refinement procedure is summarized in Appendix: Algorithm 1.

Stage1: Language shortcut removal Here we follow the method in MicroVQA to remove the language shortcut to increase MCQ complexity: first apply an evaluator agent to answer the MCQ without image and then summary the COT answer as langauge strategy which will be passed to the rewriter LLM agent. The rewriter agent revises the original question and distractors to invalidate the language strategies. To aviod the significant changes on the revised question-answering pairs, an LLM checker is applied to ensure semantic

[3]https://github.com/VikParuchuri/marker
[4]https://matportal.org/ontologies/MATONTO

equivalent with the original pair. This process will iterate if the question can still be answered correctly without image or after the pre-setting maximum iteration times.

Stage2: Caption shortcut removal After removing language shortcut in Stage 1, the questions still lack of fine-grained perception on figures since questions can be answered directly by the caption. Therefore, we introduce second stage refining MCQs by removing the caption shortcut. Similarly, we first use a evaluator agent to identify the caption shortcuts by answering the MCQs with caption only. And then reflecting on the strategies answering the question with a reflector LLM agent. The reflection results will further been passed to a rewriter to revise the question and options. Different from Stage 1, we only require the revised question-answering pairs follow the same reasoning chain used in Section 4.1, which results in larger modification on generated questions and aim to create harder problem removing the simple understanding pattern from captions.

An example of a generated reasoning chain is illustrated in Figure 3. In this sample case, provided with a paper from Arxiv. After parsing the paper pdf to markdown text and figures, we generate the reasoning path as: "10 T magnetic field (E) $\rightarrow$ collapse of the split magnetic Bragg satellites into a single peak (S) $\rightarrow$ suppression of the spin-cycloid and emergence of a collinear antiferromagnetic state (P) $\rightarrow$ redistribution of the total magnetic scattering into one commensurate reflection, doubling the peak intensity (Pe)." And for each reasoning step, we further verify it to grounded by the paper, which ensure the reliability of generated reasoning path.

4.3 Human Expert Quality Check

To ensure the scientific rigor and reliability of MatVQA, two university-affiliated postdoctoral researchers specializing in materials science were recruited to validate 50% of the dataset following a structured evaluation protocol. Each expert independently reviewed a distinct 20% subset, while an additional 10% overlap was double-annotated to measure inter-annotator agreement (IAA). The experts assessed each question for context validity, answer correctness, and reasoning fidelity. The evaluation achieved 86% overall agreement (Cohen's Kappa = 0.76 for answerability, 0.68 for correctness), indicating substantial reliability.

5 Experiments

5.1 Benchmarking MLLMs with MatVQA

To comprehensively evaluate the validity and difficulty of our proposed dataset, MatVQA, we conducted a systematic series of experiments across a range of state-of-the-art Multimodal Large Language Models (MLLMs) as shown in Table 2. The selected models encompass both open-source and proprietary systems and represent leading capabilities in visual-and-language understanding, including Grok-4(xAI, 2025) and LLaVA(met), as well as the closed-source GPT-4o(OpenAI, 2024) and Gemini(AI, 2025). We utilize standard chain-of-thought prompting(Yue et al., 2025)(details in Appendix). By comparing these models' performance on the MatVQA benchmark, we aim to elucidate the current strengths, limitations, and avenues for improvement in multimodal material visual-reasoning question answering problems.

MatVQA is uniformly challenging – The highest overall performance is achieved by CLAUDE-3.7-SONNET (52.5%) from the "large models" category, followed closely by MISTRAL-SMALL-3.2-24B (51.8%). The reasoning model o1 shows a strong baseline performance at 50.7%. In the small model category, QWEN3-VL-8B-INSTRUCT (48.3%) and GEMINI-FLASH-2.0 (46.1%) were the top performers, demonstrating that optimized smaller architectures can achieve competitive results against larger counterparts. The domain-specific fine-tuned models struggle significantly; CEPHALO-8B-ALPHA shows the lowest overall performance at 25.2%, while MOL-VL-7B achieves only 29.7%. The poor performance of these material-finetuned models might stem from their training focus on optical chemical structure recognition, which biases outputs for our heterogeneous data. The uniformly low accuracy across the board proves that MatVQA is challenging for both general-purpose and specialized models. These limitations likely stem from various factors, including the nuanced visual perception required for material-scientific figures and the sophisticated reasoning demanded by tasks such as comparative and hypothetical analysis.

Scale helps, but only up to a point – As shown in Table 2, large-parameter models generally outperform their smaller counterparts, particularly on the dominant Causal split. However, model size is no guarantee of success: the 90B LLAMA-3.2-90B-VISION (44.0%) and the multimodal PHI-4

Table 2: Performance of various vision–language models on MatVQA by task: Causal Structure-Property-Performance task(Caus),Comparative Structure-Property-Performance task(Comp), Quantitative Structure-Property-Performance task(Quan),Hypothetical Variation task(Hypo).The evaluated models are splited to four parts: Reasoning mode, Large models, Small models and Material-finetuned model(Material-FT Models).

Category	Model	Overall	Caus	Hypo	Quan	Comp
Reasoning	o1	50.7%	49.7%	48.0%	59.0%	39.3%
Large Models	Claude-3.7-Sonnet	**52.5**%	**53.1**%	42.5%	**59.6**%	48.2%
	Gemini-2.5-Pro	50.7%	49.3%	48.0%	58.2%	44.6%
	Mistral-small-3.2-24b-instruct	51.8%	49.4%	**52.3**%	60.7%	42.9%
	Grok-4	46.6%	42.7%	45.7%	59.3%	41.1%
	Qwen-2.5-VL-72b-Instruct	47.1%	46.0%	39.8%	57.1%	41.9%
	Qwen3-VL-32b-Instruct	45.6%	42.7%	48.8%	50.6%	47.3%
	GPT-4o	49.7%	50.0%	42.5%	55.6%	45.5%
	Llama-3.2-90b-Vision-Instruct	44%	40.0%	32.8%	62.1%	46.4%
	Phi-4-Multimodal-Instruct	31.5%	27.2%	30.1%	45.2%	27.6%
Small Models	Qwen-2.5-VL-7b-Instruct	42.4%	44.0%	38.7%	41.5%	40.1%
	Qwen3-VL-8b-Instruct	**48.3**%	**46.9**%	**48.4**%	52.8%	45.5%
	Claude-3.5-Haiku	39.5%	32.9%	38.3%	**58.5**%	37.5%
	Gemini-Flash-2.0	46.1%	44.4%	42.9%	50.5%	**53.5**%
	GPT-4o-mini	40.6%	41.1%	34.3%	46.3%	33.0%
	Pixtral-12b	44.0%	43.8%	35.2%	53.1%	36.6%
	Llama-3.2-11b-vision-instruct	32.3%	32.9%	28.1%	32.2%	36.6%
Material-FT Models	MOL-VL-7B	29.7%	32.9%	22.7%	34.6%	2.8%
	Cephalo-8B-Alpha	25.2%	22.4%	25.4%	30.8%	31.4%

(31.5%) lag behind the 8B QWEN3-VL (48.3%). These outliers illustrate that clever architectural design and data quality can outweigh brute-force scaling. Notably, QWEN3-VL-8B outperforms several significantly larger models, including GROK-4 and QWEN-2.5-VL-72B. In short, while bigger models generally deliver a performance premium on MatVQA, the marginal gains taper off. A likely reason is that MatVQA requires fine-grained, domain-specific cross-modal reasoning about materials, which depends less on parameter count than on how effectively the vision and language components are aligned.

Difficulty varies strongly by task – Averaging across size tiers, large-parameter models score 44.5% on the Causal split, 39.7% on Hypothetical, 42.7% on Comparative, and 56.5% on Quantitative split, while small models reach 39.9%, 36.3%, 39.6%, and 47.8% respectively. Causal reasoning is both the most fundamental skill for materials science and the backbone of the benchmark (950 of 1672 items), so modest gains here translate directly into meaningful real-world impact. Comparative questions, though not as numerous, record the worst performance; they demand precise joint perception of two entities and consistency across several reasoning hops, exposing brittleness in current vision–language pipelines.

These results collectively demonstrate that MatVQA provides a stringent and well-balanced yard-stick: it highlights genuine progress in multimodal scientific QA while exposing persistent weaknesses—particularly in counter-factual design reasoning—that future MLLMs must overcome.

Human Performance - To assess the intrinsic difficulty of MatVQA, we conducted a human baseline study with two materials-science domain experts. The experts achieved an average accuracy of 28% on randomly selected 100 samples, confirming that MatVQA tasks require deep multimodal reasoning beyond surface-level cues. This result empirically validates that MatVQA is highly challenging even for skilled practitioners.

5.2 Ablation for Two-Stage Refinement

The results in Table 3 demonstrate that both language and caption refinement stages significantly reduce model accuracy, compelling a deeper level

Table 3: Two-stage Refinement evaluation results. "Lan.Rem" represents the results after removing language shortcut and "Cap.Rem" represents the results after removing caption shortcut.

Stage	GPT-4o	GPT-4o-mini	Claude-3.7-Sonnet	Claude-3.5-Haiku	o1	Gemini-2.5-Pro
Raw	71.5%	60.7%	75.9%	56.9%	73.5%	71.6%
Lan.Rem	63.6% (**11.0%**↓)	51.6% (**15.0%**↓)	67.4% (**11.1%**↓)	51.1% (**10.3%**↓)	65.2% (**11.3%**↓)	62.9% (**12.2%**↓)
Cap.Rem	49.7% (**30.5%**↓)	40.6% (**33.1%**↓)	52.5% (**30.8%**↓)	39.5% (**30.6%**↓)	50.7% (**31.0%**↓)	50.7% (**29.2%**↓)

of reasoning. Language shortcut removal (Stage 1, or Lan.Rem) decreases accuracy by approximately 12% on average (individual model drops typically range 10%–15%). As illustrated by the sample question evolution in Appendix: Figure 4, this initial difficulty increase stems from broadening the question's scope from specific details (e.g., 'phenol') to more generalized concepts (e.g., 'aromatic substances') and removing overly assertive or simplistic textual cues. This demands more nuanced language comprehension from the models. Scientifically, this stage appropriately challenges models, as overgeneralizing concepts without full contextual understanding can lead to misinterpretations in many real-world phenomena.

Caption shortcut removal instigates a more substantial performance decline, with an additional average accuracy drop of approximately 19% from language removal (individual model drops typically range 18%–20%). By compelling models to value low-level visual perception from experimental figures—such as the "patterns and arrangements of the clusters" shown in the example—instead of relying on textual hints, Caption removal ensures a rigorous test of genuine visual grounding and fine-grained perception. This is a crucial step towards evaluating true multimodal understanding.

5.3 Error Analysis

To further explore the current shortcomings of MLLMs, we conducted a manual audit of 40 error samples produced by the 01 model. A materials-science expert spent approximately 50 minutes per item reviewing the source paper and the model's chain-of-thought response. The distribution of error types was as follows:

Visual Perception Error - 30.0%: The model incorrectly interprets information from the figure itself (e.g., reversing a plotted trend, overlooking key features, or confusing image panels).

Material Knowledge Misunderstanding - 47.5%: The model's chain-of-thought is logically coherent but rests on a faulty scientific premise (e.g., making an incorrect assumption about material behavior). This was the predominant failure mode.

Reasoning Wrong Judgement - 22.5%: The model's logic falters after correctly perceiving the visual data (e.g., misinterpreting an option's wording or applying an inconsistent elimination rule). Specific examples for each error category are provided in the Appendix A.6. The prevalence of knowledge-based misunderstandings suggests that future improvements for MLLMs in specialized domains will require deeper integration of domain-specific knowledge, beyond enhancing general visual and logical capabilities.

6 Conclusion

We presented MatVQA, a benchmark that marks a significant step towards evaluating true multimodal reasoning in materials science. MatVQA compels Multimodal Large Language Models (MLLMs) to genuinely "look" at complex material experimental figures, engaging with fine-grained visual details rather than relying on textual shortcuts. This is achieved through our fully automated MArxivAgent pipeline—a robust and adaptable methodology with strong potential for application across diverse scientific fields beyond materials science. This automated approach not only produces an initial set of 1672 high-fidelity questions grounded in visual evidence across four key structure-property-performance reasoning tasks (quantitative, comparative, causal, and hypothetical) but also ensures MatVQA's inherent scalability.

Limitations

While the proposed data generation process, MArxivAgent, is fully automated and highly scalable, we constrained the current MatVQA dataset to 1,672 samples due to API costs and the load required for human verification. Although MatVQA is currently the largest research-level multimodal science benchmark, extending the dataset size would further benefit the scientific domain.

Furthermore, although human experts quality-checked 50% of the MatVQA dataset using a structured evaluation protocol and achieved high inter-annotator agreement, the data quality still entails a degree of subjectivity. Broader validation from additional domain experts would further enhance the dataset's reliability.

Ethical Considerations

During the development of MatVQA, we prioritized the ethical and responsible use of data. The current MatVQA dataset is constructed exclusively from publicly available sources, specifically preprints from arXiv. Additionally, the released version of MatVQA is intended solely for research purposes, governed by a license that prohibits commercial use or redistribution of the source data.

Acknowledgements

This work is supported by the Canada Mitacs Accelerate Program (IT44046) and the Canada NSERC Discovery Grant (RGPIN-2021-03115).

References

Google AI. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, Mar 2025.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240, 2023.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*, 333(4):319, 2024.

Lenz Belzner, Thomas Gabor, and Martin Wirsing. Large language model assisted software engineering: prospects, challenges, and a case study. In *International Conference on Bridging the Gap between AI and Reality*, pp. 355–374. Springer, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19552–19564, 2025.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding, 2025. URL https://arxiv.org/abs/2501.16411.

Elizabeth Dinella, Satish Chandra, and Petros Maniatis. Crqbench: A benchmark of code reasoning questions, 2024. URL https://arxiv.org/abs/2408.08453.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. Cruxeval: A benchmark for code reasoning, understanding and execution, 2024. URL https://arxiv.org/abs/2401.03065.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems*, 37:134721–134746, 2024.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark, 2025. URL https://arxiv.org/abs/2501.05444.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital discovery*, 2(5):1233–1250, 2023.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Sung Jae Jung, Hajung Kim, and Kyoung Sang Jang. Llm based biological named entity recognition from scientific literature. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 433–435. IEEE, 2024.

Ehsan Latif, Ramviyas Parasuraman, and Xiaoming Zhai. Physicsassistant: An llm-powered interactive learning robot for physics lab investigations. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pp. 864–871. IEEE, 2024.

Yizhan Li, Florence Cloutier, Sifan Wu, Ali Parviz, Boris Knyazev, Yan Zhang, Glen Berseth, and Bang Liu. Mˆ 4olgen: Multi-agent, multi-stage molecular generation under precise multi-property constraints. *arXiv preprint arXiv:2601.10131*, 2026.

Alejandro Lozano, Jeffrey Nirschl, James Burgess, Sanket Rajan Gupte, Yuhui Zhang, Alyssa Unell, and Serena Yeung-Levy. Micro-bench: A microscopy benchmark for vision-language understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 30670–30685. Curran Associates, Inc., 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

Rachel K Luu and Markus J Buehler. Bioinspiredllm: Conversational large language model for the mechanics of biological and bio-inspired materials. *Advanced Science*, 11(10):2306724, 2024.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pp. 1–13, 2024.

Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, pp. 57. MDPI, 2024.

Dung Manh Nguyen, Thang Chau Phan, Nam Le Hai, Tien-Thong Doan, Nam V Nguyen, Quang Pham, and Nghi DQ Bui. Codemmlu: A multi-task benchmark for assessing code understanding & reasoning capabilities of codellms. In *The Thirteenth International Conference on Learning Representations*, 2025.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1): 210, 2023.

Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. Llm4mat-bench: benchmarking large language models for materials property prediction. *arXiv preprint arXiv:2411.00177*, 2024.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. Honeybee: Progressive instruction finetuning of large language models for materials science. *arXiv preprint arXiv:2310.08511*, 2023.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images, 2021.

Hongchen Wang, Kangming Li, Scott Ramsay, Yao Fehlis, Edward Kim, and Jason Hattrick-Simpers. Evaluating the performance and robustness of llms in materials science q&a and property predictions. *Digital Discovery*, 4(6):1612–1624, 2025.

xAI. Grok 4 model card, 2025. URL https://data.x.ai/2025-08-20-grok-4-model-card.pdf.

Angel Yanguas-Gil, Matthew T. Dearing, Jeffrey W. Elam, Jessica C. Jones, Sungjoon Kim, Adnan Mohammad, Chi Thang Nguyen, and Bratin Sengupta. Benchmarking large language models for materials synthesis: the case of atomic layer deposition, 2024. URL https://arxiv.org/abs/2412.10477.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. pp. 15134–15186, 2025.

Mohd Zaki, NM Anoop Krishnan, et al. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024a.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135*, 2024b.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024c.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models, 2024.

A Appendix

A.1 Algorithm for Two-stage Refinement

Algorithm 1 Two-stage Refinement

Require: Raw MCQ set $Q = \{(q_i, C_i)\}$, where each C_i comprises the correct answer A_i and distractors D_i, maximum iteration count T

Ensure: A refined MCQ set Q'

$Q' \leftarrow Q$
Stage 1: Language Shortcut Removal
for each (q, A, D) in Q' **do**
  $t \leftarrow 0$
  while $t < T$ **do**
    $r \leftarrow \text{Evaluator}(q, C)$
    if not r.correct **then**
      break
    end if
    $s \leftarrow \text{ExtractLangStrategy}(r)$
    $(q^{'}, C^{'}) \leftarrow \text{Rewriter}(q, C, s)$
    $v \leftarrow \text{Checker}(q, A; q^{'}, A^{'})$
    if semantically_inconsistent **then**
      FAIL_REWRITE and **continue**
    end if
    $t \leftarrow t + 1$
  end while
end for
Stage 2: Caption Shortcut Removal
for each (q, C) in Q' **do**
  $t \leftarrow 0$
  while $t < T$ **do**
    $r \leftarrow \text{Evaluator}(q, C, captions)$
    if not r.correct **then**
      break
    end if
    $s \leftarrow \text{Reflector}(r)$
    $(q^{'}, C^{'}) \leftarrow \text{Rewriter}(q, C, s)$
    $v \leftarrow \text{Checker\_caption}(q, A; q^{'}, A^{'})$
    if reasoning_path_inconsistent **then**
      FAIL_REWRITE and **continue**
    end if
    $t \leftarrow t + 1$
  end while
end for
return Q'

A.2 Prompts For MArxivAgent

All prompts for evaluation used the chain-of-thought prompt template from the MMMU-Pro code(Yue et al., 2025):

```
The following is a multiple choice question (with answers).
Think step by step and then output the answer in the format of \ "The answer is (X) \ " at the end.

{ { QUESTION } }

Options:
{ { CHOICES } }
```

An example complete question is:

```
The following is a multiple choice question (with answers).
Think step by step and then output the answer in the format of \ "The answer is (X) \ " at the end.

How does the observed binding energy affect the catalytic performance for ammonia decomposition?

Options:
    (1): Higher binding energy from Fe exposure enhances reaction efficiency by breaking ammonia bonds more effectively.
    (2): Increased binding energy improves adsorption efficiency, enhancing ammonia decomposition on the catalyst surface.
    (3): High binding energy indicates a reactive surface, facilitating rapid ammonia decomposition by destabilizing NH3.
    (4): Lower binding energy suggests structural stability, maintaining effective catalytic performance over various temperatures.
```

Candidate answers are extracted with the regular expression `answer is\s*([0-9])`.

A.3 Analysis of Overfitting Risk

A potential concern is that, given the dataset is sourced from 44 unique papers, models might achieve high performance by memorizing these sources rather than engaging in genuine reasoning. To investigate whether MatVQA evaluates reasoning over retrieval, we conducted two experiments.

Experiment 1: Fine-tuning on Source Papers. We fine-tuned the Qwen-2.5-VL-7B-Instruct model on a figure-captioning task using all figures from the 44 source papers. After fine-tuning, the model's accuracy on MatVQA increased trivially from 42.6% to 43.18%, a gain of only 0.42%. This indicates that superficial exposure to the source materials is insufficient for mastering the multi-step reasoning required by our benchmark.

Experiment 2: Providing Full Context. We provided the full text of the source paper as context to the GPT-4o model during inference. This increased performance from 48.3% to 55.0%. While access to the full context is beneficial, the performance remains modest and is far from saturated, still lagging significantly behind the 86% inter-expert agreement on answer correctness.

Together, these results strongly suggest that MatVQA questions cannot be answered by simple retrieval or overfitting to the source documents. They effectively probe deep contextual and visual reasoning capabilities.

A.4 Question Evolution Through Two-stage shortcut removal process.

The evolution of a sample question through the two-stage shortcut removal process shows in Figure 4.

A.5 Representative Examples

Here show some representative examples of MatVQA for varies material science domain in Figure 5.

A.6 Failure Analysis

Several error examples are shown in Figure 6 and Figure 7.

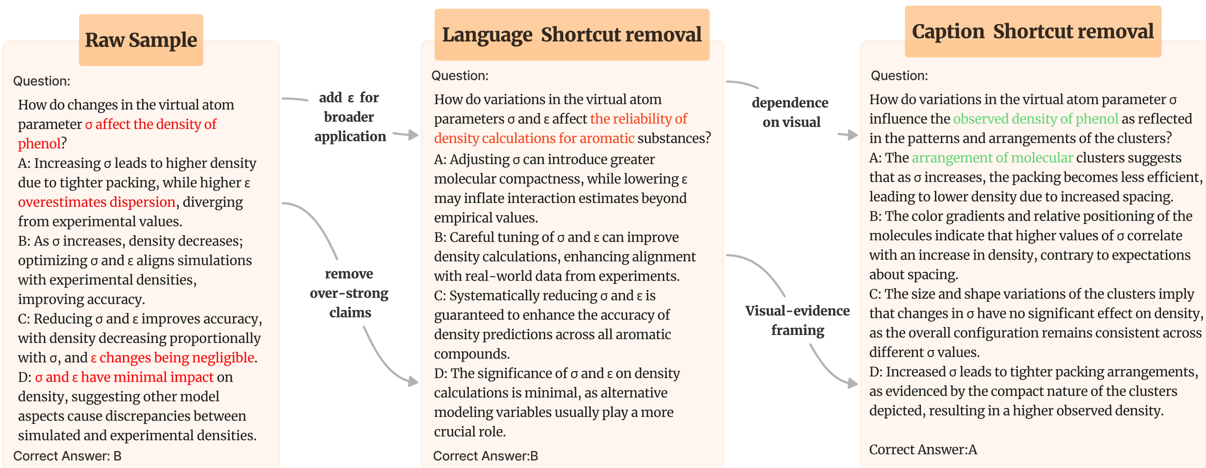
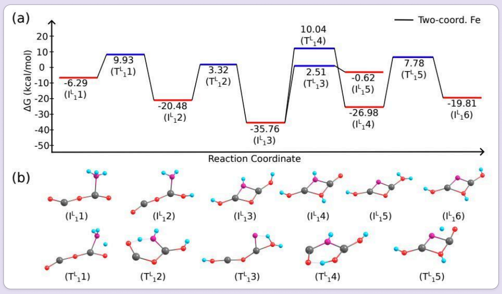


Figure 4: Evolution of a sample question through the two-stage shortcut removal process. The figure shows the transformation from: the initial 'Raw Sample,' to after 'Language Shortcut removal', and finally to after 'Caption Shortcut removal'.

Energy & Sustainable Materials

In what ways do the specific atomic arrangements, as seen in the Fe\(_2\)O\(_3\) linear-type isomer, affect the visible characteristics and reactive behavior during the NH\(_3\) decomposition?

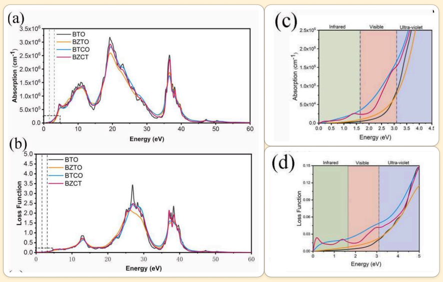
(1): The atomic arrangement leads to prominent variations in the surface texture, which correlate with distinct reactive sites influencing NH\\(_3\\) decomposition rates.
(2)The spatial distribution of atoms results in observable changes in hue, indicating the presence of intermediates that facilitate the NH\\(_3\\) decomposition process.
(3)The specific alignment of atoms creates unique energy state transitions that are visually apparent, altering the efficiency of NH\\(_3\\) decomposition.
(4)Differences in atomic positioning produce contrasting surface patterns that can be linked to different energy profiles during NH\\(_3\\) decomposition.



Electronic & Photonic Materials

What specific changes can be observed in the absorption spectrum of BTO with increased doping concentration of Zn and Co, based on the details present in the image?

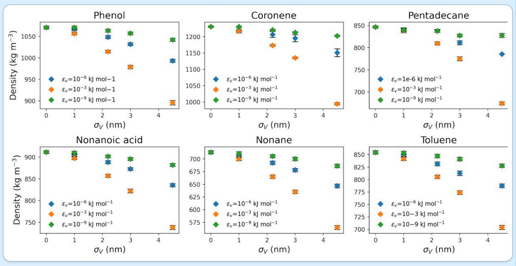
(1): The peaks in the absorption spectrum become narrower and sharper, indicating enhanced selectivity for certain wavelengths as the doping concentration increases.
(2): The shape of the absorption spectrum transforms into a series of broad, overlapping peaks that obscure distinct transitions between absorption features.
(3): The color representation of the peaks shifts distinctly to indicate a preference for absorbing in the visible region rather than the ultraviolet spectrum with increased doping.
(4): There is a noticeable displacement of some peaks towards lower energy, while others appear as flattened bands, demonstrating a unique pattern of changes related to high doping levels.



Electronic & Photonic Materials

How do variations in the virtual atom parameter σ influence the observed density of phenol as reflected in the patterns and arrangements of the clusters?

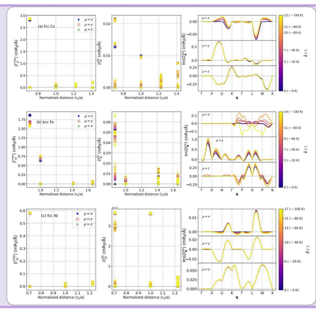
(1): The arrangement of molecular clusters suggests that as σ increases, the packing becomes less efficient, leading to lower density due to increased spacing.
(2): The color gradients and relative positioning of the molecules indicate that higher values of σ correlate with an increase in density, contrary to expectations about spacing.
(3): The size and shape variations of the clusters imply that changes in σ have no significant effect on density, as the overall configuration remains consistent across different σ values.
(4): Increased σ leads to tighter packing arrangements, as evidenced by the compact nature of the clusters depicted, resulting in a higher observed density.'



Energy & Sustainable Materials

How do spin-lattice parameters depend on the polar angle in bcc Fe?

(1): The polar angle causes a uniform increase in exchange coupling, with bcc Fe showing weak temperature sensitivity due to its atomic mass.
(2): The polar angle affects orbital overlap, making fcc Co and Ni more temperature-sensitive, while bcc Fe remains stable due to its lattice symmetry.
(3): The polar angle causes significant exchange coupling variability in bcc Fe due to non-Heisenberg interactions, increasing its temperature sensitivity.
(4): The polar angle primarily affects magnetic anisotropy, with minimal impact on exchange coupling, leading to similar temperature sensitivities in bcc Fe, Co, and Ni.



Electronic & Photonic Materials

What is the impact of phonon replica energies on the optical characteristics of hexagonal boron nitride (hBN)?

(1): Phonon replica energies influence transmission rates in photodetectors, leading to enhanced signal clarity and precision.
(2): Broad energy distributions from phonon replicas enhance the resolution of emission peaks, improving optical performance.
(3): The presence of various phonon replica energies creates interference patterns that sharpen emission features and improve detection sensitivity.
(4): Certain phonon replica energies, including levels around 4.10 eV and 3.70 eV, contribute to spectral overlap that blurs emission peaks, degrading temporal resolution.

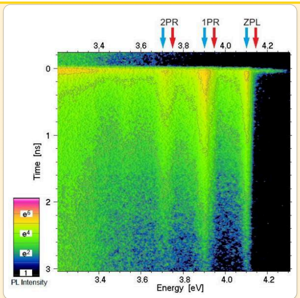
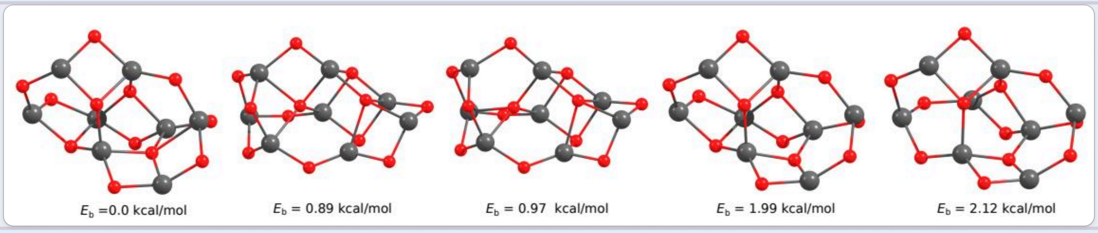


Figure 5: Representative Examples for varies material science domain.

Q: How does the observed binding energy affect the catalytic performance for ammonia decomposition?

(1): Higher binding energy from Fe exposure enhances reaction efficiency by breaking ammonia bonds more effectively.
(2): Increased binding energy improves adsorption efficiency, enhancing ammonia decomposition on the catalyst surface.
(3): High binding energy indicates a reactive surface, facilitating rapid ammonia decomposition by destabilizing NH_3.
(4):Lower binding energy suggests structural stability, maintaining effective catalytic performance over various temperatures.



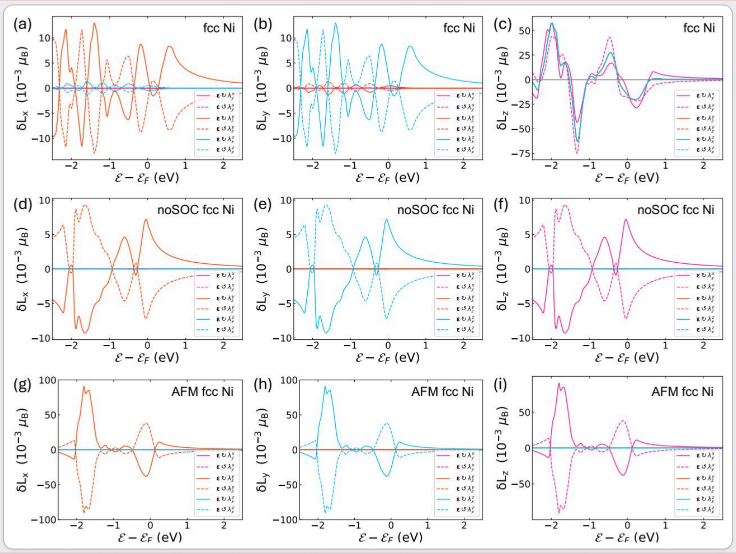
Llama-90b Response:To determine the effect of observed binding energy on the catalytic performance for ammonia decomposition, we need to understand how binding energy influences the reaction process. Binding energy in this context refers to the energy required to adsorb reactants (in this case, ammonia) onto the catalyst surface. * **Adsorption Process**: The adsorption of ammonia onto the catalyst surface is a crucial step in its decomposition. A higher binding energy generally indicates that ammonia molecules are more strongly attracted to the catalyst surface, which can enhance the efficiency of the decomposition process by ensuring that more reactant molecules are available for the reaction. * **Reaction Mechanism**: The decomposition of ammonia into nitrogen and hydrogen involves breaking the nitrogen-hydrogen bonds. Higher binding energy can also imply that the dissociation process is facilitated because the reactant is more strongly bound to the catalyst surface, making the breaking of bonds (to form products) more feasible. Given these considerations: * Option 1 suggests that higher binding energy enhances reaction efficiency by facilitating the breaking of ammonia bonds. This aligns with the general understanding that a stronger interaction between the reactant and the catalyst often benefits the reaction process. * Option 2 mentions improved adsorption efficiency due to increased binding energy, which supports better ammonia decomposition. This is consistent with the need for effective adsorption for the reaction to proceed efficiently. * Option 3 indicates that high binding energy signifies a reactive surface beneficial for destabilizing NH3 and enhancing decomposition. This perspective also aligns with the idea that stronger binding conditions can promote the reaction by weakening the ammonia bonds. * Option 4 implies that lower binding energy could suggest structural stability, maintaining effective catalytic performance across various temperatures. However, this option seems less directly related to how binding energy impacts the decomposition process compared to the other options. Considering these points, all options (1, 2, 3) correctly describe how the observed binding energy could positively affect the catalytic performance for ammonia decomposition by enhancing the adsorption and facilitating the bond-breaking process necessary for the reaction. However, the task requires selecting one option that best describes the effect without contradicting principles. Since the question seeks a singular best answer and all positive-effect descriptions were provided as options, the answer that encompasses the underlying principle most broadly applicable without specification would be: The answer is (2).

Error Reason: For ammonia decomposition over Fe-based catalysts, experimental spectra showing lower N 1s binding energies are interpreted as a sign that the surface nitrides are less tightly held, i.e., the structure is more stable and the catalytic cycle proceeds more readily over a temperature range.

Figure 6: Error sample for Material Knowledge Misunderstanding error type, the response missed catalytic know-how—namely the Sabatier optimum and the distinction between Fe–support binding vs. NH_3 adsorption

Q: In observing the variations in light-induced orbital δL^z magnetic moments in fcc Ni, what specific visual characteristics or patterns related to the textures, shapes, and color differences in the corresponding image indicate their impact on optical manipulation applications?

(1): The smooth gradients in color indicate uniform absorption levels across all magnetic configurations, suggesting minimal influence on optical manipulation.
(2): The different shapes and relative positions of peaks in the graph correlate with variations in orbital δL^z moments, showing enhanced control under specific light geometries.
(3): Textural variations observed indicate that ferromagnetic fcc Ni exhibits distinct light interaction patterns, enhancing performance in optical applications compared to non-relativistic configurations.
(4): The presence of unexpected patterns or irregularities in the image reflects the underlying spin-orbit coupling differences, emphasizing the importance of structural features in manipulating δL^z moments.



GPT-4o Response: To answer the question systematically: 1. **Analyze the images and their context**: The images display changes in orbital magnetic moments \(\delta L_x, \delta L_y, \delta L_z \) across various configurations (e.g., fcc Ni, noSOC fcc Ni, and AFM fcc Ni) as a function of energy relative to the Fermi level (\(E - E_f \)). 2. **Understand the visual information**: - Variations in **shapes, peaks, and patterns** across the plots indicate differences in the magnetic moments for different light-induced geometries. - The importance of **spin-orbit coupling (SOC)** can be inferred by comparing the "normal" fcc Ni (a-c) and noSOC fcc Ni (d-f) panels, as SOC-related features seem to be dominant in the former. - Textural and color-coded distinctions correspond to different magnetic projection contributions (\(\epsilon x \lambda^2\), etc.) impacting these variations. 3. **Relate findings to the options**: - Option (1): Incorrect. The gradients or absorption patterns are not "smooth" or uniform across configurations. - Option (2): Correct. The relative positions and shapes of peaks directly correlate with changes in \(\delta L_z \), showing distinctions for optical manipulation. - Option (3): Incorrect. While textural variations exist, they do not clearly suggest superiority in optical applications. - Option (4): Incorrect. Spin-orbit coupling differences are evident but do not reflect unexpected or irregular patterns based on the given data. The correct answer focuses on the relationship between variations and the impact on optical applications. The answer is (2).

Error Reason: The failure happened in the decision step: they judged "peak shape" to be the decisive cue and ignored the more salient texture/colour contrast that the question emphasised.

Figure 7: Error sample for Reasoning-Wrong-Judgement error type. In this example, optical control schemes exploit those extra SOC-enabled orbital channels; the richer the texture, the more knobs you have. Therefore the figure is telling us that ferromagnetic fcc Ni, with its unique textural signature, offers superior optical-manipulation capability compared with the non-relativistic case. The respondent saw the figure correctly (they talked about peaks and SOC). They knew material concepts (spin-orbit coupling, ferromagnetism). The failure happened in the decision step: they judged "peak shape" to be the decisive cue and ignored the more salient texture/colour contrast that the question emphasised.