

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

Anonymous ACL submission

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. MatVQA benchmark data, along with evaluation code, is publicly available in <https://anonymous.4open.science/r/matvqa-1E01/README.md> to catalyze further research in 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), 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., 2024), 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.

MatVQA is built by **MARxivAgent**, an auto-

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					(b) MatVQA benchmark attributes	
Benchmark	Level	Domain	Source	Size	MatVQA feature	Value
ScienceQA	High-school	Science	Exams	16.8k	Total questions	1,672
MicroBench	Graduate	Microscopy	Datasets	17.2k	Causal SPP questions	950
MMSci	PhD	Science	Figures	7,132	Comp SPP questions	112
MacBench	Research	Chem&Mat	Lab	628	Hypo SPP questions	256
LabBench	Research	Biology	WebQA	181	Quan SPP questions	354
MicroVQA	Research	Microscopy	Expert	1,042	Unique images	731
MatVQA (Ours)	Research	Materials	Paper	1,672	Unique papers	120
					Research areas	32

085 mated and verifiable pipeline engineered for the
086 efficient generation of challenging multiple-choice
087 questions (MCQs) directly from arXiv materials-
088 science papers. This automated construction under-
089 pins MatVQA’s inherent scalability, facilitating on-
090 going expansion and adaptation. After generation
091 by advanced LLMs, 50% of the MCQs are vetted
092 by domain experts to ensure quality. Grounded in
093 real-world literature, the benchmark targets four
094 research-critical structure–property–performance
095 (SPP) tasks that map how a material’s structure
096 governs its measurable properties and ultimate per-
097 formance. These tasks—Quantitative SPP, Com-
098 parative SPP, Causal SPP, and Hypothetical Vari-
099 ation—are derived from core components of sci-
100 entific inquiry and by their very nature necessitate
101 the complex reasoning MatVQA aims to evaluate.
102 Solving them requires a tight integration of visual
103 and textual evidence, collectively probing the core
104 cognitive operations of materials research.

105 These tasks mirror the questions materials sci-
106 entists pose when characterising, designing, and
107 optimising materials, and they rigorously test an
108 MLLM’s ability to perform fine-grained visual per-
109 ception, numerical reasoning, comparative analysis,
110 causal inference, and forward prediction. As shown
111 in Table 1, MatVQA contains 1672 multiple-choice
112 questions automatically generated by advanced lan-
113 guage models and randomly verified by domain
114 experts carefully.

115 Motivated by the forensic study of multimodal
116 biases in (Burgess et al., 2025), we identify
117 two textual artifacts that can subvert genuine vi-
118 sion–language evaluation: **(i) Language shortcuts**
119 appear when the prompt provides verbose image
120 descriptions, weak distractors, or stylistic cues that
121 allow the answer to be inferred from the text alone.
122 **(ii) Caption shortcuts** happen when the informa-
123 tion embedded in its stem or options—typically

124 paraphrased from the figure caption—suffices to
125 answer the question without inspecting the figure
126 itself. Caption shortcuts are especially prevalent
127 in materials-science corpora because captions ex-
128 plicitly condense the key morphological, crystallo-
129 graphic, or spectroscopic observations. Their inad-
130 vertent reuse limits MCQs to high-level descriptors
131 (e.g., phase label, mean grain size) and excludes
132 the low-level visual cues—diffraction peaks, defect
133 textures, subtle contrast variations—crucial for au-
134 thentic reasoning. Additional leakage stems from
135 distractors generated with caption vocabulary, re-
136 liance on caption-embedded numerical values, and
137 minimal emphasis on spatial or pixel-level patterns.

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

143 To excise these artifacts, MARxivAgent executes
144 an *iterative shortcut-elimination loop*. After initial
145 question synthesis, an evaluator agent answers the
146 MCQ using (a) only the stem and options and (b)
147 the stem, options plus caption but *without* the im-
148 age. Success in either mode triggers a rewriter that
149 removes or rephrases the incriminating text while
150 a consistency checker enforces fidelity to the origi-
151 nal scientific claim. By progressively eliminating
152 both language and caption shortcuts, we elevate
153 the benchmark from coarse- to fine-grained diffi-
154 culty: solving the final questions requires precise,
155 low-level visual scrutiny (e.g., counting diffrac-
156 tion spots, discerning lattice fringes) coupled with
157 multi-hop scientific reasoning. This refinement is
158 essential for measuring the true multimodal com-
159 petence that front-line materials research demands.

160 We evaluated a suite of 19 open- and closed-
161 source MLLMs on MatVQA and compared the per-
162 formance 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 (Manh 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., 2025; 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 materials. The benchmark includes a vast dataset of

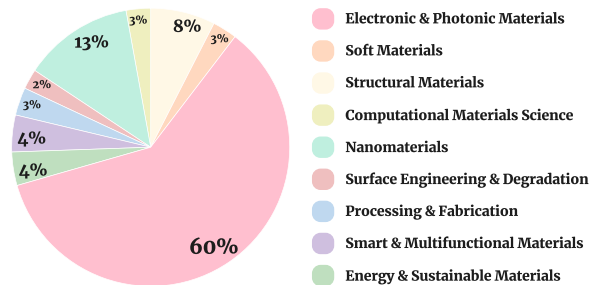


Figure 1: Domain attribution for **MatVQA**

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 topical relevance and cutting-edge difficulty.

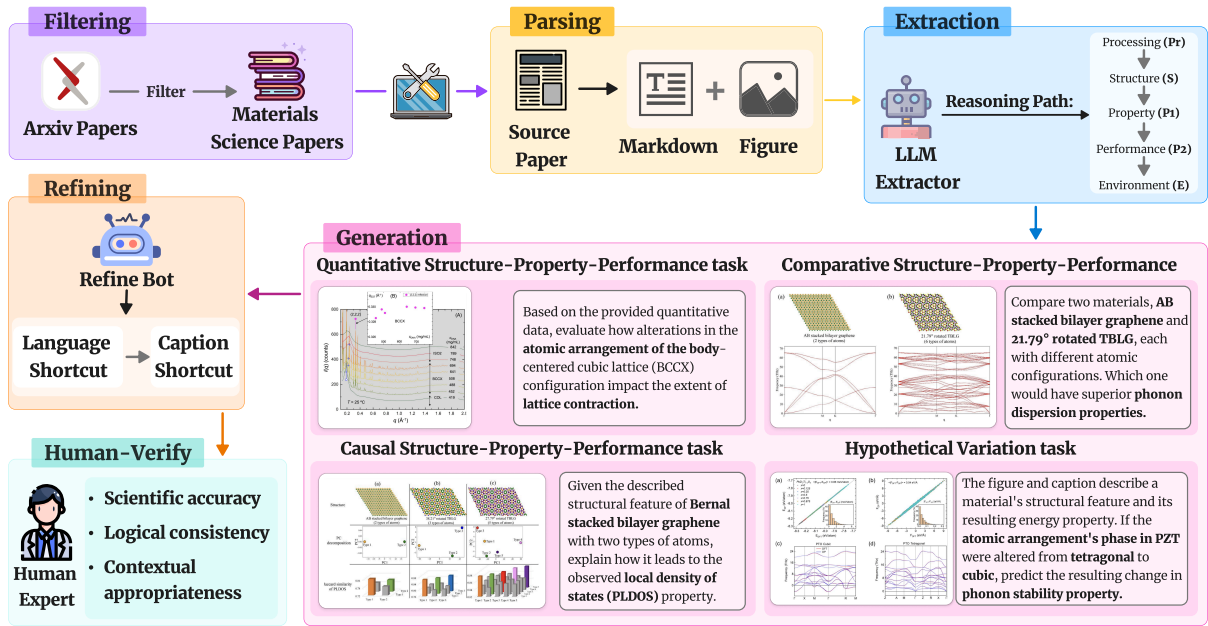


Figure 2: Construction Pipeline of MatVQA

(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 platform¹, 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 dissemination of state-of-the-art research, and compatibil-

¹<https://arxiv.org/>

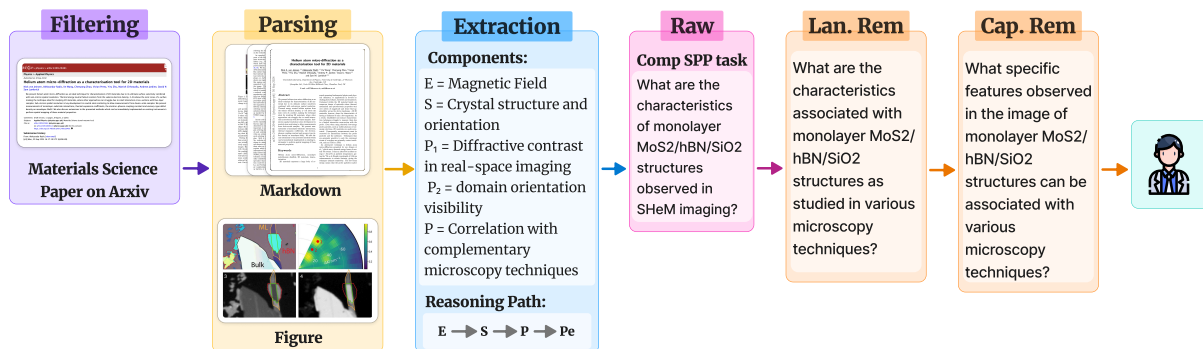


Figure 3: MARxivAgent Pipeline for MCQ automatically Generation. "Lan. Rem" represents the question after language shortcut removal. "Cap. Rem" represents the question after removing caption.

ity 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 Marker² 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 consistency identification of these elements, we incorporate the MatOnto³ 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 Choices 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 op-

tion. 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 language 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 avoid the significant changes on the revised question-answering pairs, an LLM checker is applied to ensure semantic equivalent with the original pair. This process will

²<https://github.com/VikParuchuri/marker>

³<https://matportal.org/ontologies/MATONTO>

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(AI, 2024), as well as the closed-source GPT-4o(OpenAI, 2024) and Gemini(AI, 2025). We utilize standard chain-of-thought prompting(Yue et al., 2024)(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 o1 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

621	a degree of subjectivity. Broader validation from	Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor	674
622	additional domain experts would further enhance	Guizilini, and Yue Wang. Physbench: Benchmarking	675
623	the dataset’s reliability.	and enhancing vision-language models for physical	676
624	Ethical Considerations	world understanding, 2025. URL https://arxiv.org/abs/2501.16411 .	677
625	During the development of MatVQA, we prioritized	Elizabeth Dinella, Satish Chandra, and Petros Maniatis.	679
626	the ethical and responsible use of data.	Crqbench: A benchmark of code reasoning questions,	680
627	The current MatVQA dataset is constructed exclu-	2024. URL https://arxiv.org/abs/2408.08453 .	681
628	sively from publicly available sources, specifically	Alex Gu, Baptiste Rozière, Hugh Leather, Armando	683
629	preprints from arXiv. Additionally, the released	Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang.	684
630	version of MatVQA is intended solely for research	Cruxeval: A benchmark for code reasoning, under-	685
631	purposes, governed by a license that prohibits com-	standing and execution, 2024. URL https://arxiv.org/abs/2401.03065 .	686
632	mercial use or redistribution of the source data.		687
633	References	Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng	688
634	Google AI. Gemini 2.5: Our most intelligent ai	Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang,	689
635	model. https://blog.google/technology/google-deepmind/	Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang.	690
636	gemini-model-thinking-updates-march-2025/ ,	Can llms solve molecule puzzles? a multimodal	691
637	Mar 2025. Accessed May 6, 2025; Experimental	benchmark for molecular structure elucidation. In	692
638	Version.	A. Globerson, L. Mackey, D. Belgrave, A. Fan,	693
639		U. Paquet, J. Tomczak, and C. Zhang (eds.),	694
640	Meta AI. Llama 3: Open foundation and instruction-	<i>Advances in Neural Information Processing Systems</i> ,	695
641	tuned language models. https://ai.meta.com/	volume 37, pp. 134721–134746. Curran Asso-	696
642	research/llama , 2024. Version 3.2-90b-Vision-	ciates, Inc., 2024. URL https://proceedings.	697
643	Instruct; Accessed May 6, 2025.	neurips.cc/paper_files/paper/2024/file/	698
644	Sören Auer, Dante AC Barone, Cassiano Bartz, Ed-	f2b9e8e7a36d43ddfd3d55113d56b1e0-Paper-Datasets_	699
645	uardo G Cortes, Mohamad Yaser Jaradeh, Oliver	and_Benchmarks_Track.pdf .	700
646	Karras, Manolis Koubarakis, Dmitry Mouromtsev,	Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie	701
647	Dmitrii Plukhin, Daniil Radyush, et al. The sciqa sci-	Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng.	702
648	entific question answering benchmark for scholarly	Can mllms reason in multimodality? emma: An	703
649	knowledge. <i>Scientific Reports</i> , 13(1):7240, 2023.	enhanced multimodal reasoning benchmark, 2025.	704
650	Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash,	URL https://arxiv.org/abs/2501.05444 .	705
651	Sanmi Koyejo, Alison Callahan, Jason A Fries,	Noah Hollmann, Samuel Müller, Lennart Purucker,	706
652	Michael Wornow, Akshay Swaminathan, Lisa So-	Arjun Krishnakumar, Max Körfer, Shi Bin Hoo,	707
653	leymani Lehmann, et al. Testing and evaluation of	Robin Tibor Schirrmeyer, and Frank Hutter. Ac-	708
654	health care applications of large language models: a	curate predictions on small data with a tabular foun-	709
655	systematic review. <i>JAMA</i> , 2024.	dation model. <i>Nature</i> , 637(8045):319–326, 2025.	710
656	Lenz Belzner, Thomas Gabor, and Martin Wirsing.	Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-	711
657	Large language model assisted software engineering:	Feghali, Shruti Badhwar, Joshua D Bocarsly, An-	712
658	prospects, challenges, and a case study. In <i>Internat-</i>	dres M Bran, Stefan Bringuier, L Catherine Brin-	713
659	<i>ional Conference on Bridging the Gap between AI</i>	son, Kamal Choudhary, Defne Circi, et al. 14 ex-	714
660	<i>and Reality</i> , pp. 355–374. Springer, 2023.	amples of how llms can transform materials science	715
661	Tom Brown, Benjamin Mann, Nick Ryder, Melanie	and chemistry: a reflection on a large language model	716
662	Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind	hackathon. <i>Digital discovery</i> , 2(5):1233–1250, 2023.	717
663	Neelakantan, Pranav Shyam, Girish Sastry, Amanda	Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fan-	718
664	Askell, et al. Language models are few-shot learners.	jia Yan, Tianjun Zhang, Sida Wang, Armando	719
665	<i>Advances in neural information processing systems</i> ,	Solar-Lezama, Koushik Sen, and Ion Stoica. Live-	720
666	33:1877–1901, 2020.	codebench: Holistic and contamination free eval-	721
667	James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez,	uation of large language models for code. <i>arXiv</i>	722
668	Alejandro Lozano, Sanket Rajan Gupte, Jesus G	<i>preprint arXiv:2403.07974</i> , 2024.	723
669	Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha	Sung Jae Jung, Hajung Kim, and Kyoung Sang Jang.	724
670	Bhowmik, Zachary Coman, et al. Microvqa: A mul-	Llm based biological named entity recognition from	725
671	timodal reasoning benchmark for microscopy-based	scientific literature. In <i>2024 IEEE International Con-</i>	726
672	scientific research. <i>arXiv preprint arXiv:2503.13399</i> ,	<i>ference on Big Data and Smart Computing (Big-</i>	727
673	2025.	<i>Comp)</i> , pp. 433–435. IEEE, 2024.	728
		Ehsan Latif, Ramviyas Parasuraman, and Xiaoming	729
		Zhai. Physicsassistant: An llm-powered interactive	730

731	learning robot for physics lab investigations. In <i>2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)</i> , pp. 864–871. IEEE, 2024.	786
732		787
733		788
734		789
735	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.), <i>Advances in Neural Information Processing Systems</i> , volume 37, pp. 30670–30685. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/36b31e1bb8ecd4f4081686448e9eff2d-Paper-Datasets_and_Benchmarks_Track.pdf .	790
736		791
737		792
738		793
739		794
740		795
741		796
742		797
743		798
744		799
745		800
746		
747	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 .	801
748		802
749		
750		803
751		804
752		805
753	Rachel K Luu and Markus J Buehler. Bioinspiredllm: Conversational large language model for the mechanics of biological and bio-inspired materials. <i>Advanced Science</i> , 11(10):2306724, 2024.	806
754		807
755		808
756		
757	Dung Nguyen Manh, Thang Phan Chau, Nam Le Hai, Thong T. Doan, Nam V. Nguyen, Quang Pham, and Nghi D. Q. Bui. Codemmlu: A multi-task benchmark for assessing code understanding & reasoning capabilities of codellms, 2025. URL https://arxiv.org/abs/2410.01999 .	809
758		810
759		811
760		812
761		813
762		814
763	Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. Using an llm to help with code understanding. In <i>Proceedings of the IEEE/ACM 46th International Conference on Software Engineering</i> , pp. 1–13, 2024.	815
764		816
765		817
766		818
767		
768	Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. In <i>Informatics</i> , volume 11, pp. 57. MDPI, 2024.	819
769		820
770		821
771	OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ . Accessed May 6, 2025.	822
772		823
773		824
774		825
775	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. <i>NPJ digital medicine</i> , 6(1): 210, 2023.	826
776		827
777		828
778		829
779		830
780		831
781	Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bouso Dieng. Llm4matbench: benchmarking large language models for materials property prediction. <i>arXiv preprint arXiv:2411.00177</i> , 2024.	832
782		833
783		834
784		835
785		836
		837
		838
		839
		840
		841
		842
		843
		844
		845
		846
		847
		848
		849
		850
		851
		852
		853
		854
		855
		856
		857
		858
		859
		860
		861
		862
		863
		864
		865
		866
		867
		868
		869
		870
		871
		872
		873
		874
		875
		876
		877
		878
		879
		880
		881
		882
		883
		884
		885
		886
		887
		888
		889
		890
		891
		892
		893
		894
		895
		896
		897
		898
		899
		900

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'

```

1:  $Q' \leftarrow Q$ 
   Stage 1: Language Shortcut Removal
2: for each  $(q, A, D)$  in  $Q'$  do
3:    $t \leftarrow 0$ 
4:   while  $t < T$  do
5:      $r \leftarrow \text{Evaluator}(q, C)$ 
6:     if not  $r.\text{correct}$  then
7:       break
8:     end if
9:      $s \leftarrow \text{ExtractLangStrategy}(r)$ 
10:     $(q', C') \leftarrow \text{Rewriter}(q, C, s)$ 
11:     $v \leftarrow \text{Checker}(q, A; q', A')$ 
12:    if  $\text{semantically\_inconsistent}$  then
13:      FAIL_REWRITE and continue
14:    end if
15:     $t \leftarrow t + 1$ 
16:  end while
17: end for
   Stage 2: Caption Shortcut Removal
18: for each  $(q, C)$  in  $Q'$  do
19:    $t \leftarrow 0$ 
20:   while  $t < T$  do
21:      $r \leftarrow \text{Evaluator}(q, C, \text{captions})$ 
22:     if not  $r.\text{correct}$  then
23:       break
24:     end if
25:      $s \leftarrow \text{Reflector}(r)$ 
26:      $(q', C') \leftarrow \text{Rewriter}(q, C, s)$ 
27:      $v \leftarrow \text{Checker\_caption}(q, A; q', A')$ 
28:     if  $\text{reasoning\_path\_inconsistent}$  then
29:       FAIL_REWRITE and continue
30:     end if
31:      $t \leftarrow t + 1$ 
32:   end while
33: end for
34: 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., 2024):

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.

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

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

881 **A.5 Representative Examples**

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

885 **A.6 Failure Analysis**

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

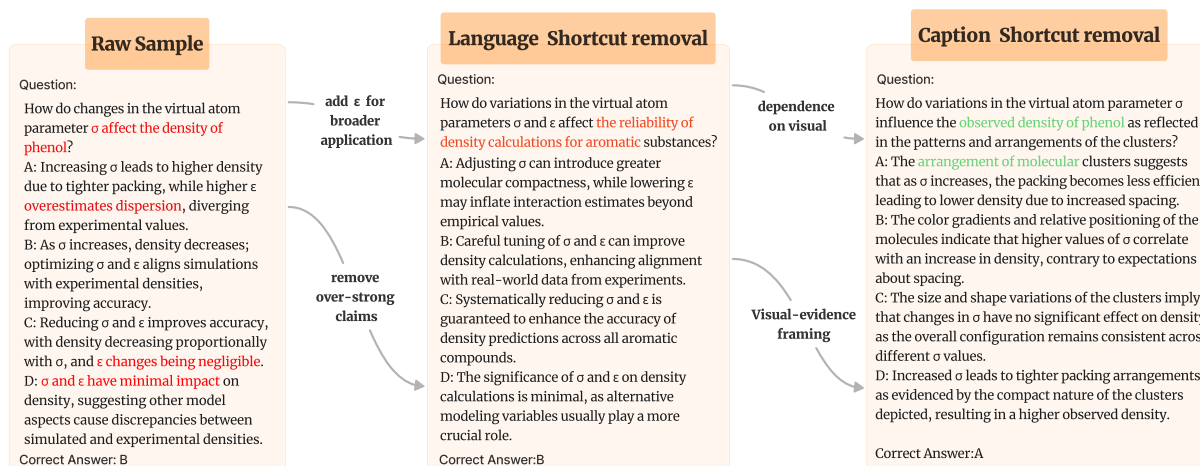


Figure 4: Evolution of a sample question through the two-stage shortcut removal process. The figure shows the transformation from: the initial 'Raw Sample,' to after 'Language Shortcut removal', and finally to after 'Caption Shortcut removal'.

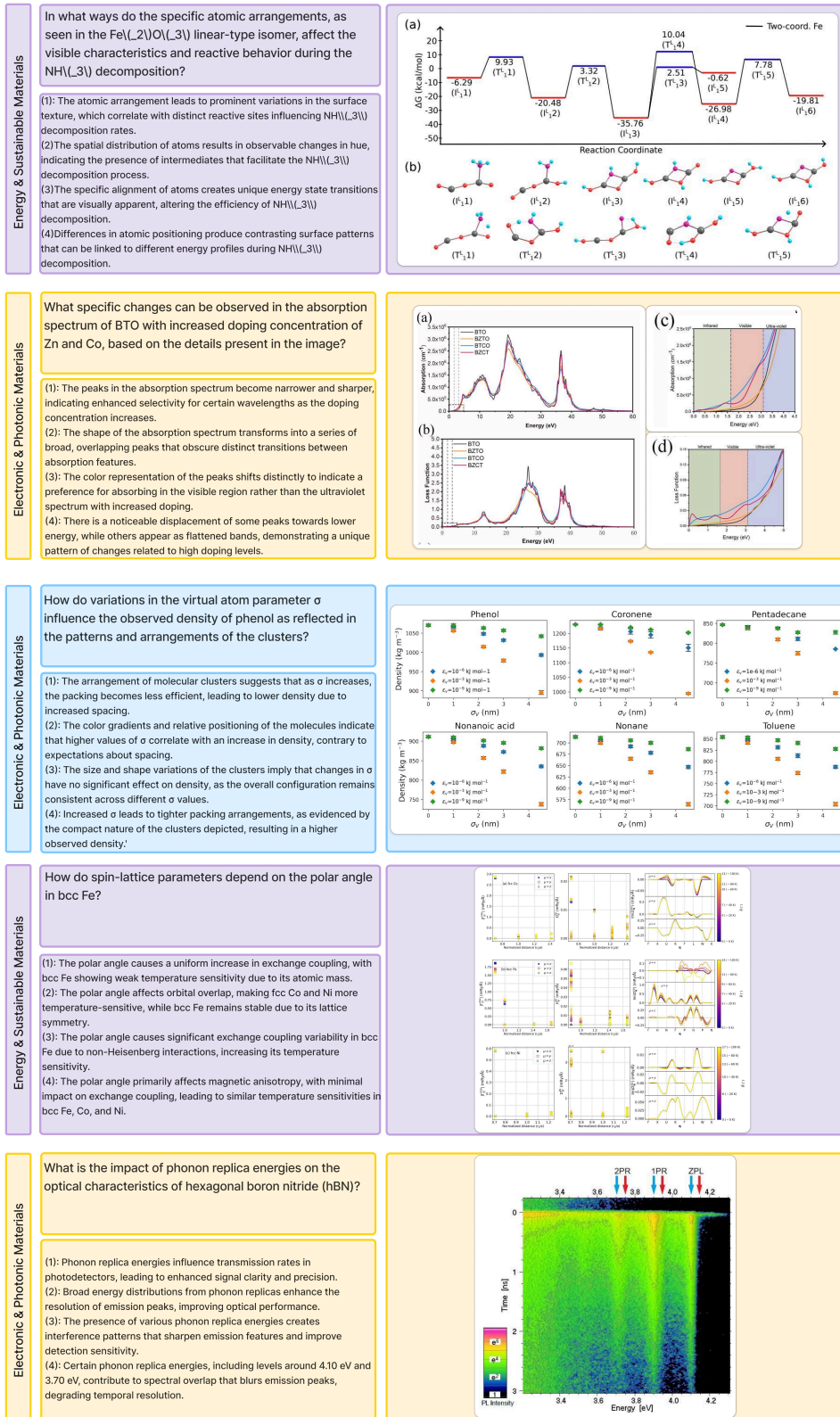


Figure 5: Representative Examples for varies material science domain.

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.



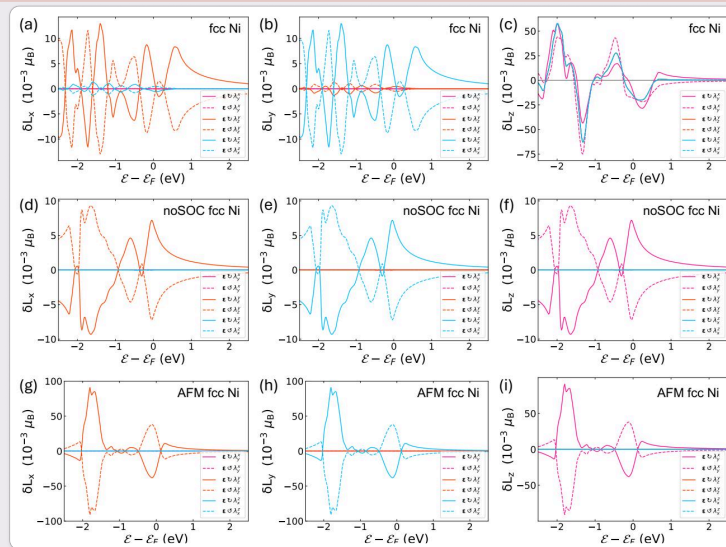
Llama-90b Response: To determine the effect of observed binding energy on the catalytic performance for ammonia decomposition, we need to understand how binding energy influences the reaction process. Binding energy in this context refers to the energy required to adsorb reactants (in this case, ammonia) 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 NH_3 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 (δL_x , δL_y , δL_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 ($(\epsilon \times \lambda^2)$, 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 (δL_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/color 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/color contrast that the question emphasised.