MSQA: Benchmarking LLMs on Graduate-Level Materials Science Reasoning and Knowledge

Anonymous ACL submission

Abstract

Despite recent advances in large language models (LLMs) for materials science, there is a lack of benchmarks for evaluating their domain-specific knowledge and complex reasoning abilities. To bridge this gap, we introduce MSQA, a comprehensive evaluation benchmark of 1,757 graduate-level materials science questions in two formats: detailed explanatory responses and binary True/False assessments. MSQA distinctively challenges LLMs by requiring both precise factual knowledge and multi-step reasoning across seven materials science sub-fields, such as structure-property relationships, synthesis processes, and computational modeling. Through experiments with 10 state-of-the-art LLMs, we identify significant gaps in current LLM performance. While API-based proprietary LLMs achieve up to 84.5% accuracy, open-source (OSS) LLMs peak around 60.5%, and domain-specific LLMs often underperform significantly due to overfitting and distributional shifts. MSQA represents the first benchmark to jointly evaluate the factual and reasoning capabilities of LLMs crucial for LLMs in advanced materials science.

1 Introduction

Recent advances in large language models (LLMs) hold considerable promise for accelerating materials research by automating complex analysis (Miret and Krishnan, 2024), prediction (Rubungo et al., 2024), and design tasks (Jablonka et al., 2023). However, progress in applying LLMs to materials science has been constrained by the scarcity of high-quality domain-specific annotated datasets (Miret and Krishnan, 2024) and robust evaluation benchmarks (Zaki et al., 2023).

Existing datasets for materials science primarily focus on conventional natural language processing (NLP) tasks, such as named entity recognition (NER) (Weston et al., 2019; Shetty et al., 2023), classification (Venugopal et al., 2021; Gupta et al., 2022b), and relation extraction (RE) (Cheung et al., 2023; Song et al., 2023a). Although valuable, these datasets do not sufficiently assess models' materials science knowledge grounded reasoning and question-answering abilities. Recent efforts (Zaki et al., 2023) have introduced benchmarks featuring questions derived from graduate admission exams; however, the short-answer format (*e.g.*, multiple choice) limits the assessment of complex reasoning and in-depth explanatory capabilities essential for real-world applications in materials science, such as multistep synthesis planning and detailed property evaluations.

Question:
What structural transitions occur in tetraphenylethylene carboxylate aggregates as the size of the liquid phase decreases from micro to submicro dimensions?
Long-answer Response:
Tetraphenylethylene carboxylate (TPEC) aggregates undergo a two-step structural transition as the size of the liquid phase decreases from micro- to submicrometer dimensions due to ice confinement. Initially, in relatively large confined spaces (greater than 2 µm), TPEC aggregates are loosely packed, and their molecular arrangement changes continuously, leading to increased fluorescence intensity and a hypsochromic (blue) shift in emission spectra. This structural change stabilizes when the space size reduces to 2 µm, where the aggregates adopt a more consistent arrangement.

As the liquid phase shrinks further to submicrometer dimensions (below 0.5 µm), a second structural transition occurs, where the aggregates are compressed into a more compact molecular arrangement. This densification is driven by enhanced solubility in the confined phase, which exhibits properties akin to low-density liquid water, as well as the structured configuration of water molecules at the ice-liquid interface. These structural transitions significantly impact fluorescence properties, highlighting the role of confinement effects in influencing aggregate behavior.

Question:
Do tetraphenylethylene carboxylate aggregates undergo a densification transition due to enhanced solubility in submicrometer liquid phases?
Binary-answer Response:
YES

Figure 1: Example question-answer pairs from MSQA. The long-answer response (top) demonstrates multi-step reasoning with explanations, while the binary-answer response (bottom) shows a concise yes/no determination on a related concept.

To bridge this critical gap, we introduce MSQA, a graduate-level materials science benchmark specifically crafted to evaluate both factual knowledge and complex reasoning capabilities of LLMs. MSQA includes two complementary evaluation modes: *long-answer* and *binary-answer* (Figure 1). The long-answer questions demand detailed, multi-step explanations spanning seven challenging subfields, including structure-property relationships, polymer synthesis, and computational material modeling. In contrast, the binary-answer questions offer balanced True/False queries that require LLMs to assess complex domain-specific judgments on material properties, applications, and technical claims. Together, MSQA comprehensively test the depth of factual understanding and advanced reasoning skills.

To ensure high-quality and domain-grounded questions and answers, we employ advanced LLMs, including `gpt-4o` (Hurst et al., 2024), `gemini-2.0-pro` (Team et al., 2023), and `Deepseek-v3` (DeepSeek-AI, 2025), guided by expert-curated materials science literature. The dataset generation process incorporates rigorous three-stage quality assurance: (1) regular expression-based filtering, (2) LLM-driven refinement, and (3) expert annotation.

In our experiments, we systematically benchmark seven leading open-source and black-box LLMs alongside three domain-specific fine-tuned models. Our results reveal that commercial black-box LLMs consistently outperform open-source alternatives, achieving accuracy as high as 84.5%. Incorporating retrieved contextual data notably enhances model performance, showing retrieval augmentation as a crucial adaptation strategy. Conversely, domain-specific fine-tuned models underperform relative to general-purpose models, likely due to distribution shifts and overfitting, underscoring critical limitations in current domain-adaptation approaches. We summarize our main contributions as follows:

- We present MSQA, one of the first materials science benchmarks explicitly designed to rigorously test complex reasoning and explanatory abilities of LLMs beyond factual knowledge;
- We provide a thorough empirical evaluation of leading general-purpose and domain-specific LLMs. We conduct detailed analyses of challenging scenarios to catalyze the development of more reasoning-intensive, domain-adapted LLMs for the materials science community.
- We open-source our curated dataset and benchmark code to foster community-driven innovation towards LLM-driven advanced material science discovery: `https://anonymous.4open.science/r/MSQA-C6C5/MSQA/`.

2 Related Works

Materials Science Datasets for LLMs. Prior research in materials science NLP primarily targets structured extraction tasks such as NER and RE. Expert-curated datasets, such as Wang et al. (2021) and Weston et al. (2019), focus explicitly on identifying and extracting material names, properties, and their interrelations. Additional specialized datasets emphasize tasks like property prediction; for instance, Friedrich et al. (2020) annotated a corpus of scholarly articles related to solid oxide fuel cells, tagging entities such as `materials`, `values`, and `devices`, while Panapitiya et al. (2021) provided annotations on chemical entities (`CHEM`), numerical values (`VALUE`), and measurement units (`UNIT`) from studies on soluble materials. More recent datasets aimed at evaluating LLMs include question-answer (QA) pairs to test domain knowledge. Zaki et al. (2023) created a dataset with 650 questions derived from graduate-level admissions exams in India, while Song et al. (2023a) aggregated multiple previously published datasets into a meta-dataset. However, they primarily utilize short-answer formats such as multiple-choice or numerical values, which inadequately capture the nuanced reasoning and explanatory capabilities required in real-world materials science applications.

Synthetic Data Generation for Benchmarks. LLMs increasingly serve as tools for creating evaluation benchmarks, especially when manual curation requires domain expertise or is prohibitively expensive. For example, SciBench (Wang et al., 2024c) and BioGPTQA (Sarwal et al., 2025) employ LLM-generated content subsequently validated through expert reviews and structured filtering mechanisms. Other benchmarks, including MT-Bench (Zheng et al., 2023) and HELM (Liang et al., 2023), similarly rely on synthetically generated data to evaluate model performance across diverse tasks. Synthetic data generation markedly decreases annotation expenses; however, it is important to ensure the validity and reliability of generated benchmarks. Researchers employ quality

control measures such as expert validation, statistical filtering, and alignment with reference materials to minimize factual inaccuracies and reasoning shortcuts.

3 MSQA: A Graduate-Level Materials Science QA Dataset

We present MSQA, a comprehensive materials science benchmark for evaluating LLMs (Figure 2). We begin by describing our scientific literature collection process in section 3.1. We then detail the procedures for generating long-answer questions and their corresponding answers in section 3.2 and section 3.3, respectively. Next, we outline our three-stage quality assurance process in section 3.4. We then describe the generation process for binary-answer questions in section 3.5. Finally, we summarize key dataset statistics for MSQA in section 3.6.

3.1 Data Collection

Previous research has highlighted the inherent limitations of LLMs in effectively addressing questions grounded in specialized materials science knowledge (Miret and Krishnan, 2024; Zaki et al., 2023; Wang et al., 2024a). To overcome this limitation, we curate a comprehensive collection of materials science literature to serve as a context for generating high-quality QA pairs. Specifically, we source 3,000 diverse articles from the extensive dataset compiled by Shetty et al. (2023), encompassing 2.4 million publications from seven primary materials science publishers spanning from 2000 to 2021. We employ sentence transformers (Reimers and Gurevych, 2019) to generate embeddings for abstracts from 50,000 randomly selected articles to ensure diversity and representativeness. These embeddings are subsequently clustered using K-means to achieve the highest silhouette score (Shahapure and Nicholas, 2020), resulting in 10 distinct clusters, each with 5000 papers on average. From each cluster, we randomly sample 300 articles, culminating in our final set of 3,000 context articles.

Instead of using PDF documents, we collected papers in XML format and parsed them using a publicly available `chemistry-paper-parser` tool.[1] This allows accurate information extraction from these papers by preserving the integrity of mathematical formulas and chemical representations.

[1] https://github.com/Yinghao-Li/ChemistryHTMLPaperParser

3.2 Candidate Question Generation

To ensure that the generated questions accurately reflect the complexity and depth of real-world materials science research, we utilize the abstracts of selected articles, as they succinctly encapsulate research objectives, methodologies, and significant findings. Following Zhong et al. (2024), we first prompt `gpt-4o` to generate concise summaries highlighting key findings from each abstract. This initial summarization step reduces the influence of specialized scholarly language, thereby facilitating more precise and targeted question generation. Subsequently, each article is categorized based on its primary objective: introducing a new synthesis method ("method") or presenting novel experimental observations ("result"). Guided by this categorization, `gpt-4o` is then prompted to formulate candidate questions specifically aligned with the focus identified in the article.

3.3 Question Selection and Candidate Answer Generation

Previous studies have shown that context relevance (Wang et al., 2024b) and factual precision (Ram et al., 2023) are crucial factors for response quality. To enhance alignment between questions and provided context, we employ a backward selection approach, where we use `gpt-4o` to select the candidate questions based on the provided context. Specifically, for articles categorized under "methods", we provide the `Experimental Method` sections containing detailed descriptions of research protocols and materials synthesis procedures. For articles categorized as "results", we supply the `Results` sections, which include comprehensive interpretations of experimental outcomes.

Candidate answers were then generated using `gpt-4o`, `gemini-2.0-pro`, and `deepseek-v3`. Initial assessments highlight that LLM-generated answers frequently included ambiguous references such as "`the K0 Samples` and `the SCAs Units`", diminishing clarity and self-contained informativeness. To address these issues, we refine the prompts by explicitly discouraging the use of definite articles, significantly enhancing answer clarity. This adjustment results in more precise and contextually anchored responses, explicitly referencing chemical entities (*e.g.*, `hexamethyldisilane`, `copper phthalocyanine`) and specific methodologies (*e.g.*, `CryoTEM Imaging`, `dynamic light scattering`) from the original studies.

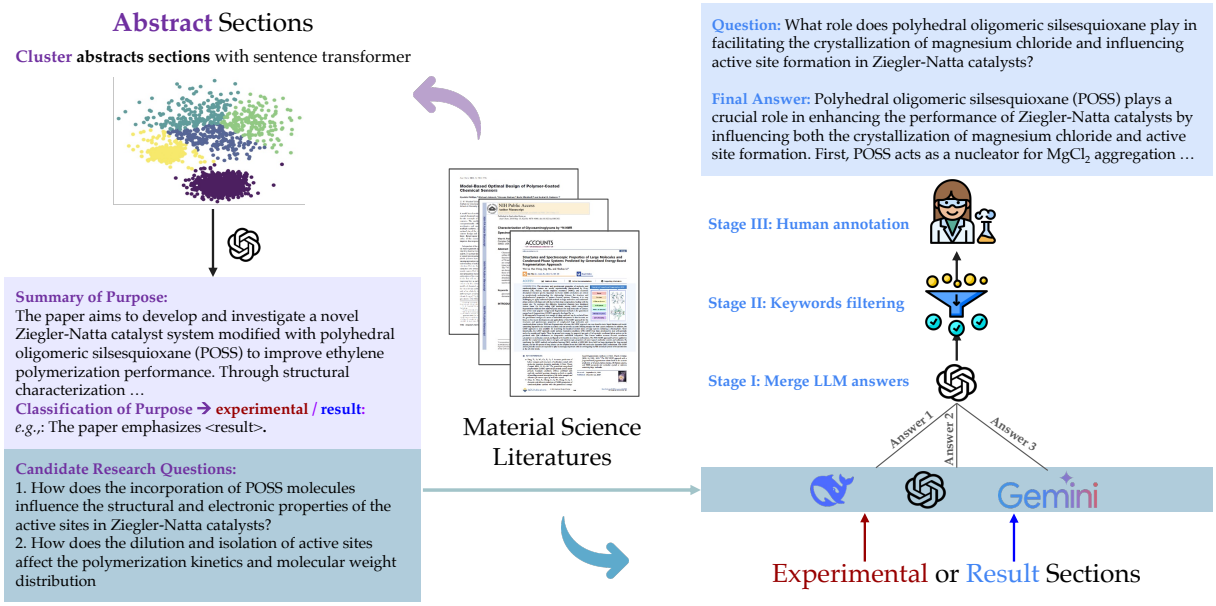

Figure 2: Overview of data generation and quality evaluation in MSQA.

3.4 Three-Stage Quality Assurance

Prior research (Wang et al., 2022; Huang et al., 2022) indicates that self-consistency among LLMs significantly enhances answer accuracy and coherence. Extending this idea, Li et al. (2025) applies self-consistency to handle open-ended tasks effectively. Inspired by these findings, we adopt a multi-model self-consistency approach to enhance answer quality. Specifically, candidate answers produced by `gpt-4o` (Hurst et al., 2024), `gemini-2.0-pro` (Team et al., 2023), and `deepseek-v3` (DeepSeek-AI, 2025) (section 3.3) are aggregated using `gpt-4o`. This aggregation explicitly accounts for inter-model agreement, leveraging consensus among multiple sophisticated models.

To ensure question clarity and relevance, we first employ automated filtering methods, using regular expressions and keyword matching to remove ambiguous and overly context-dependent questions. Subsequently, materials science domain experts manually review the remaining questions to exclude unclear, incorrect, or irrelevant queries.

We recruit two materials science PhD students to rigorously evaluate the quality of a representative subset of the generated answers. Specifically, we randomly select 50 question-answer pairs for assessment. Each evaluator independently applied their expert domain knowledge to assess whether the provided answers are: (1) factually correct, (2) directly relevant and precisely addressed the questions, and (3) logically coherent. The evaluators show that 92.86% of answers fulfilled all three quality criteria.

3.5 Binary Question Generation

Given the computational expense associated with evaluating detailed long-form answers using advanced LLMs such as `gpt-4o`, we develop a set of 1,757 binary True/False questions derived from previously generated question-answer pairs. Initial efforts to directly convert existing pairs into binary format reveal several issues: (1) questions frequently included overly detailed clues revealing the correct answer; (2) LLM-generated questions demonstrated a pronounced bias towards "True" responses; and (3) the generated questions lacked complexity, often omitting nuanced reasoning involving approximations or comparative thresholds (*e.g.*, "exact value" versus "around" a value).

To tackle these issues, we instruct `gpt-4o` to (1) explicitly generate questions with a predefined True or False label, yielding 878 "True" and 879 "False" labeled questions; (2) favor approximations (*e.g.*, "around", "more or less than") for rigid details that are not essential; and (3) avoid unnecessary detail without adding challenge. Detailed prompts are included in appendix B.

3.6 Dataset Statistics

Table 1 summarizes the basic statistics of MSQA.

Question Types. We manually annotate a subset of questions to identify and categorize various

Question & Answer Pairs	Counts
Long-answer Q&A pairs	1,757
Binary-answer Q&A pairs	1,757
- # w/ "True" label	878
- # w/ "False" label	879
Avg. question length (in words)	19
Avg. long answer length (in words)	150

Table 1: Data statistics of MSQA.

Question Types	Counts
Structure-property relationships	818
Synthesis and processing	257
Computational	216
Material analysis techniques	187
Material modeling	125
Failure analysis and degradation	93
Material properties	61

Table 2: Question composition in MSQA.

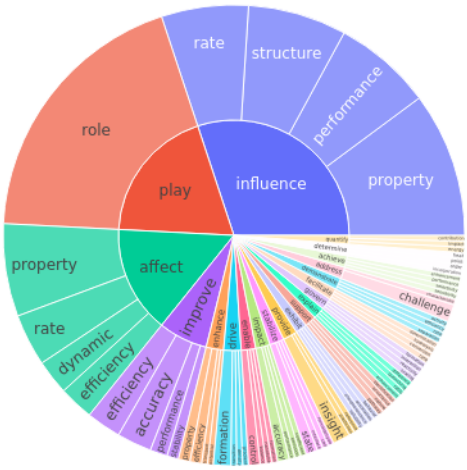

Figure 3: The top 20 most common root verbs (inner circle) and their top 4 direct noun objects (outer circle) in the generated instructions.

question types pertinent to general materials science tasks. Using `gpt-4o` (Hurst et al., 2024), we then classify the remaining questions into these predefined categories. For questions with ambiguous types, `gpt-4o` is prompted to label them as "Other", which are subsequently manually reviewed and categorized by domain experts. In total, we identify seven distinct question categories (Table 2) that comprehensively assess LLM capabilities within the materials science domain.

Question Semantics. To examine question semantics, we analyze their verb-noun structures following the methodology of Wang et al. (2023). We employ the Berkeley Neural Parser[2] to parse each question, extracting the primary verb (closest to the root) and its direct noun object. The most frequent root verbs, along with their associated direct noun objects, are visualized in Figure 3. This analysis highlights the broad topical coverage and complex conceptual nature of questions in MSQA, particularly emphasizing relationships between material designs, their properties, and relevant experimental methodologies.

Answer Semantics. To further explore the diversity of the dataset, we analyze the semantic content of the long-form answers by identifying references to specific materials and chemical compounds using ChemDataExtractor (Swain and Cole, 2016). Extracted chemical entities are embedded using MatSciBERT (Gupta et al., 2022a), a specialized language model trained on extensive materials science literature. These embeddings are then visualized through t-SNE clustering, presented in Figure 4. Upon examining the resulting clusters, we confirm that the answers encompass diverse material categories, including `Polymers & Copolymers` and `Inorganic Complexes`, illustrating the comprehensive topical diversity inherent in our dataset.

[2] `https://parser.kitaev.io/`

4 Experiments

4.1 Experimental Setup

Tasks. We evaluate the performance of open-source and domain-specific LLMs on both the long-answer and binary-answer variants of the MSQA dataset using two distinct prompting strategies: direct generation and retrieval-augmented generation. In the direct-generation scenario, models are presented only with the question, without supplementary context. In contrast, in the retrieval-augmented setting, we first build a contextual database by segmenting the Methods and Results sections described in section 3.2 into separate paragraphs. We then use BM25 (Robertson et al., 2009) to retrieve the top five most relevant paragraphs, which serve as additional context provided to the model. For binary-answer questions, we also investigate if chain-of-thought (Wei et al., 2022) improves model performance. Black-box LLMs are evaluated exclusively under the direct generation setting. Evaluation of the long-answer responses is conducted through GPT-4o acting as an LLM judge, assessing responses as either "correct" or "mostly correct", both categories counted as correct in our metrics. For binary-answer evaluations, accuracy is determined by exact keyword matching for responses

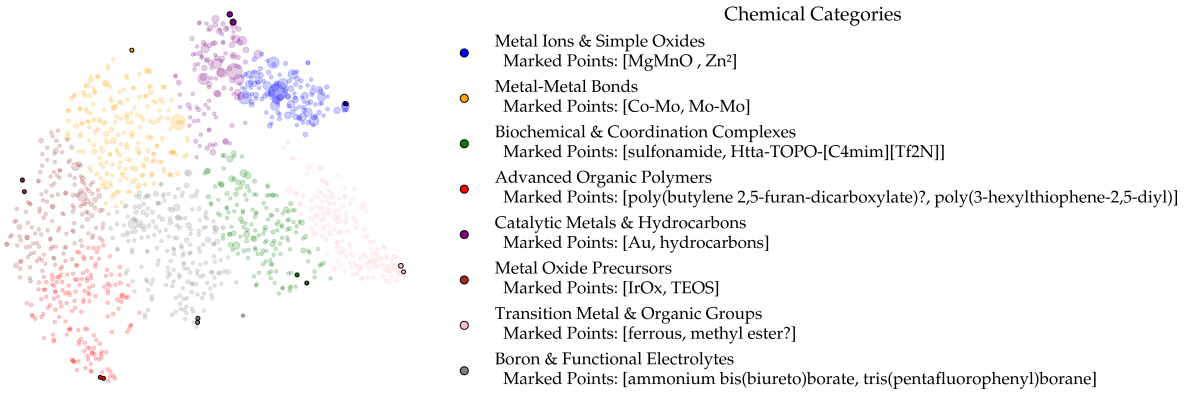

Figure 4: T-SNE visualization of material and chemical mentions from long-form answers, embedded using MatSciBERT. Legend shows manually labeled categories with example compounds.

containing either "YES" or "NO". Performance results are uniformly reported as accuracy percentages across all experiments.

Evaluated LLM models. We comprehensively evaluate several LLMs on MSQA. For black-box models, our analysis includes `Claude-3.7-Sonnet` (Claude), `Gemini-2.0-Flash`, and `Grok-3` (Grok). Additionally, we assess the performance of prominent open-source models, specifically `Llama-3-8B` (Grattafiori et al., 2024), `Phi-4-mini` (Abdin et al., 2024), `Qwen-2.5-7B` (Yang et al., 2024), and `Deepseek-R1-distilled-Llama-3`. Furthermore, we benchmark several domain-specific LLMs specialized for materials science. These include `Honeybee` (Song et al., 2023b), an LLM fine-tuned iteratively on synthesized materials science data; `Mol-Instructions-Molecule` (Fang et al., 2023), a model fine-tuned explicitly on chemical reaction and molecular design datasets tailored for small molecules; and `Llasmol` (Yu et al., 2024), a specialized chemistry domain LLM instruction-tuned across 14 chemistry-specific tasks utilizing a dataset exceeding three million samples.

LLM-as-Judge. Prior studies have consistently validated the effectiveness of employing LLMs for pairwise comparisons across diverse applications (Qin et al., 2023; Liu et al., 2024; Liusie et al., 2023). Moreover, Zeng et al. (2023) demonstrate that incorporating rule-based or self-generated evaluation criteria from LLMs further enhances the accuracy and reliability of these assessments. Given that our dataset involves comprehensive long-answer responses comprising detailed explanations of synthesis processes and materials modeling, we leverage `gpt-4o` as an evaluator with a gold-standard reference answer, a model-generated inference response, and a structured evaluation rubric. `gpt-4o` then evaluates the inference response as "correct", "mostly correct", or "incorrect". We further verify the validity of `gpt-4o`'s judgment by comparing it against human expert evaluations detailed in section 4.3.

4.2 Main Results

Tables 3 and 4 present the results for open-source, domain-specific, and black-box LLMs on MSQA, respectively.

Long-Answer Questions. From Table 3, we highlight several key observations: (1) among the evaluated open-source models, `Deepseek-R1-distilled-Llama3` (DeepSeek-AI, 2025) achieves the highest accuracy (60.50%), outperforming `Qwen-2.5-7B` (Yang et al., 2024) (51.28%) and `Phi-4-mini` (Abdin et al., 2024) (46.39%). The superior performance of `Deepseek-R1-distilled-Llama3` may be attributed to the model's inherent self-correction nature in its thought process, allowing it to review and refine its outputs; (2) retrieval-augmented generation (RAG) notably improves performance for `Llama-3` and `Qwen-2.5-7B`. However, `Phi-4-mini` exhibits only a marginal improvement from 46.39% to 51.28%, likely due to limited exposure to long-context and retrieval-augmented training data in its alignment corpus. This highlights the crucial role of post-training model alignment; and (3) domain-specific LLMs surprisingly underperform compared to general-purpose models in producing coherent long-form answers. This underperformance is

Tasks (→)	Long-answer		Binary-answer		
Baselines (↓)	DP	RAG	DP	COT	RAG
Open-source LLMs					
Llama-3-8B (Grattafiori et al., 2024)	39.39 (16/676/1065)	85.20 (330/1167/260)	63.97	57.37	73.71
Phi-4-mini (Abdin et al., 2024)	46.39 (15/800/942)	51.28 (207/694/856)	68.24	60.39	64.43
Qwen-2.5-7B (Yang et al., 2024)	51.28 (41/860/856)	87.48 (504/1033/220)	72.34	69.89	83.84
Deepseek-R1-distilled-Llama3 (DeepSeek-AI, 2025)	60.50 (37/1026/694)	85.71 (362/1144/251)	52.74	51.91	65.40
Domain-specific LLMs					
Honeybee (Song et al., 2023b)	19.53 (0/343/1413)	2.73 (3/45/1708)	8.82	0.23	0.68
Mol-Instructions-Molecule (Fang et al., 2023)	0.23 (0/4/1753)	6.66 (16/101/1640)	22.82	0.11	11.84
Llasmol (Yu et al., 2024)	4.84 (0/85/1672)	6.26 (12/98/1647)	28.34	5.41	29.82

Table 3: Main results of open-source and domain-specific LLMs on MSQA. Long-answer results are presented as "accuracy" in % (Correct/Mostly Correct/Incorrect). Binary-answer results are presented as "accuracy" in %. Notations are consistent across tables. DP refers to "direct prompting". COT refers to "direct prompting with chain of thoughts".

potentially due to distributional shifts between their specialized finetuning datasets and our more general domain-focused dataset, alongside evident overfitting. For instance, `LlamaSmol` (Yu et al., 2024) model frequently outputs chemical names encapsulated within `<SMILE>` tags, reflecting such training limitations.

Binary-Answer Questions. As shown in Table 3, binary-answer questions present considerable difficulty, with two out of four general-purpose models (`Llama-3-8B` and `Deepseek-R1-distilled-Llama3`) performing only slightly above random guessing levels at 63.97% and 52.74%, respectively. Interestingly, chain-of-thought (Wei et al., 2022) decreases performance for all open-source and domain-specific LLMs. Our analysis reveals that this decline is due to LLMs generating factually incorrect intermediate steps, likely stemming from their limited materials science knowledge, as confirmed by Wang et al. (2024a). Domain-specific models again perform worse than general-purpose models, with the best domain-specific model, `Llasmol`, achieving only 28.34% accuracy. We attribute this performance gap to two causes: 1) the domain-specific models likely overfitted on the finetuning data, decreasing their ability to output "True" and "False" answers; 2) the distribution shift between our dataset and their finetuning corpus.

Black-Box LLMs. Results presented in Table 4 clearly demonstrate that black-box LLMs substantially outperform open-source models in answering long-form questions. Specifically, Grok-3 (Grok) achieves an impressive accuracy of 84.46% without supplementary contextual data. However, performance on binary-answer tasks remains comparable between black-box and open-source models, with slight improvement after chain-of-thought (Wei et al., 2022) prompting.

Model	Long-ans	Binary-ans	Binary-cot
Claude	66.35 (136/840/495)	68.58	70.18
Grok	84.84 (363/885/223)	65.05	71.37
Gemini	77.63 (254/888/329)	72.85	71.54

Table 4: Experimental results of black-box LLMs on MSQA with direct prompting.

4.3 LLM as Judge

We investigate the reliability of using an LLM as a judge by comparing its annotation decisions directly against those made by human annotators. Considering human annotations as the gold standard, we quantified agreement using a confusion matrix, as illustrated in Figure 5. Results indicate that `gpt-4o`'s evaluations align with human judgments in approximately 77.38% of cases, demonstrating particularly high consistency for answers rated as "correct".

GPT-4o-mini as Judge. Due to the computational expense associated with GPT-4o, we explore the viability of using GPT-4o-mini as an alternative judge for assessing long-answer responses generated by open-source LLMs (Table 5). We observe that GPT-4o-mini demonstrates several distinct bi-

Figure 5: Confusion matrix between human and GPT-4o judgments.

Model	Accuracy (%)
Llama-3-8B	77.18
Phi-4-mini	92.26
Qwen-2.5-7B	95.11
Deepseek-R1-distilled-Llama3	96.07

Table 5: Results on long-answer questions with GPT-4o-mini as judge.

ases compared to GPT-4o. Notably, GPT-4o-mini exhibits a pronounced *verbosity bias* (Saito et al., 2023), often incorrectly rating longer responses as accurate, particularly evident in evaluations involving Deepseek-distilled-r1-llama3. Additionally, GPT-4o-mini frequently *accepts vague or irrelevant explanations* as correct. For instance, when asked "How does the presence of an SH-OC hydrogen bond influence the geometry and stability of the global minimum conformer of methyl 3-mercaptopropionate?", GPT-4o-mini deemed a general explanation related to hydrogen bonding sufficient. Moreover, GPT-4o-mini *fails to consistently identify scientific inaccuracies*. Specifically, it incorrectly classified an inference answer that labeled Li_4SnS_4 as an *anode material* rather than correctly as a *solid electrolyte*, a mistake accurately detected by GPT-4o.

4.4 Question Type Difficulties

We conducted a case study to evaluate the performance of Qwen-2.5-7B (Yang et al., 2024) across various categories of materials science questions, as summarized in Table 6. The model exhibited the lowest accuracy (36.07%) on questions related to *material properties*, likely attributable to the lack of symbolic understanding of numerical values. This finding aligns with prior research by Miret and Krishnan (2024), which similarly underscores the difficulties LLMs encounter with materials science numerical tasks. The *failure analysis and degradation* category emerged as the second most challenging, probably due to the sparse representation of these topics within its pre-training dataset. Conversely, Qwen-2.5-7B demonstrated superior performance on questions involving *structure-property relationships*, indicating stronger foundational knowledge likely driven by the broader availability of material structure data in textbooks and journal articles.

Question Type	Accuracy (%)
structure-property relationships	49.27
synthesis and processing	40.86
computational	49.07
material analysis techniques	45.45
material modeling	45.6
failure analysis and degradation	39.78
material properties	36.07

Table 6: Results of Qwen-2.5-7B on each question type.

4.5 Error Analysis

We conduct a detailed manual analysis of inference errors made by LLMs on the MSQA dataset and identify several recurrent failure patterns: (1) LLMs explicitly recognize their own limitations in domain-specific knowledge and thus fail to provide complete answers; (2) LLMs deliver partially accurate responses, neglecting critical sub-questions or necessary qualifying statements; and (3) LLMs generate scientifically incorrect or misleading facts due to hallucination. We include specific examples illustrating each of these error types in appendix D.

5 Conclusion

In this study, we introduce MSQA, a comprehensive benchmark explicitly designed to assess LLMs on complex, domain-specific reasoning and explanatory capabilities in materials science. Comprising 1,757 rigorously crafted long-answer and binary-answer questions, MSQA addresses a significant gap in current evaluation resources by simulating realistic scientific inquiry scenarios. Our extensive evaluation of ten advanced LLMs highlights substantial performance gaps, particularly revealing limitations in accurately generating coherent, nuanced responses to complex materials science queries. MSQA serves as a robust platform for benchmarking and advancing the development of specialized LLMs tailored to the demanding context of materials science research.

Limitations

We rely on GPT-4 as a judge for evaluating long-answer responses. However, this approach introduces evaluation costs for future researchers wishing to replicate or extend our work. Due to constraints on annotation resources, we performed manual annotations on a randomly sampled subset of the questions. This process, while necessary, may result in the inclusion of a small number of low-quality question pairs due to the inherent variability in LLM-generated content. Third, our computational limitations restricted us from evaluating open-source LLMs exceeding 8 billion parameters. We acknowledge that this may limit the generalizability of our findings. We encourage future research to overcome these limitations by assessing larger-scale LLMs.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Jerry Junyang Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. 2023. Polyie: A dataset of information extraction from polymer material scientific literature. *arXiv preprint arXiv:2311.07715*.

Claude. Claude 3.7 Sonnet and Claude Code — anthropic.com. https://www.anthropic.com/news/claude-3-7-sonnet. [Accessed 03-05-2025].

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grok. Grok 3 Beta âĂŤ The Age of Reasoning Agents | xAI — x.ai. https://x.ai/news/grok-3. [Accessed 03-05-2025].

Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. 2022a. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. 2022b. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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, and 1 others. 2023. 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.

Zichong Li, Xinyu Feng, Yuheng Cai, Zixuan Zhang, Tianyi Liu, Chen Liang, Weizhu Chen, Haoyu Wang, and Tuo Zhao. 2025. Llms can generate a better answer by aggregating their own responses. *arXiv preprint arXiv:2503.04104*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher RÃl', Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Adian Liusie, Potsawee Manakul, and Mark JF Gales. 2023. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. *arXiv preprint arXiv:2307.07889*.

Santiago Miret and Nandan M Krishnan. 2024. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*.

Gihan Panapitiya, Fred Parks, Jonathan Sepulveda, and Emily Saldanha. 2021. Extracting material property measurement data from scientific articles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5393–5402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, and 1 others. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Varuni Sarwal, Gaia Andreoletti, Viorel Munteanu, Ariel Suhodolschi, Dumitru Ciorba, Viorel Bostan, Mihai Dimian, Eleazar Eskin, Wei Wang, and Serghei Mangul. 2025. A benchmark for large language models in bioinformatics. *bioRxiv*.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748.

Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kuenneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. 2023. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials*, 9(1):52.

Yu Song, Santiago Miret, and Bang Liu. 2023a. Matscinlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. *arXiv preprint arXiv:2305.08264*.

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

Matthew C Swain and Jacqueline M Cole. 2016. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Vineeth Venugopal, Sourav Sahoo, Mohd Zaki, Manish Agarwal, Nitya Nand Gosvami, and NM Anoop Krishnan. 2021. Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, 2(7).

Hongchen Wang, Kangming Li, Scott Ramsay, Yao Fehlis, Edward Kim, and Jason Hattrick-Simpers. 2024a. Evaluating the performance and robustness of llms in materials science q&a and property predictions. *arXiv preprint arXiv:2409.14572*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, and 1 others. 2024b. Searching for best practices in retrieval-augmented generation. *arXiv preprint arXiv:2407.01219*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024c. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *Preprint*, arXiv:2307.10635.

Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021. ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

Mohd Zaki, NM Krishnan, and 1 others. 2023. Mascqa: A question answering dataset for investigating materials science knowledge of large language models. *arXiv preprint arXiv:2308.09115*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Xianrui Zhong, Yufeng Du, Siru Ouyang, Ming Zhong, Tingfeng Luo, Qirong Ho, Hao Peng, Heng Ji, and Jiawei Han. 2024. Actionie: Action extraction from scientific literature with programming languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12656–12671.

A License

This dataset is licensed under the MIT License. Future works are free to use, modify, and distribute this dataset in accordance with the terms of the MIT License.

B Prompt Details

Prompt for generating candidate questions:

```
Here is the "abstract" of a materials science
paper. Please complete the following tasks:

1. Summarize the purpose of the paper in clear
and concise terms.
2. Classify the purpose as emphasizing
"<method>" or "<result>".
3. Identify research questions relevant to the
abstract's themes and materials science
interests.

"Abstract": {paper[key_abstract]}
```

Prompt for generating candidate answers:

```
I will provide the purpose of a materials
science paper, related research questions, and
a detailed section of the paper.

Your tasks:
    1. Select the Most Relevant Question:
Choose the research question that is most
specific, clearly phrased, and directly
related to the provided section.
    2. Refine the Question: Modify the selected
question to ensure it is:
    - Grounded on information from the
provided section, but answerable even
without using the provided section.
    - Standalone and unambiguous. Do not
use definite articles when referring to
compounds.
    - Clearly phrased for precision.
    3. Generate a Direct Answer: Provide a
concise and well-structured response that:
    - Directly answers the question.
    - Is based on the provided section but
remains meaningful out of context.
    - Avoids vague references such as
"this study" or "this paragraph."
    - Clearly conveys the information
without requiring the reader to see the
original section.

Present the output as a JSON shown below:
{{
    "question": "A clear and specific
question.",
    "answer": "A concise and relevant answer
that remains meaningful without additional
context."
}}

    Input data:
    - "Purpose and related questions":
{llm_curate_abstract}
    - "Detailed section":
{paper[key_detail]}
```

Prompt for merging candidate answers:

```
### Instructions:
1. Review the above solutions.
2. Generate an improved and refined solution by
aggregating the strengths from the provided
solutions. Enclose the solution within
<SOLUTION> and </SOLUTION> tag.
3. Provide a brief explanation of your
reasoning.
```

```
4. Ensure your answer is clear, concise, and
structured logically.

### Here is the problem:
"question": {question},

### Reference Solutions:
Solution 1: {answer1}

Solution 2: {answer2}

Solution 3: {answer3}
```

Prompt for LLM-as-Judge:

```
Your task is to evaluate the accuracy of
LLM-generated answers to materials science
questions by comparing them to expert-validated
"gold" answers.

For each evaluation, you will receive:
        - A materials science question
        - A gold answer, based on authoritative
domain knowledge
        - An LLM-generated inference answer,
which you must assess

Your goal is to evaluate how well the inference
answer aligns with the gold answer in terms of
factual accuracy, conceptual completeness, and
relevance.

Use the following evaluation rubric:
    - Correct: The inference answer fully
captures all essential concepts from the gold
answer, with no significant omissions or
factual errors.
    - Mostly Correct: The inference answer
conveys the main idea or correct conclusion,
even if minor details are missing or slight
inaccuracies are present. Additional
non-conflicting information is acceptable.
    - Incorrect: The inference answer
demonstrates substantial misunderstanding,
includes major factual errors, or omits core
concepts present in the gold answer.

Provide a short justification for your rating,
highlighting key similarities or discrepancies
between the inference and gold answers. Output
your response in the following JSON format:
{{
    "reasoning": "A concise explanation
supporting your judgment.",
    "judgment":
"correct|mostly correct|incorrect"
}}
```

C Keywords for candidate question filtering

We use the following keywords to filter vague candidate questions that make direct reference to the context:
["this", "article", "paper", "study", "abstract", "experimen", "studied", "investigat", "discuss", "describ", "example", "document", "section", "the finding","the sample"].

D Error Analysis

D.1 Cases where LLM admits that it does not have knowledge about the question

Question: How does destruction of Mn4CaO5 cluster affect conformational flexibility of CP43 lumenal domain?
LLM Answer:
I'm not aware of any information regarding how the destruction of the Mn4CaO5 cluster affects the conformational flexibility of the CP43 lumenal domain. The Mn4CaO5 cluster is a critical component of the photosynthetic water-oxidizing complex (WOC) in photosystem II, a complex protein in photosynthetic organisms. The CP43 lumenal domain is part of this WOC.

The Mn4CaO5 cluster is essential for the water-oxidation reaction during photosynthesis, where light energy is used to split water into oxygen, protons, and electrons. The destruction of the Mn4CaO5 cluster can be caused by various factors, such as oxidative stress, mutations, or chemical modifications, and it can disrupt the water-oxidation process.

However, I am unable to provide any detailed information regarding how this affects the conformational flexibility of the CP43 lumenal domain. To answer this question, I would need to look into more specific research related to this topic.

D.2 Case where LLM only partially address the question (Unaddressed part is bolded)

Question: How does the band structure of methylammonium lead iodide influence its intrinsic semiconductor behavior **despite the presence of defects and impurities?**
LLM Answer:
The band structure of methylammonium lead iodide ($MAPbI_3$), a widely studied perovskite material, is characterized by a direct band gap at the Brillouin zone center (τ-point), with the valence band maximum primarily derived from the iodine 5p orbitals and the conduction band minimum from the lead 6p orbitals. This direct band gap enables efficient light absorption and emission, making $MAPbI_3$ an excellent absorber for photovoltaic applications. The relatively small band gap (1.55 eV) allows absorption across the visible spectrum, while strong spin-orbit coupling due to lead

further influences the band structure, enhancing carrier mobility. The presence of organic cations (methylammonium) has minimal direct impact on the electronic states near the band edges but contributes to structural stability, allowing for flexible tuning of the material's optoelectronic properties.

D.3 Case where LLM introduces incorrect scientific facts (Incorrect facts are bolded)

Question: How do variations in polysarcosine block length affect protein adsorption and surface passivation in block copolymers?

LLM Answer:

Polysarcosine (PSA) is a water-soluble, biocompatible, and biodegradable polymer that has been studied extensively in the context of biomedical applications, including drug delivery, tissue engineering, and surface modification. Block copolymers comprising polysarcosine and another polymer (often referred to as PSA-based block copolymers) are of particular interest due to their ability to self-assemble into nanostructures with unique properties.

Variations in polysarcosine block length in PSA-based block copolymers can significantly affect protein adsorption and surface passivation. Here's a general overview of how different block lengths can impact these properties:

1. **Protein Adsorption:**

* Shorter PSA blocks: When the PSA block is relatively short (e.g., 4-10 monomers), the block copolymer **tends to adsorb more proteins** from the surrounding environment due to the relatively hydrophobic nature of the core block (often a hydrophobic polymer like poly(L-lysine) or poly(ethylene oxide)). This is because the hydrophobic core can interact more strongly with hydrophobic parts of proteins, leading to enhanced protein adsorption.

* Longer PSA blocks: Conversely, when the PSA block is longer (e.g., 20-50 monomers), the block copolymer tends to exhibit reduced protein adsorption. This is because the increased hydrophilicity of the PSA block dominates the surface, creating a more hydrophilic interface that repels proteins.

2. **Surface Passivation:**

* Shorter PSA blocks: Short PSA blocks can lead to poor surface passivation due to the underlying hydrophobic core being more accessible to the environment, resulting in a higher likelihood of protein adsorption and biofouling.

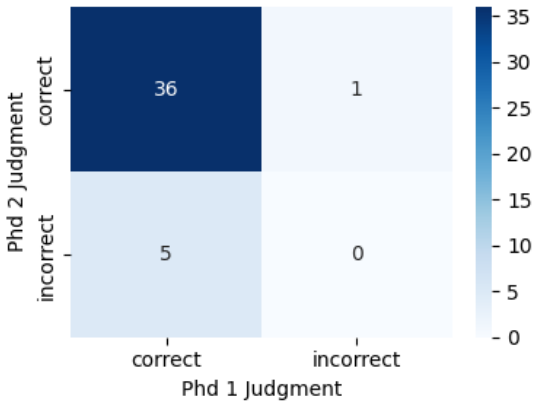

Figure 6: Confusion matrix between two material science expert annotators.

* Longer PSA blocks: Longer PSA blocks are **more effective at surface passivation**. They create a more stable and hydrophilic surface that reduces protein adsorption, minimizing the risk of biofouling and promoting a more stable interface with the surrounding environment.

In summary, the length of the polysarcosine block in PSA-based block copolymers plays a significant role in determining protein adsorption and surface passivation properties. Shorter PSA blocks tend to facilitate protein adsorption, while longer PSA blocks enhance surface passivation.

E Annotation Details

E.1 Annotation Guideline

The annotation guideline is shown in Figure 7. We detail the criteria for the annotation of QA pairs in MSQA.

E.2 Annotation Format

Screenshot of the annotation survey is shown in Figure 8. For question quality, we ask the annotator to annotate whether the question is a "Good Question" or "Bad Question". For answer quality, we ask them to identify whether the answers are "Correct", "Mostly Correct", or "Incorrect", matching the judgment choices of LLM judge. If the annotator selected "Mostly Correct" or "Incorrect", they are required to specify their reasoning.

E.3 Annotation Agreement

Confusion matrix for annotation agreement between two PhD students on the "gold" generated data for MSQA is shown in Figure 6.

Annotating the Question instruction:
Select "Good Question" or "Bad Question".

A Good Question appears reasonable at first glance and does not exhibit any of the following issues.
A Bad Question has one or more of the following problems:

- **Unclear**: Lacks necessary context (e.g., mentions “this study” or uses undefined labels for compounds).
- **Irrelevant**: Not related to material science.
- **Too Broad / Oversimplified**: Lacks specificity or depth (e.g., “What is metal?”).
- **Factually Incorrect**: Contains factual inaccuracies (e.g., “Why does copper melt at 10°C?”).
- **Other**: If the question is flawed but doesn't fall into the categories above, please also mark it as Bad Question.

Annotating the Answer instruction:
Select "Correct", "Mostly Correct", "Incorrect". If you select "Incorrect" or "Mostly Correct", please check all the reasons why the answer is not correct.

A Correct Answer: Clearly and directly answers the question and has no factual errors.
A Mostly Correct or Incorrect Answer may include following problems:
(Note: “Mostly correct answer” may have problems during explanation but doesn't affect the final answer correctness)

- **Irrelevant Response**: Does not address the question (e.g., Question: “What is Young’s modulus?” Answer: “Steel is strong.”)
- **Factual Errors**: Contains incorrect scientific information (e.g., wrong values, misstated principles).
- **Unclear Logic**: Reasoning is confusing or disjointed.
- **Misses Key Details**: Omits critical information needed for a full explanation.
- **Other:** Specify any additional issues in the “Other” field.

Figure 7: Annotation Survey Guideline

Question:
"What are the energetic and structural differences between the C3v and Oh conformers of xenon hexafluoride, and how does the transition state facilitate interconversion?"

- ◯ Good Question
- ◯ Bad Question

Answer:
"The C_{3v} conformer of xenon hexafluoride (XeF_6) is the global minimum structure, lying 186 cm^{-1} lower in energy than the O_h conformer when zero-point vibrational energy is included. Structurally, the O_h conformer exhibits equivalent Xe-F bond lengths of 1.922 Å, whereas the C_{3v} conformer distorts these bonds, with three elongating to 1.9548 Å and the opposite three shortening to 1.8512 Å. The interconversion between these conformers occurs via a low-lying transition state with a barrier of only 36 cm^{-1}, making XeF_6 highly fluxional and rapidly interconverting between the eight degenerate C_{3v} structures through the O_h minimum even at low temperatures.The O_h conformer is typically absent in noble gas matrices at low temperatures (e.g., 5 K) due to matrix interactions stabilizing the C_{3v} structure and the inability to overcome the low interconversion barrier. Vibrational spectroscopy further supports the dominance of the C_{3v} conformer, as the absence of the O_h t_{1u} band near 590 cm^{-1} indicates its instability under these conditions. Overall, the low energy barrier and fluxional nature of XeF_6 contribute to its dynamic behavior and structural variability."

- ◯ Correct
- ◯ Mostly Correct
- ◯ Incorrect

If you selected "Mostly Correct" or "Incorrect" above, please check **all** reasons why the answer is not correct.

- ☐ Irrelevant Response
- ☐ Factual Errors
- ☐ Unclear Logic
- ☐ Miss Key Details
- ☐ Other

Figure 8: Annotation Survey Format