Probing the limitations of multimodal language models for chemistry and materials research

Nawaf Alampara [1], Mara Schilling-Wilhelmi [1], Martiño Ríos-García [1],
Indrajeet Mandal [2], Pranav Khetarpal [3], Hargun Singh Grover[3],
N. M. Anoop Krishnan [3,4, ✉], and Kevin Maik Jablonka [1,5,6,7, ✉]

[1]Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany
[2]School of Interdisciplinary Research, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
[3]Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
[4]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
[5]Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany
[6]Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstrasse 12-14, 07743 Jena, Germany
[7]Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany
✉krishnan@iitd.ac.in and mail@kjablonka.com

November 27, 2024

Abstract

Recent advancements in artificial intelligence have sparked interest in scientific assistants that could support researchers across the full spectrum of scientific workflows, from literature review to experimental design and data analysis. A key capability for such systems is the ability to process and reason about scientific information in both visual and textual forms—from interpreting spectroscopic data to understanding laboratory setups. Here, we introduce MaCBench, a comprehensive benchmark for evaluating how vision-language models handle real-world chemistry and materials science tasks across three core aspects: data extraction, experimental understanding, and results interpretation. Through a systematic evaluation of leading models, we find that while these systems show promising capabilities in basic perception tasks—achieving near-perfect performance in equipment identification and standardized data extraction—they exhibit fundamental limitations in spatial reasoning, cross-modal information synthesis, and multi-step logical inference. Our insights have important implications beyond chemistry and materials science, suggesting that developing reliable multimodal AI scientific assistants may require advances in curating suitable training data and approaches to training those models.

1

1 Introduction

The practice of science has always required assimilating and integrating diverse forms of information, from visual observations in the laboratory and measurements to theoretical frameworks and prior literature. While automation has traditionally excelled at repetitive tasks such as high-throughput experimentation,[1–4] capturing the fundamental characteristic of scientific work — the ability to interpret and connect multiple modes of information flexibly — has remained a central challenge for scientific discovery.

Recent advances in artificial intelligence, particularly in large language models (LLMs), have sparked renewed interest in developing more flexible computational systems for scientific workflows. These models can orchestrate specialized tools and combine general reasoning capabilities with domain-specific functions, suggesting a path toward more adaptable scientific automation.[5–11] However, a fundamental challenge persists: bridging the gap between human scientists' natural ability to seamlessly integrate visual, numerical, and textual information and the current limitations of computational systems in processing these different data types. This gap becomes particularly apparent in tasks that require combining visual interpretation with scientific reasoning, such as analyzing spectroscopic data,[12] interpreting experimental setups,[13] or evaluating safety conditions in laboratories.[14,15]

Recent work has shown promising capabilities of LLMs in scientific tasks, from literature mining[16–23] and property prediction[10,24–30] to experiment planning.[31–34] Similarly, Vision Large Language Models (VLLMs) have demonstrated increasing capabilities in general visual reasoning tasks.[35–39] While recent benchmarks have evaluated either the scientific reasoning capabilities of language models[40,41] or general multimodal abilities,[35,36,42,43] a systematic evaluation of how these models handle the interplay of different modalities across the entire scientific process has been missing. This raises a crucial question: What are the limits of these models as copilots accelerating materials and chemistry research involving multimodal information extraction, simulations or experiments, and data analysis?

To address this gap, we present MaCBench (materials and chemistry benchmark), a comprehensive benchmark that evaluates multimodal capabilities across three fundamental pillars of the scientific process: information extraction from the literature, experiment execution, and data interpretation. By focusing on these pillars, we can assess models' abilities across the full spectrum of scientific tasks, from understanding published results to executing and interpreting new experiments. Our benchmark is distinctively designed to not only measure performance but also to uncover the underlying failure modes of current models systematically. Through carefully constructed ablation studies, we investigate how performance varies across different modalities, levels of domain expertise required, reasoning complexity, and the distance to the training data corpus. This systematic approach allows us to test the hypothesis that current models might rely on superficial pattern matching rather than deeper scientific understanding. Our results reveal that while models can handle certain modalities individually, they often fail when tasks require flexible integration of information types—a core capability required for scientific work. For instance, models might

correctly perceive information but struggle to connect these observations in scientifically meaningful ways.

These insights have important implications for developing AI-powered scientific assistants and self-driving laboratories. Our results highlight the specific capabilities needing improvement for these systems to become reliable partners in scientific discovery. They also suggest that fundamental advances in multimodal integration and scientific reasoning may be needed before these systems can truly assist in the creative aspects of scientific work.

2 Results

2.1 The MaCBench framework

Our benchmark design is guided by the observation that scientific work requires not only access to multiple modalities of information but the ability to flexibly integrate them. To probe these capabilities of VLLMs meaningfully — rather than creating artificial question-answer-based challenges — we focus on tasks that mirror real scientific workflows, from interpreting scientific literature to evaluating laboratory conditions and analyzing experimental data (see Figure 1). This approach allows us to evaluate the models' ability to process different types of information and their capacity to use this information to support scientific discovery.

The benchmark is structured around three key aspects that form the basis of many scientific workflows: information extraction, in silico or laboratory experiments, and data interpretation. Within each pillar, we include tasks spanning various scientific activities (see Figure 2). The information extraction pillar analyzes the performance in parsing scientific literature, including extracting data from tables and plots and interpreting chemical structures. The experiment execution pillar evaluates the models' ability to understand laboratory protocols, identify equipment, assess safety conditions, and understand crystal structures (as potential simulation artifacts). The data interpretation pillar tests models' capability to analyze various types of scientific data, from spectral analysis to electronic structure interpretation.

2.2 Performance landscape

There is significant variation in model performance across different task types and modalities (Figure 3, see Table A.1 for detailed descriptions of all tasks). However, when averaged over different tasks, Claude 3.5 Sonnet is the leading model on all three task families. In addition, it is interesting to note that the models do not fail at one specific part of the scientific process but struggle in all of them, suggesting that broader automation is not hindered by one bottleneck but requires advances on multiple fronts. Interestingly, even for the first step of the scientific process — data extraction — some models do not perform much better than random guessing (e.g., Llama 3.2 90B Vision in Figure 3). Current systems
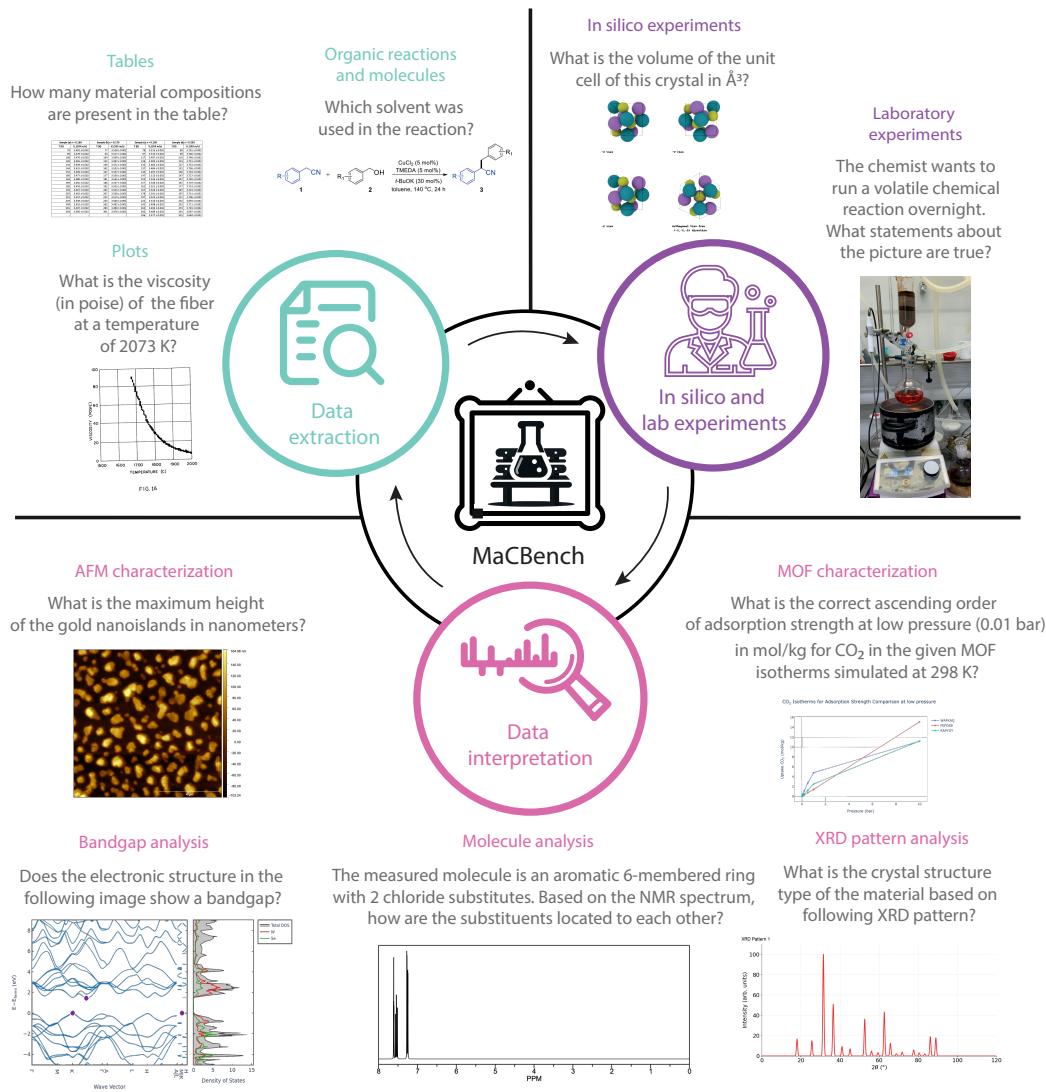
— Example questions —

Tables

How many material compositions are present in the table?

Organic reactions and molecules

Which solvent was used in the reaction?

In silico experiments

What is the volume of the unit cell of this crystal in Å³?

Laboratory experiments

The chemist wants to run a volatile chemical reaction overnight. What statements about the picture are true?

Plots

What is the viscosity (in poise) of the fiber at a temperature of 2073 K?

Data extraction

In silico and lab experiments

MaCBench

AFM characterization

What is the maximum height of the gold nanoislands in nanometers?

Data interpretation

MOF characterization

What is the correct ascending order of adsorption strength at low pressure (0.01 bar) in mol/kg for CO_2 in the given MOF isotherms simulated at 298 K?

Bandgap analysis

Does the electronic structure in the following image show a bandgap?

Molecule analysis

The measured molecule is an aromatic 6-membered ring with 2 chloride substitutes. Based on the NMR spectrum, how are the substituents located to each other?

XRD pattern analysis

What is the crystal structure type of the material based on following XRD pattern?

Figure 1: Overview of the MaCBench framework, covering the multimodal chemistry and materials science research life cycle. The framework evaluates VLLM performance across three key domains: data extraction (teal), in silico and laboratory experiments (purple), and data interpretation (pink). The benchmark includes diverse tasks spanning tables, plots, organic chemistry diagrams, crystal structures, atomic force microscopy (AFM) imaging, spectroscopy, and materials characterization. Each task requires domain-specific visual understanding and scientific reasoning, from extracting numerical values to analyzing complex experimental setups and interpreting spectroscopic data. We use icons created by Rainy Ting (on `svgrepo.com`).
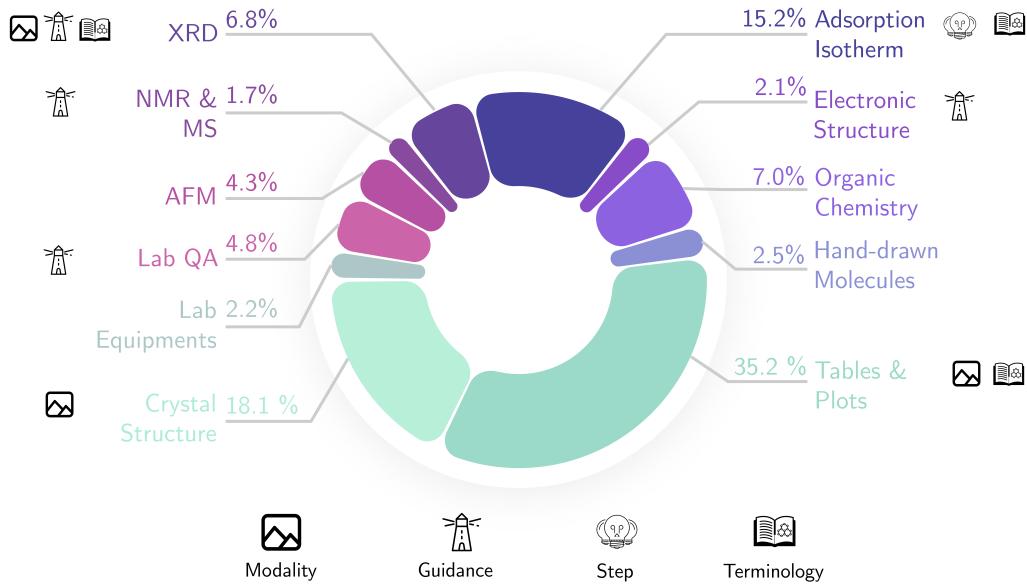
Figure 2: Distribution of tasks in the MaCBench dataset. MaCBench comprises nine distinct task categories with their respective proportions, ranging from Tables & Plots (35.2 %) to mass spectrometry (MS) & nuclear magnetic resonance (NMR) analysis (1.7 %). Each segment is annotated with relevant icons indicating the ablations we conducted on those tasks: modality understanding (image icon), guidance requirements (lighthouse icon), reasoning steps (lightbulb icon), and terminology complexity (book icon). The chart illustrates the benchmark's comprehensive coverage of chemistry and materials tasks.

tend to perform best on multiple-choice-based perception tasks (e.g., lab equipment and hand-drawn molecules in Figure 3).
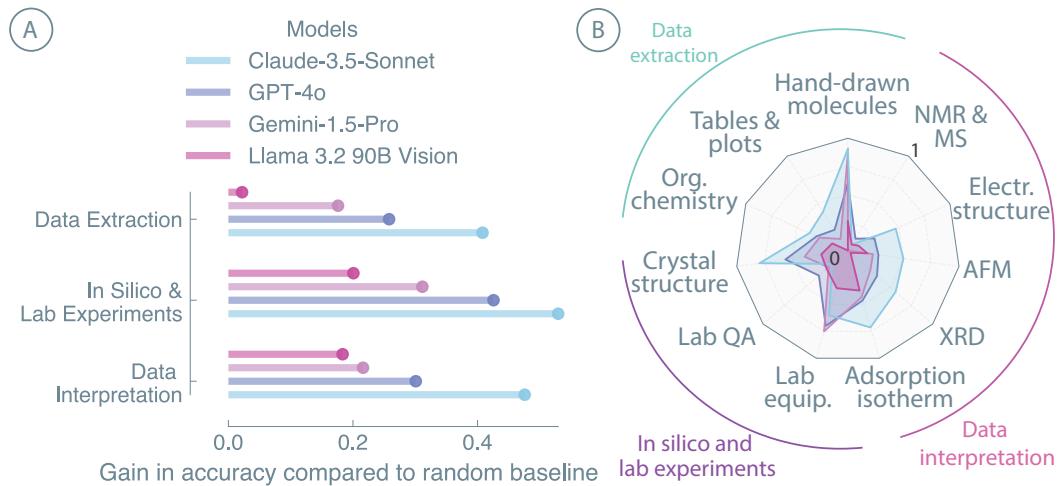

Figure 3: Performance of frontier VLLMs. a. Accuracy gains compared to random baseline across three core scientific tasks, showing varied performance of Claude 3.5 Sonnet, GPT-4o, Gemini Pro, and Llama 3.2 90B Vision in averaged across all task in the three focus areas of MaCBench: data extraction, experimental understanding, and interpretation tasks. We show the performance as the fraction of correctly answered questions relative to a random baseline. A performance of 0 means that the model is indistinguishable from random guessing. **b.** Radar plot demonstrating the relative model performance across ten specialized scientific domains. Again, we show the fraction of correctly answered questions relative to a random baseline (the plots without the normalization are shown in Figure A.1). We can observe substantial differences in performance across topics.

Data extraction   Interestingly, our analysis shows that the first step of the scientific workflow, data extraction, already poses considerable challenges for the models we tested. This is particularly the case for extracting science-specific data, for instance, about organic reactions and molecules. While the best models perform well at extracting information about reaction diagrams, they fail to correctly describe the relationship between isomers (see Figure A.3). As discussed below, this is likely caused by models struggling with spatial reasoning. In addition, even the extraction of compositions from tables still shows room for improvement for the VLLMs we tested (average accuracy of 0.48), performing not distinguishable from random guessing for Llama 3.2 90B Vision.

In silico and lab experiments   A similar variance in performance is observed for tasks related to the execution of laboratory or in silico experiments. While models show good

performance in recognizing laboratory equipment (average accuracy of 0.80), reasoning about lab scenarios, for example, comparing the safety hazards of two similar lab setups, shows low performance (average accuracy of 0.43).

The disparity between equipment identification and safety assessment performance suggests that while models can learn to recognize standard laboratory equipment, they still struggle with the more complex reasoning required for safe laboratory operations, questioning their ability to assist in real-world experiment planning and execution. This finding also implicates that current models cannot bridge gaps in tacit knowledge frequently discussed in biosafety scenarios.[44,45]

Also, the interpretation of crystal structure renderings, a crucial step for in silico experiments, shows performance that is indistinguishable from random guessing in some cases, such as the assignment of space groups (see Figure A.2).

Data interpretation   Interpreting experimental results often proves challenging to all models, including Claude 3.5 Sonnet. While most models can interpret capacity values (average accuracy of 0.61), compare Henry constants (average accuracy of 0.88) from MOF isotherms, or identify basic features such as bandgaps in bandstructure with acceptable performance (average accuracy of 0.43, 30% improvement over baseline), they struggle to interpret AFM images (average accuracy of 0.24) and often fail with tasks involving measurements like width and length (despite the presence of clear legends). They also fail to reliably interpret MS and NMR spectra (average accuracy of 0.3) or to make inferences on X-ray diffraction (XRD) pattern. In the latter case, it is particularly striking that while some models perform very well in identifying the positions of the most intense reflexes, they perform poorly in determining relative orderings, crucial for interpreting XRD patterns.

2.3   Understanding model limitations

To further understand the failure modes of VLLMs, we designed a comprehensive suite of ablation studies. Our approach isolates specific aspects of scientific tasks, from the complexity of reasoning required to how information is presented. We probe two distinct categories of limitations (Figure 4): first, core reasoning limitations that appear fundamental to current model architectures, and second, sensitivities to inference choices.

Core Reasoning Limitations   Some limitations appear intrinsic to current model architectures and are unlikely to be overcome regardless of how tasks are presented or prompted. These fundamental constraints manifest in three key areas.

Spatial reasoning   While one might expect VLLMs to excel at processing spatial information, our results reveal significant limitations in this capability. For example, while models achieve high performance in matching hand-drawn molecules to simplified molecular input line-entry system (SMILES) strings (average accuracy of 0.77, three times better
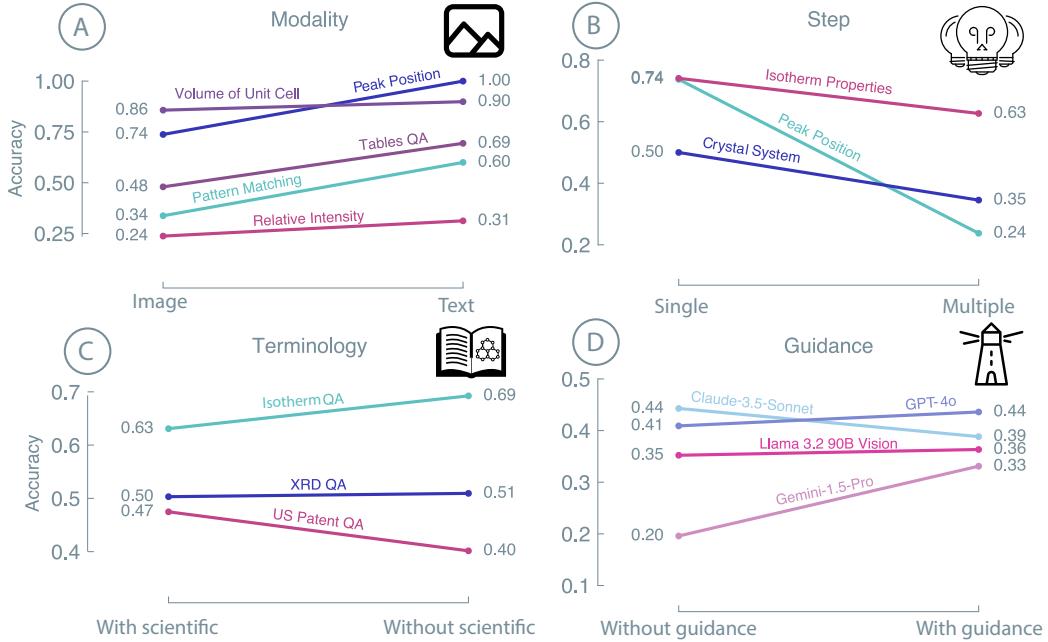
Figure 4: Ablation study results across four key dimensions of VLLMs performance in chemistry and materials science tasks. a. Modality analysis compares performance between image-only and text-only inputs across different task types, with typically higher performance when the same information is shown in text form. **b.** Step complexity analysis demonstrates performance degradation as tasks require multiple reasoning steps. **c.** Terminology impact shows how scientific language specificity affects model accuracy, comparing performance with and without domain-specific terminology. We found the behavior on US Patent QA to be mostly due to the sensitivity of Gemini Pro to the prompt template (see Appendix A.6) **d.** The guidance study compares performance across different VLLMs (Claude 3.5 Sonnet, GPT-4o, Llama 3.2 90B Vision, and Gemini Pro) with and without additional task guidance, revealing model-specific sensitivity to prompting strategies.

than baseline), they perform almost indistinguishably from random guessing for naming the isomeric relationship between two compounds (e.g., enantiomer, regioisomer, average accuracy, and baseline is 0.25) and when assigning stereochemistry (average accuracy of 0.51, baseline is 0.16). Similarly, models perform well in simple perception tasks on crystal structures (e.g., counting the number of different species, average accuracy of 0.82) but struggle at assigning the crystal system (average accuracy of 0.5) or space groups (average accuracy of 0.38).

These striking performance drops for tasks requiring spatial reasoning suggest that current VLLMs cannot reliably be used for any tasks requiring this capability — even though this might be one of the most intuitive use cases of these models.

Synthesis across modalities   Given that models input visual and textual input the same way, one might expect that the same information is processed in the same way regardless of how it is presented to the model.

To probe the ability of models to integrate information across modalities, we presented identical information in both text and image. In Figure 4, we find that for all tasks where we show the same information as images and text, the performance in the text modality is better than when the information is provided as an image. A striking example emerges when calculating the volume of crystal structures. Models show a four percentage point difference in performance when presented with the same structural information in visual form (unit cell parameters shown in the image) versus textual form (unit cell parameters shown in text). These results suggest that current models have not yet developed robust strategies for cross-modal information synthesis.

Multi-step reasoning   Motivated by the fact that the overall performance analysis indicated that perception tasks tended to be best, we designed experiments in which we probe, with the same inputs,[46] the performance on very similar tasks but requiring different numbers of reasoning steps (or different numbers of tool calls when implemented in an agentic framework).

Our analysis reveals consistent degradation in performance as tasks require more reasoning steps. Figure 4 shows that in all our experiments, the tasks requiring multiple steps perform significantly worse than those requiring only one step. For instance, in XRD pattern analysis, models perform significantly better at identifying the highest peak than at ranking relative peak intensities (average accuracy of 0.74 for identification of the highest peak against 0.24 for ranking). Similarly, for the interpretation of adsorption isotherms, accuracy in finding the highest value notably exceeds performance in ordering multiple values. This pattern suggests fundamental limitations in chaining logical steps, a crucial capability for scientific reasoning.

Sensitivity to inference choices   While addressing these core limitations will require novel training approaches, we also identified several factors that significantly influence model performance through inference choices rather than fundamental capabilities. Those factors present an actionable way to improve the performance of current systems directly without retraining them.

    Scientific terminology   One might hypothesize that models struggle with some tasks because they are unfamiliar with the scientific terminology used in the questions. Figure 4 shows that removing scientific terminology improves performance across some tasks, including the analysis of adsorption isotherms of metal-organic framework (MOF), XRD pattern interpretation. Similarly, using International Union of Pure and Applied Chemistry (IUPAC) names instead of SMILES notation for chemical compound identification leads to better results. This suggests models might be overly sensitive to specific technical vocabularies rather than understanding underlying concepts. In fact, some models like Gemini Pro (and the surrounding refusal mechanisms) are very sensitive to the exact wording of the prompt. In Appendix A.6 we show that for some questions, large variations in performance can be

    Guidance following   Given that chemists receive instructions on interpreting various experimental characterizations, we hypothesized that similar guidance might also help the models perform better on such tasks. Interestingly, adding step-by-step instructions improves performance for most models in spectral analysis, electronic structure interpretation, and XRD pattern matching—with the notable exception of Claude 3.5 Sonnet, whose performance tends to drop when provided with guidance. This variation in response to instruction suggests different underlying approaches to problem-solving across models.

2.4   Performance as a function of frequency on the internet

The varying impact of guidance across models led us to investigate whether models truly engage in scientific reasoning or primarily match patterns from their training data.[46] To probe this question, we measured the number of Google search results for various crystal structures as a proxy for the frequency of those structures in the training corpus (Figure 5).

    Our analysis reveals a striking correlation between the prominence of crystal structures on the Internet and task performance. Figure 5 shows that for all cases in our benchmark, the structures for which the models solve the tasks are more prominent on the Internet. This suggests that models might rely more on pattern matching than genuine scientific reasoning. Interestingly, we observe this effect even for tasks that depend solely on perception, such as counting the number of distinct atomic species.
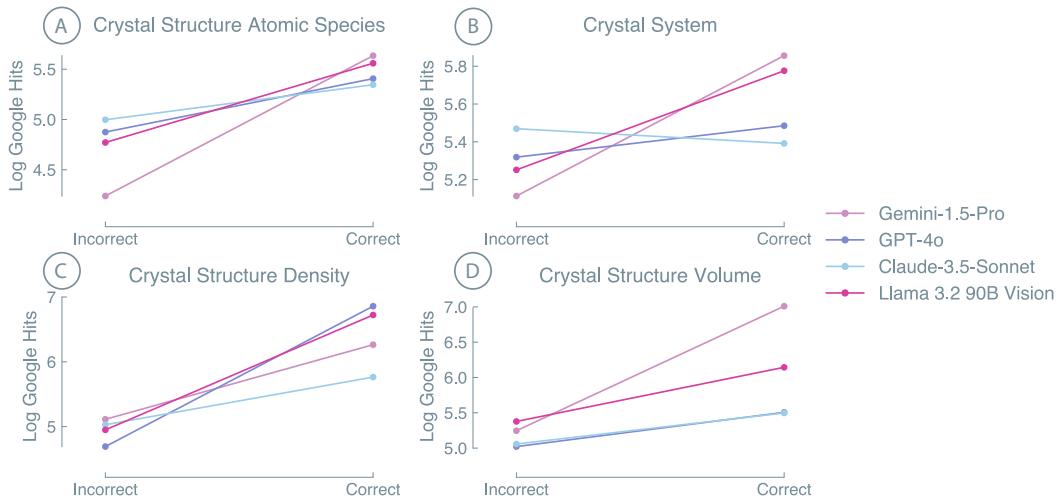
Figure 5: VLLM performance as a function of number of search hits. The plots compare four leading VLLMs across different crystallographic tasks: **a.** atomic species identification, **b.** crystal system classification, **c.** density calculation, and **d.** volume determination. For each property, the log-scale Google hit counts are plotted against the correctness of model responses, revealing correlations between answer accuracy and the prevalence of information in online sources. Higher hit counts for correct answers suggest models may not solely rely on reasoning in their responses to crystal structure analysis tasks.

11

2.5 Toward robust multimodal assistants

Our analysis reveals the promise and limitations of state-of-the-art VLLMs in scientific tasks. Compared to text-only benchmarks such as the one of Mirza *et al.*[40], we observe substantially higher performance variability across tasks, suggesting that multimodal systems are more fragile than LLMs. This fragility manifests in several ways: the striking performance gap between visual and textual representations of identical information indicates incomplete integration of modalities, while the strong correlation between model performance and the Internet presence of specific crystal structures raises questions about true reasoning capabilities versus pattern matching. The sensitivity to prompting choices (see Appendix A.6) and the counterintuitive finding that guidance can degrade performance for top models further underscore reliability concerns. However, our findings also point to actionable paths forward. Many observed limitations, particularly in spatial reasoning, could potentially be addressed through synthetic training data generation. When pursuing such approaches, we recommend incorporating generalization tests (e.g., evaluating spatial reasoning on larger compounds than those in training[47]) to ensure robust capability development. Furthermore, the significant performance differences between modalities suggest opportunities for improved training strategies, such as incorporating modality transformation tasks (e.g., automated conversion between spectral data representations). These targeted interventions could help bridge the gap between current capabilities and the needs of scientific workflows.

3 Conclusions

Scientific reasoning is fundamentally a multimodal process. Current vision-language models show promising capabilities in simple cases, such as identifying laboratory equipment or extracting explicit numerical values from plots. For standardized representations like SMILES notations or simple spectra, models can even achieve high accuracy in information extraction. However, model performance becomes unreliable when tasks require the integration of visual and conceptual understanding—as in complex laboratory safety assessments or crystal structure analysis.

Through careful ablation studies, we found that despite their impressive scale and training, current VLLMs require significant improvement in their vision modality as they seem to perform drastically better when the same information is shown in text instead of as an image. Moreover, the models seem to rely on pattern matching rather than developing robust scientific understanding. This becomes particularly evident in the observation that model performance correlates strongly with online prominence.

Yet, our benchmark also demonstrates the remarkable progress in AI systems' ability to process scientific information, with (almost) perfect performance achieved in several tasks. The observation that performance can be improved through careful terminology choice and task guidance (though with model-specific variations) suggests practical paths forward.

More broadly, our findings indicate that advancing AI in science requires not just model improvements but also better ways of representing scientific knowledge—particularly in addressing the observed gaps in spatial reasoning and cross-modal integration capabilities.

While current VLLMs cannot yet serve as autonomous scientific reasoners, they show promise as assistive tools when their limitations are well understood and their deployment is carefully structured around their demonstrated strengths. As we continue to develop these systems, our work suggests that advancing from pattern matching—demonstrated by the strong correlation between model performance and internet presence of crystal structures—–to true scientific reasoning may require fundamental advances in both training data curation and model architectures that can better handle spatial relationships and cross-modal information synthesis.

4   Methods

Our question curation and model evaluation methodology leverages the ChemBench framework.[40] For curation, we manually sourced questions and then created ablations based on error analyses to systematically understand failure modes (Figure 6). For most tasks, we created new images, e.g., by building and photographing lab setups or by plotting experimental data. Similar to Mirza *et al.*[40], all questions have been reviewed by multiple scientists before being entered into the corpus. In the curation process, we also recorded tolerances for each question. That is, for each numerical answer, we recorded windows within which an answer would still be deemed correct to account for natural uncertainties and noise.

Dataset   Our questions in the dataset are stored in an extended BigBench format.[48] Each question, along with its corresponding base64-encoded image, is stored in separate JSON files. To prevent potential data leakage during future model training, the BigBench canary string is included in each file. Our pipeline employs a robust templating system, allowing for the dynamic insertion of multiple images and other text template elements into questions using placeholders. This enables our pipeline to interleave images directly into question prompts in designated locations.

Evaluation   We employ ChemBench's prompt templates for instruction-tuned models, which also impose specific response formats on the models. The parsing workflow, also based on ChemBench, utilizes regex-based functions to extract answers from various scientific notations, handling both multiple-choice responses and numerical values. The regex-based parsing is backed up with an LLM extractor (e.g., Claude 3.5 Sonnet) for cases where standard parsing fails. We included the encoded images in the prompt. We used the default quality setting for each provider. That is, for Gemini Pro images will be automatically scaled up or down to fit into the allowed range (768×768 - 3072×3072), while for Claude 3.5 Sonnet if the image's long edge is more than 1568 pixels it is scaled down. For Llama 3.2 90B Vision,
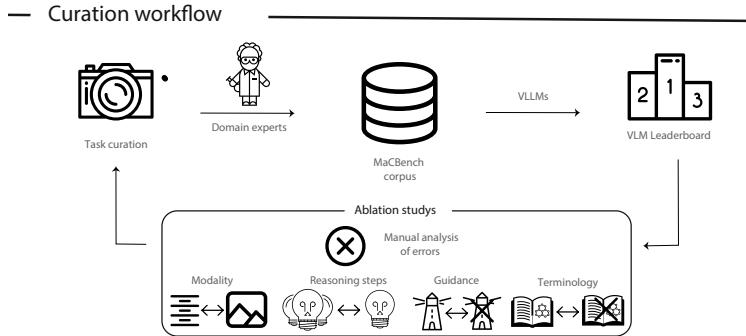
Figure 6: The MaCBench curation workflow. Tasks are initially collected and curated through manual selection, followed by validation by domain experts in chemistry and materials science. The validated tasks form the MaCBench corpus, which is used to evaluate various VLLMs, resulting in a performance leaderboard. Ablation studies are conducted through manual error analysis focusing on four key aspects: modality understanding, reasoning steps, guidance requirements, and terminology usage. Results from these analyses feed back into the task curation process, enabling continuous benchmark refinement.

an application programming interface (API) error will be raised if the images are bigger than allowed. For GPT-4o, the default configuration is set to "auto", meaning that the quality of the images is automatically selected by the API. For low-resolution images, they are set to 512×512 pixels. For the high-resolution mode, the model first sees the 512×512 image, then crops the image into 512×512 pixels tiles that are studied individually.

Refusal We implement a comprehensive framework combining regular expression-based detection from LLM Guard and a fine-tuned BERT model[49] to identify potential LLM refusals. We employed an interval-based retry mechanism to mitigate refusals on the initially refused questions. A count on the refusal by different models is shown in Table A.5.

Relative performance To account for the fact that for multiple-choice questions (MCQs) a non-zero performance can be achieved that depends on the number of options, we report the metrics in the main text as performance gains over the performance this random baseline would achieve:

$$\mathrm{acc_{rel}} = \mathrm{acc} - \mathrm{acc_{baseline}} \tag{1}$$

Correlation of performance with the number of search results For analyzing the correlation between the performance of the models and the prominence of the web, we used

the total number of results for querying the common name of crystal structures returned by the Serp API.

Acknowledgments

Data and code availability

The code and data can be found at `https://github.com/lamalab-org/mac-bench`.

Author contributions

	Nawaf Alampara	Mara Schilling-Wilhelmi	Martino Rios-Garcia	Indrajeet Mandal	Pranav Khetarpal	Hargun Singh Grover	N. M. Anoop Krishnan	Kevin Maik Jablonka
Conceptualization	■	■	■	■	■	■		■
Data curation	■	■	■	■	■	■		■
Formal analysis	■	■	■					■
Funding acquisition								■
Investigation	■	■	■	■	■	■		■
Methodology	■	■	■	■	■	■		■
Project administration								■
Resources							■	■
Software	■		■					
Supervision							■	■
Validation	■	■	■	■	■	■		■
Visualization	■	■	■					
Writing – original draft							■	
Writing – review & editing	■	■	■	■	■	■		■

References

1. Mahjour, B. *et al.* Rapid planning and analysis of high-throughput experiment arrays for reaction discovery. *Nature Communications* **14** (2023).

2. Lu, J. & Leitch, D. C. Organopalladium Catalysis as a Proving Ground for Data-Rich Approaches to Reaction Development and Quantitative Predictions. *ACS Catalysis* **13,** 15691–15707 (2023).

3. Gesmundo, N. *et al.* Miniaturization of popular reactions from the medicinal chemists' toolbox for ultrahigh-throughput experimentation. *Nature Synthesis* **2,** 1082–1091 (2023).

4. Wagen, C. C., McMinn, S. E., Kwan, E. E. & Jacobsen, E. N. Screening for generality in asymmetric catalysis. *Nature* **610,** 680–686 (2022).

5. Microsoft Research AI4Science and Microsoft Azure Quantum. The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4. *arXiv preprint arXiv:2311.07361* (2023).

6. Jimenez, C. E. *et al.* SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *arXiv preprint arXiv:2310.06770* (2024).

7. Laurent, J. M. *et al.* LAB-Bench: Measuring Capabilities of Language Models for Biology Research. *arXiv preprint arXiv:2407.10362* (2024).

8. Miret, S. & Krishnan, N. Are LLMs Ready for Real-World Materials Discovery? *arXiv preprint arXiv:2402.05200* (2024).

9. White, A. D. The future of chemistry is language. *Nature Reviews Chemistry* **7**, 457–458 (2023).

10. Jablonka, K. M. *et al.* 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digit. Discov.* **2**, 1233–1250 (2023).

11. Ramos, M. C., Collison, C. J. & White, A. D. A Review of Large Language Models and Autonomous Agents in Chemistry. *arXiv preprint arXiv:2407.01603* (2024).

12. Bushuiev, R. *et al.* MassSpecGym: A benchmark for the discovery and identification of molecules. *arXiv preprint arXiv:2410.23326* (2024).

13. Intelligent.com. *One-third of college students used CHATGPT for schoolwork during the 2022-23 academic date* https://www.intelligent.com/one-third-of-college-students-used-chatgpt-for-schoolwork-during-the-2022-23-academic-date/. 2023.

14. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nat. Mach. Intell.* **4**, 189–191 (2022).

15. Campbell, Q. L., Herington, J. & White, A. D. Censoring chemical data to mitigate dual use risk. *arXiv preprint arXiv:2304.10510* (2023).

16. Schilling-Wilhelmi, M. *et al.* From Text to Insight: Large Language Models for Materials Science Data Extraction. *arXiv preprint arXiv:2407.16867* (2024).

17. Polak, M. P. & Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications* **15**, 1–13 (2024).

18. Schilling-Wilhelmi, M. & Jablonka, K. M. *Using machine-learning and large-language-model extracted data to predict copolymerizations* in *AI for Accelerated Materials Design - Vienna 2024* (2024).

19. Ai, Q., Meng, F., Shi, J., Pelkie, B. & Coley, C. W. Extracting structured data from organic synthesis procedures using a fine-tuned large language model. *Digital Discovery* **3**, 1822–1831 (9 2024).

20. Dagdelen, J. *et al.* Structured information extraction from scientific text with large language models. *Nature Communications* **15**, 1–12 (2024).

21. Caufield, J. H. *et al.* Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. *Bioinformatics* **40** (ed Wren, J.) (2024).

22. Skarlinski, M. D. *et al.* Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740* (2024).

23. Gupta, T., Zaki, M., Krishnan, N., *et al.* DiSCoMaT: distantly supervised composition extraction from tables in materials science articles. *arXiv preprint arXiv:2207.01079* (2022).

24. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **6,** 161–169 (2024).

25. Ramos, M. C., Michtavy, S. S., Porosoff, M. D. & White, A. D. Bayesian optimization of catalysts with in-context learning. *arXiv preprint arXiv:2304.05341* (2023).

26. Zhong, Z., Zhou, K. & Mottin, D. Benchmarking Large Language Models for Molecule Prediction Tasks. *arXiv preprint arXiv:2403.05075* (2024).

27. Xie, Z. *et al.* Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules. *Chem. Sci.* **15,** 500–510 (2024).

28. Kristiadi, A. *et al.* A Sober Look at LLMs for Material Discovery: Are They Actually Good for Bayesian Optimization Over Molecules? *arXiv preprint arXiv:2402.05015* (2024).

29. Gruver, N. *et al.* Fine-Tuned Language Models Generate Stable Inorganic Materials as Text. *arXiv preprint arXiv:2402.04379* (2024).

30. Alampara, N., Miret, S. & Jablonka, K. M. MatText: Do Language Models Need More than Text & Scale for Materials Modeling? *arXiv preprint arXiv:2406.17295* (2024).

31. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624,** 570–578 (2023).

32. Darvish, K. *et al.* ORGANA: A Robotic Assistant for Automated Chemistry Experimentation and Characterization. *arXiv preprint arXiv:2401.06949* (2024).

33. M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6,** 525–535 (2024).

34. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation. *bioRxiv,* 2024–11 (2024).

35. Lu, P. *et al.* Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *arXiv preprint arXiv:2209.09513* (2022).

36. Gupta, H. *et al.* Polymath: A Challenging Multi-modal Mathematical Reasoning Benchmark. *arXiv preprint arXiv:2410.14702* (2024).

37. Cheng, K. *et al. Vision-Language Models Can Self-Improve Reasoning via Reflection* 2024.

38. Zou, C. *et al.* DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models. *arXiv preprint arXiv:2411.00836* (2024).

39. Shao, H. *et al.* Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. *arXiv preprint arXiv:2403.16999* (2024).

40. Mirza, A. *et al.* Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475* (2024).

41. Zaki, M., Jayadeva, Mausam & Krishnan, N. M. A. MaScQA: investigating materials science knowledge of large language models. *Digital Discovery* **3,** 313–327 (2024).

42. Wang, X. *et al.* SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. *arXiv preprint arXiv:2307.10635* (2024).

43. Zhang, R. *et al.* MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? *arXiv preprint arXiv:2403.14624* (2024).

44. Barrett, A. M., Jackson, K., Murphy, E. R., Madkour, N. & Newman, J. Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. *arXiv preprint arXiv:2405.10986* (2024).

45. Sandbrink, J. B. *Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools* 2023. arXiv: `2306.13952` `[cs.CY]`.

46. McCoy, R. T., Yao, S., Friedman, D., Hardy, M. & Griffiths, T. L. *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve* 2023. arXiv: `2309.13638` `[cs.CL]`.

47. Anil, C. *et al.* *Exploring Length Generalization in Large Language Models* 2022. arXiv: `2207.04901` `[cs.CL]`.

48. Srivastava, A. *et al.* Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2023).

49. ProtectAI.com. *Fine-Tuned DistilRoberta-Base for Rejection in the output Detection* 2024.

A Appendix

A.1 Desired properties of a chemistry and materials based multimodal benchmark

- *Evaluation of the cognitive abilities of VLLMs.* The main requirement of a benchmark is to evaluate the performance of the current leading models in a set of robust, extensive, and representative tasks.

- *Generalization on all real-world problems.* For fields such as chemistry or materials science, VLLMs are intended to help the scientists in their daily tasks, going from lab safety assistant to assisting in the planning and interpretation of the experimental work.

- *Help to identify the limitations of the models.* To make future VLLMs more useful to the scientists, the benchmarks must identify current limitations and light the path to more practical models.

- *Highlight strengths of the models.* Many of the current capabilities of VLLMs are still undiscovered. Showing light on these capacities can increase the usefulness of the current models.

- *Image-text integration.* A key indicator of the performance of VLLMs is how well they can join and understand image and text inputs to produce meaningful outputs.

- *Evaluation of the performance in noisy images.* To test the models' performance in complex tasks, include out-of-distribution tasks that evaluate the models' robustness against noise and atypical data.

- *Task versatility*. Include tasks that include the possible scientific scenarios; these can include visual reasoning, visual data extraction, or visual interpretation.

A.2 Related work

The rapid development of VLLMs,[1–4] has led to the publication of numerous benchmarks focused on some domains such as the medical,[5] math,[6–8] general science,[9,10] or general knowledge benchmarks.[11–16] In addition, some interesting benchmarks have been published focusing on chemistry, materials science, and related fields. Therefore, Laurent *et al.*[17] created a benchmark to evaluate LLM-powered agents. In the benchmark, they defined some tasks as multimodal images and tables, which evaluate the agents' capabilities in biological settings. Li *et al.*[9] created a broad scientific benchmark by extracting figures from some open-source general science journals and prompting the models with questions about them. Thus, the authors designed different visual tasks to evaluate its ChemVLM model and enhance their textual benchmark.[18] Roberts *et al.*[19] created a benchmark focused on

evaluating the interpretation and understanding of different scientific figures. Similarly, Khalighinejad *et al.*[20] build a benchmark that is specifically focused on evaluating the data extraction capabilities of VLLMs in extracting polymers data from full scientific articles. While the tasks and areas studied by the previous benchmarks reveal important insights, we target the focus of MaCBench on the uncovered areas and tasks, such as the Lab Protocols, to fill the gaps in our comprehension of the models' capabilities in chemistry and materials science.

A.3 Tasks in the MaCBench corpus

To unveil the proficiency of the models, we carefully designed a set of specific tasks that we consider essential parts of the scientific workflow in the chemical sciences. Table A.1 include the name, number of questions, and descriptions for all the main tasks in the MaCBench corpus.

Model performance on the main tasks   As mentioned in the main text, we evaluated some leading VLLMs. Table A.2 collects the overall performance of the models along the different tasks. In that table, we also include the random baseline results, which are used as the base for the overall performance figure of the main text (see Figure 3).

Similarly, to better illustrate the overall results, Figure A.1 visually describes the performance of the models along all the MaCBench tasks, including the random baseline as the fifth model.

Table A.1: Number of questions and description of all the tasks in the MaCBench corpus. We grouped tasks in themes corresponding to typical tasks in the scientific workflow in the chemical sciences. Those groups correspond to the ones shown in the radar plots.

Topic	Nº of Questions	Description
Data extraction	517	
Hand-drawn Molecules	29	Systematic naming of hand-drawn organic molecules
Organic Chemistry		
Chirality	25	Determination of the number of chiral centers in molecules, including their configuration, spatial orientation, and priority groups
Isomers	20	Identification of isomeric relationships between two molecules
Organic Molecules	15	Systematic naming of organic molecules following IUPAC nomenclature
Organic Reactions Schema	4	Extraction of components such as solvents, temperature, or yield from organic reaction schemas
Organic Schema without SMILES	17	Analysis of organic reaction schemas with visual references for molecule identification.
Tables and plots		
Tables	308	Analysis of composition tables
US Patent Figures	63	Extraction of information from scientific figures in US patents
US Patent Plots	36	Interpretation of 2D plots presented in US patents
In silico and lab experiments	289	
Lab QA		
Lab Protocol Comparison	17	Comparison of laboratory images to identify correct practices and violations of good laboratory standards
Lab Protocol	38	Review of images taken in a chemistry lab focusing on safety protocols and proper laboratory practices
Lab Equipments	25	Identification and classification of laboratory glassware and other equipment
CIF QA		
Crystal Structure Atomic Species	41	Determination of the number of different atomic species from crystal structure images
Crystal Structure Density	42	Determination of the density from crystal structure images
Crystal Structure Symmetry	42	Determination of the point group from crystal structure images
Crystal Structure Volume	42	Determination of the volume from crystal structure images
Crystal System	42	Determination of the crystal system from crystal structure images
Data interpretation	349	
AFM Image Analysis	50	Analysis of topography in various specimens using an atomic force microscope.
Adsorption Isotherm		
Adsorption Isotherm Capacity Comparison	19	Comparison of the capacities of adsorption isotherms
Adsorption Isotherm Capacity Order	20	Ordering of capacities of adsorption isotherms
Adsorption Isotherm Capacity Value	20	Determination of the capacity value from adsorption isotherms
Adsorption Isotherm Henry Constant Comparison	10	Comparison of the Henry's constants of adsorption isotherms
Adsorption Isotherm Henry Constant Order	12	Ordering of Henry's constants of adsorption isotherms
Adsorption Isotherm Strength Comparison	15	Comparison of the adsorption strengths of isotherms
Adsorption Isotherm Strength Order	19	Ordering of adsorption strengths of isotherms
Adsorption Isotherm Working Capacity Comparison	20	Comparison of the working capacity of adsorption isotherms
Adsorption Isotherm Working Capacity Order	20	Ordering of working capacities of adsorption isotherms
Adsorption Isotherm Working Capacity Value	20	Determination of the working capacity value from adsorption isotherms
Electronic Structure	24	Analysis of the electronic structure of materials, such as direct or indirect bandgap and metallic characteristics
NMR and MS Spectra	20	Identification of halide atoms using MS isotope patterns and substitution positions on benzene rings using 1H NMR spectra
XRD QA		
XRD Pattern Matching	20	Determination of crystal type from a XRD pattern
XRD Pattern Shape	20	Selection of the crystalline or amorphous nature from a XRD pattern
XRD Peak Position	20	Determination of the peak position of most intense peak from a XRD pattern
XRD Relative Intensity	20	Ordering of the peak positions of the three most intense peaks from XRD pattern
Overall	1155	

Table A.2: Absolute performance of the models on the MaCBench corpus classified by the three pillars considered in MaCBench. Note that in this table, the random baseline is included as a model.

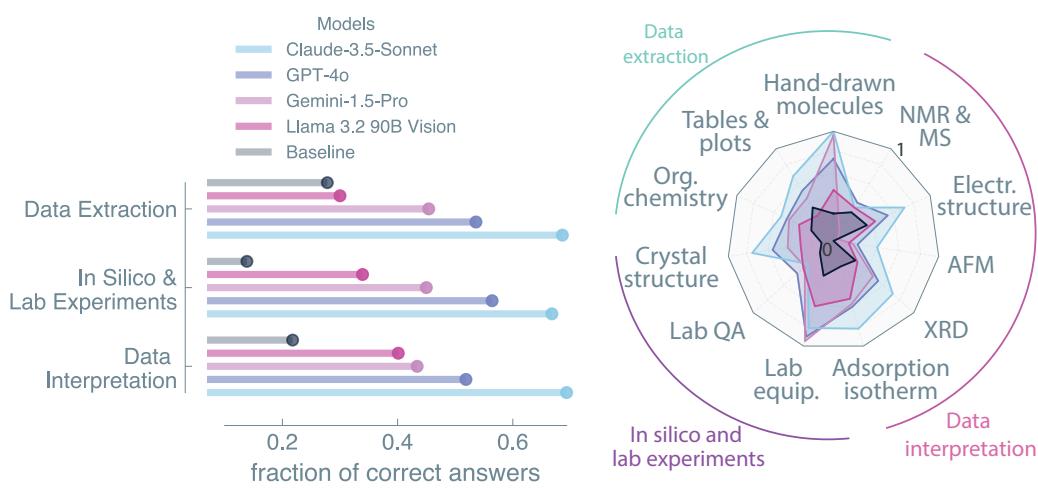|	Baseline	Claude-3.5-Sonnet	GPT-4o	Gemini-1.5-Pro	Llama 3.2 90B Vision
Data extraction					
Hand-drawn Molecules	0.24	**0.97**	0.72	0.93	0.45
Organic Chemistry					
Chirality	0.16	**0.64**	0.52	0.56	0.32
Isomers	0.25	**0.3**	0.25	0.25	0.2
Organic Molecules	0.27	**0.8**	0.73	0.6	0.47
Organic Reactions Schema	0.5	**1.0**	0.75	0.75	0.0
Organic Schema without SMILES	0.18	**0.82**	0.71	0.76	0.53
Tables and plots					
Tables	0.35	**0.68**	0.53	0.45	0.27
US Patent Figures	0.1	**0.67**	0.56	0.38	0.38
US Patent Plots	0.08	**0.72**	0.5	0.19	0.33
In silico and lab experiments					
Lab QA					
Lab Protocol Comparison	0.24	0.41	**0.53**	0.35	0.41
Lab Protocol	0.13	0.26	**0.39**	0.34	0.34
Lab Equipments	0.32	0.8	0.88	**0.92**	0.6
CIF QA					
Crystal Structure Atomic Species	0.0	**0.95**	0.85	0.78	0.71
Crystal Structure Density	0.05	**0.52**	0.33	0.26	0.24
Crystal Structure Symmetry	0.26	**0.64**	0.24	0.48	0.19
Crystal Structure Volume	0.02	**0.9**	0.81	0.19	0.07
Crystal System	0.21	**0.71**	0.57	0.4	0.31
Data interpretation					
AFM Image Analysis	0.0	**0.4**	0.22	0.18	0.14
Adsorption Isotherm					
Adsorption Isotherm Capacity Comparison	0.37	**1.0**	0.89	0.95	0.68
Adsorption Isotherm Capacity Order	0.25	**0.85**	0.55	0.65	0.65
Adsorption Isotherm Capacity Value	0.2	**0.7**	0.55	**0.7**	0.5
Adsorption Isotherm Henry Constant Comparison	0.2	**1.0**	0.9	0.7	0.9
Adsorption Isotherm Henry Constant Order	0.25	**0.83**	0.75	0.5	0.67
Adsorption Isotherm Strength Comparison	0.13	**0.93**	0.6	0.57	0.6
Adsorption Isotherm Strength Order	0.26	**0.74**	**0.74**	0.63	0.32
Adsorption Isotherm Working Capacity Comparison	0.25	**0.75**	0.5	0.55	0.55
Adsorption Isotherm Working Capacity Order	0.15	**0.75**	0.6	0.45	0.45
Adsorption Isotherm Working Capacity Value	0.25	**0.65**	0.2	0.15	0.25
Electronic Structure	0.33	**0.71**	0.54	0.04	0.42
NMR and MS Spectra	0.3	0.35	**0.4**	0.1	0.35
XRD QA					
XRD Pattern Matching	0.2	**0.45**	0.3	0.3	0.3
XRD Pattern Shape	0.3	**0.95**	0.85	0.7	0.3
XRD Peak Position	0.25	**1.0**	0.8	0.7	0.45
XRD Relative Intensity	0.3	**0.45**	0.2	0.2	0.1
Overall Score	0.22	**0.71**	0.57	0.49	0.4

Figure A.1: Performance of frontier vision-language models across scientific tasks, organized by the three pillars of the scientific process: information extraction, experiment execution, and data interpretation. While models show strong performance in certain basic tasks, their capabilities vary significantly when deeper scientific reasoning is required.

A.4 Ablation studies and systematic elucidation of failure modes

To further elucidate the capabilities and limitations of VLLMs we created a set of tests intended to shed light on the strengths and limitations of these models. Most of these tests were created using the same images as for the main corpus of MaCBench, but changing the textual part of the questions. Table A.3 describes each test, highlighting the differences from the original tasks.

Performance   Table A.4 lists the performance in all our systematic failure mode elucidation experiments.
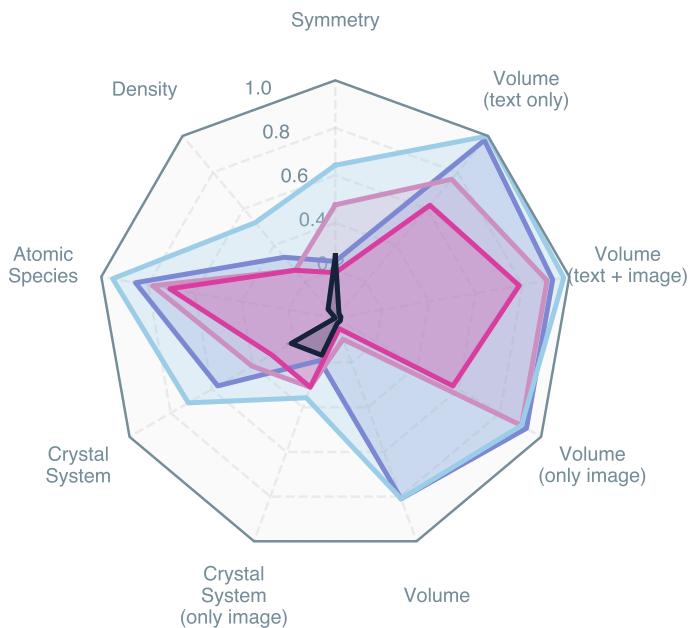

Figure A.2: VLLMs performance in tasks dealing with the interpretation of crystal structure renderings.

Crystal structure analysis   In Figure A.2 we show the performance of VLLMs on tasks concerning the analysis of crystal structures. To probe the influence of the modality on the performance, we showed the lattice parameters in only the text, only in the image, or in text and image. Interestingly, the performance changes depending on the modalities in which information is shown. In addition, the plot highlights that models show low performance for tasks requiring spatial reasoning, e.g., the assignment of the space group or crystal system. In the case of the assignment of the crystal system, we see that adding the lattice

Table A.3: Descriptions for the different ablations performed. Note that multi-step tasks are the same as some tasks in MaCBench corpus. This is because multi-step reasoning is needed to solve the questions associated with these tasks.

Ablation	Nº of Questions	Description
Modality	410	
Composition Tables (Ablation)	308	Evaluation of tabular data with text-based tuple representations instead of images.
Crystal Structure Volume Text	42	calculation of crystal structure volume with lattice parameters given in text and image
XRD Pattern Matching Text	20	Determination of crystal type from a XRD pattern given as text
XRD Peak Position (Ablation)	20	Determination of the peak position of most intense peak from a XRD pattern given the intensity and theta values as text
XRD Relative Intensity (Ablation)	20	Ordering of the peak positions of the three most intense peaks from XRD pattern indicating in the text part of the question the intensity and theta values.
Step	133	
Adsorption Isotherm Capacity Order	20	Ordering of capacities of adsorption isotherms
Adsorption Isotherm Henry Constant Order	12	Ordering of Henry's constants of adsorption isotherms
Adsorption Isotherm Strength Order	19	Ordering of adsorption strengths of isotherms
Adsorption Isotherm Working Capacity Order	20	Ordering of working capacities of adsorption isotherms
Crystal Structure Density	42	Determination of the density from crystal structure images
XRD Relative Intensity	20	Ordering of the peak positions of the three most intense peaks from XRD pattern
Terminology	337	
Adsorption Isotherm Capacity Comparison (Ablation)	20	Comparison of capacity of adsorption isotherms, avoiding scientific terminology
Adsorption Isotherm Capacity Order (Ablation)	20	Ordering of capacity of adsorption isotherms, avoiding scientific terminology
Adsorption Isotherm Capacity Value (Ablation)	20	Determination of the capacity value of adsorption isotherms, avoiding scientific terminology
Adsorption Isotherm Henry Constant Comparison (Ablation)	10	Comparison of Henry constants of adsorption isotherms, avoiding scientific terminology
Adsorption Isotherm Henry Constant Order (Ablation)	11	Ordering of the Henry constants of adsorption isotherms, avoiding scientific terminology
Adsorption Isotherm Strength Comparison (Ablation)	19	Comparison of the adsorption strength of isotherms, avoiding scientific terminology
Adsorption Isotherm Strength Order (Ablation)	18	Ordering of adsorption strength of isotherms, avoiding scientific terminology
Adsorption Isotherm Working Capacity Comparison (Ablation)	20	Comparison of working capacities of adsorption isotherms, avoiding scientific terminology
Adsorption Isotherm Working Capacity Order (Ablation)	20	Ordering of the working capacity of adsorption isotherms, avoiding scientific terminology
Adsorption Isotherm Working Capacity Value (Ablation)	20	Determination of working capacity of adsorption isotherms, avoiding scientific terminology
Pattern Shape (Ablation)	20	Adsorption isotherm pattern shape log (Ablation), avoiding scientific terminology
Peak Position (Ablation)	20	Determination of the peak position in an XRD pattern with explanation on how to get this
Relative Intensity (Ablation)	20	Ordering of the peak positions of the three most intense peaks from XRD pattern, avoiding scientific terminology
US Patent Figures (Ablation)	63	Interpretation of patent figures avoiding the use of technical jargon.
US Patent Plots (Ablation)	36	Interpretation of patent plots with plain language, avoiding complex terminology.
Guidance	102	
Electronic Structure with Knowledge	24	Investigation of electronic structures with instructions on how to solve the specific tasks
Lab Protocol (Ablation)	38	Examination of chemistry lab images with an emphasis on safety protocols, proper practices, and adherence to laboratory safety rules.
NMR and MS Spectra with Explanation	20	Analysis of halogen atom patterns with detailed instructions for interpreting MS isotope data.
Pattern Matching (Ablation)	20	Determination of crystal type from a XRD pattern, avoiding scientific terminology
Overall	982	

Table A.4: Absolute performance of the different models in all the failure mode elucidation experiments.

	Baseline	Claude-3.5-Sonnet	GPT-4o	Gemini-1.5-Pro	Llama 3.2 90B Vision
Modality					
Composition Tables (Ablation)	0.69	**0.80**	0.69	0.64	0.65
Crystal Structure Volume Text	0.02	**0.95**	0.83	0.90	0.55
XRD Pattern Matching Text	0.25	**0.70**	0.45	0.65	0.60
XRD Peak Position (Ablation)	0.20	**1.00**	**1.00**	**1.00**	**1.00**
XRD Relative Intensity (Ablation)	0.35	**0.40**	0.30	0.25	0.30
Steps					
Adsorption Isotherm Capacity Order	0.25	**0.85**	0.55	0.65	0.65
Adsorption Isotherm Henry Constant Order	0.25	**0.83**	0.75	0.50	0.67
Adsorption Isotherm Strength Order	0.26	**0.74**	**0.74**	0.63	0.32
Adsorption Isotherm Working Capacity Order	0.15	**0.75**	0.60	0.45	0.45
Crystal Structure Density	0.05	**0.52**	0.33	0.26	0.24
XRD Relative Intensity	0.30	**0.45**	0.20	0.20	0.10
Terminology					
Adsorption Isotherm Capacity Comparison (Ablation)	0.35	**0.80**	0.70	**0.80**	0.55
Adsorption Isotherm Capacity Order (Ablation)	0.20	**0.95**	**0.95**	0.75	0.65
Adsorption Isotherm Capacity Value (Ablation)	0.20	**0.80**	0.65	0.55	0.35
Adsorption Isotherm Henry Constant Comparison (Ablation)	0.40	**1.00**	0.90	0.70	0.70
Adsorption Isotherm Henry Constant Order (Ablation)	0.36	**1.00**	0.91	0.36	0.91
Adsorption Isotherm Strength Comparison (Ablation)	0.26	**0.84**	0.58	0.74	0.42
Adsorption Isotherm Strength Order (Ablation)	0.33	**1.00**	0.94	0.78	0.83
Adsorption Isotherm Working Capacity Comparison (Ablation)	0.50	**0.90**	0.80	0.75	0.55
Adsorption Isotherm Working Capacity Order (Ablation)	0.35	**0.85**	0.70	0.35	0.25
Adsorption Isotherm Working Capacity Value (Ablation)	0.25	**0.95**	0.30	0.50	0.20
Pattern Shape (Ablation)	0.35	**0.75**	**0.75**	0.70	0.20
Peak Position (Ablation)	0.30	**0.90**	**0.90**	0.80	0.75
Relative Intensity (Ablation)	0.30	**0.40**	0.35	0.20	0.05
US Patent Figures (Ablation)	0.14	**0.65**	0.52	0.22	0.32
US Patent Plots (Ablation)	0.06	**0.64**	0.33	0.14	0.31
Guidance					
Electronic Structure with Knowledge	0.38	0.46	**0.58**	0.46	0.42
Lab Protocol (Ablation)	0.05	**0.39**	0.21	0.32	0.24
NMR and MS Spectra with Explanation	0.15	0.35	**0.60**	0.25	0.40
Pattern Matching (Ablation)	**0.40**	0.35	0.35	0.30	**0.40**

parameters to the image (which is by default included in all questions) helps the model perform better compared to only having access to the rendering of the structure.
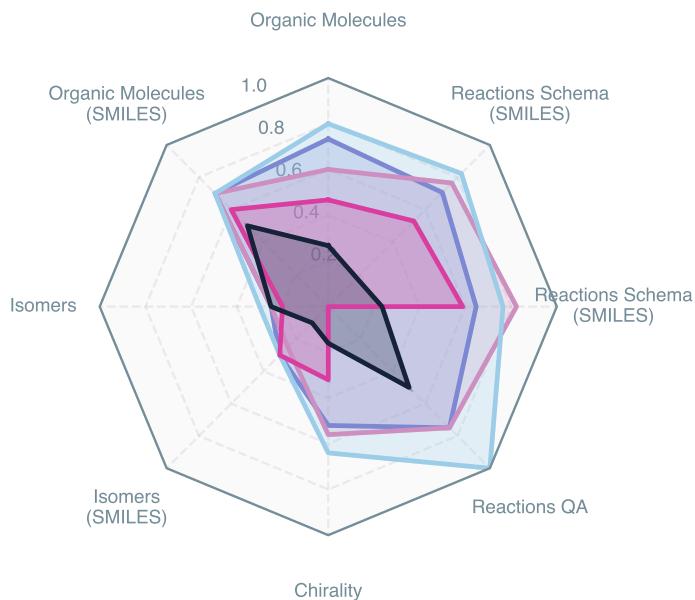

Figure A.3: VLLMs performance for questions related to organic molecules and reactions in MaCBench.

Organic chemistry performance   In Figure A.3 we show the performance in tasks related to renderings of organic molecules and reactions. One of the most striking observations is the low performance in tasks related to identifying isomeric relationships between molecules. Here, the models perform comparably to the baseline in the vision modality and only slightly better than the baseline when provided with SMILES as text. Similar limitations in spatial reasoning are probably the reason for low performance in tasks related to the assignment of chiral centers.

A.5   Refusals

By manually checking some of the answers for MaCBench, we observed that some models refused to answer some of the questions, claiming they could not answer that type of question. This is probably a consequence of the safety alignment that the models go through.[21] As a result of these observations, we counted the number of refusal response occurrences, which results are described in Table A.5. Note that the results shown in the table include the original tasks and tests. Only the tasks for which some models refused are shown. Similarly, only the models that showed refusals are shown (Claude 3.5 Sonnet, GPT-4o and Gemini Pro). Interestingly, we observe that GPT-4o refuses to answer many of the Lab Protocol questions (26% for the original questions and 55% for the ablation).

Table A.5: Number of refused answers per topic for the MaCBench corpus. The percentages for each topic are relative to each task, while the overall percentage is relative to the total number of questions and ablations in the MaCBench corpus. The topics in the table are the only ones for which refusal was observed. Similarly, note that the only models present are Claude 3.5 Sonnet, GPT-4o and Gemini Pro because only these models showed refusals. For GPT-4o, a great percentage of the refusals are observed in Lab Protocol questions and are probably triggered because of the safety training of this model. Note that prompt fragility refusals are not included in this table.

	Claude-3.5-Sonnet		GPT-4o		Gemini-1.5-Pro	
	N° of refusals	%	N° of refusals	%	N° of refusals	%
Main Corpus						
Organic Chemistry						
Chirality	1	4.00	0	0.00	0	0.00
Tables and plots						
Tables	1	0.32	0	0.00	0	0.00
US Patent Plots	0	0.00	4	11.11	8	22.22
Lab QA						
Lab Protocol Comparison	0	0.00	3	17.65	0	0.00
Lab Protocol	0	0.00	10	26.32	0	0.00
CIF QA						
Crystal Structure Density	13	30.95	0	0.00	0	0.00
Crystal Structure Symmetry	0	0.00	1	2.38	0	0.00
Adsorption Isotherm						
Adsorption Isotherm Working Capacity Value	1	5.00	0	0.00	0	0.00
AFM Image Analysis	0	0.00	1	2.00	0	0.00
NMR and MS Spectra	1	5.00	0	0.00	0	0.00
Ablations						
Guidance						
Lab Protocol (Ablation)	0	0.00	21	55.26	0	0.00
Terminology						
US Patent Figures (Ablation)	0	0.00	0	0.00	2	3.17
US Patent Plots (Ablation)	1	2.78	4	11.11	2	5.56
Overall	18	0.84	44	2.06	12	0.56

A.6 Sensitivity to prompt template

To study the sensitivity of VLLMs to prompt variations, we conducted a study in which we tested six template variations, differing only by a single word: "image" (original word), "diagram", "plot", "figure", "photograph", and "None" (leaving a space). In Figure A.4 we show variation in mean absolute error (MAE) for the task considered for these tests.
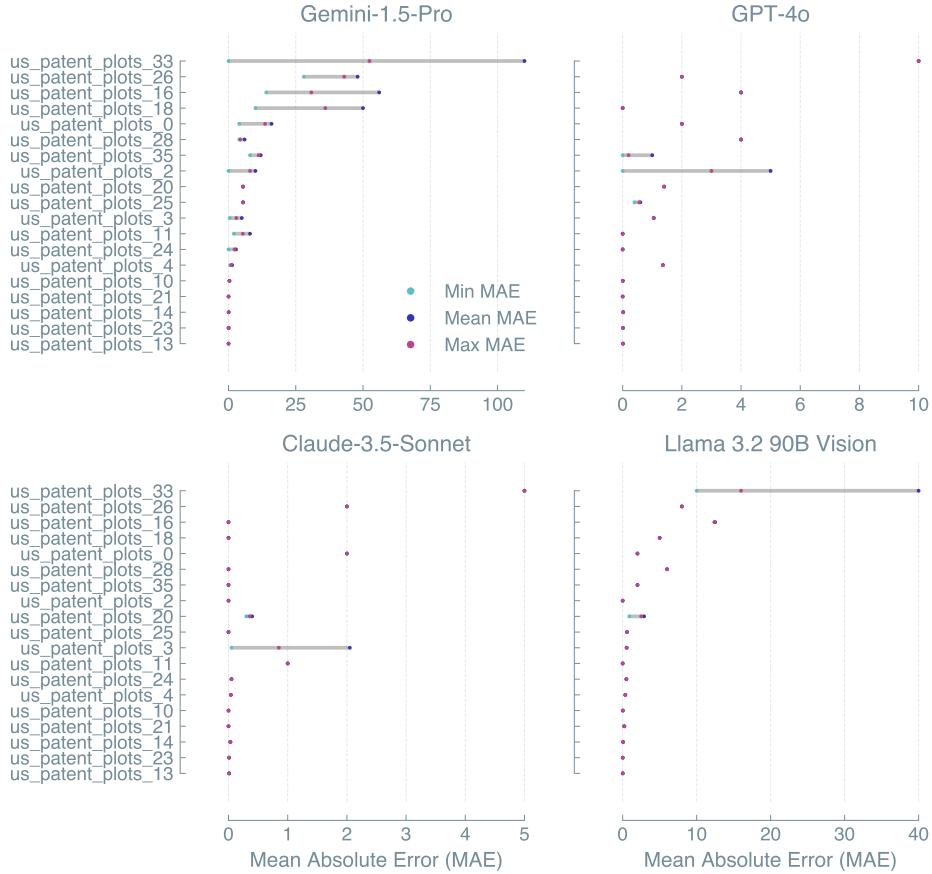

Figure A.4: VLLMs variation in performance with six different prompt templates for US Patent plots questions in MaCBench. The rows list different questions from the US Patent plots subset. The dumbbell plots show the statistics of the MAE for responses in different templates. The gray bars illustrate the variance; larger bars hence indicate larger sensitivity to the prompt template. Only questions that show variation for any model are included. We observe that Gemini Pro is highly sensitive to the prompt template.

During this prompt fragility test, we observed some refusal variations from the original task. Table A.6 summarizes those observations related to how prompt templates impact

Table A.6: Number of refused answers per different template for the US Patent plots questions. The percentages for each topic are relative to the number of questions in this task (36 questions), while the overall percentage is relative to the total number of questions for this experiment (288 questions). In the "¡None¿" template, instead of the word "image", a space was used.

	Claude-3.5-Sonnet		GPT-4o		Gemini-1.5-Pro	
	N° of refusals	%	N° of refusals	%	N° of refusals	%
¡None¿	0	0.00	4	11.11	21	58.33
Diagram	0	0.00	4	11.11	9	25.00
Figure	0	0.00	4	11.11	11	30.56
Graphic	0	0.00	4	11.11	8	22.22
Image	0	0.00	4	11.11	8	22.22
Image (Ablation)	1	2.78	4	11.11	2	5.56
Photograph	0	0.00	4	11.11	14	38.89
Plot	0	0.00	4	11.11	10	27.78
Overall	1	0.35	32	11.11	83	28.82

refusal rates. So far, the prompt template has impacted only the number of refusals of Gemini Pro, for which a high variance was observed among the different templates.

A.7 Leaderboard

To summarize the results of MaCBench, similar to ChemBench,[22] we created a leaderboard based on the toolchain developed for MatBench.[23] The online leaderboard is available at `https://lamalab-org.github.io/mac-bench/leaderboard/`.

References

1. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485* (2023).

2. OpenAI. *Hello GPT-4o* https://openai.com/index/hello-gpt-4o/. 2024.

3. Team, G. *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

4. Anthropic. *Claude 3.5 Sonnet* https://www.anthropic.com/news/claude-3-5-sonnet. 2024.

5. Jeong, D. P., Garg, S., Lipton, Z. C. & Oberst, M. Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress? *arXiv preprint arXiv:2411.04118* (2024).

6. Gupta, H. *et al.* Polymath: A Challenging Multi-modal Mathematical Reasoning Benchmark. *arXiv preprint arXiv:2410.14702* (2024).

7. Zhang, R. *et al.* MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? *arXiv preprint arXiv:2403.14624* (2024).

8. Zou, C. *et al.* DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models. *arXiv preprint arXiv:2411.00836* (2024).

9. Li, Z. *et al.* MMSci: A Dataset for Graduate-Level Multi-Discipline Multimodal Scientific Understanding. *arXiv preprint arXiv:2407.04903* (2024).

10. Liang, Z. *et al. SceMQA: A Scientific College Entrance Level Multimodal Question Answering Benchmark* in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (eds Ku, L.-W., Martins, A. & Srikumar, V.) (Association for Computational Linguistics, Bangkok, Thailand, 2024), 109–119.

11. Yue, X. *et al.* MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv preprint arXiv:2311.16502* (2024).

12. Chia, Y. K., Han, V. T. Y., Ghosal, D., Bing, L. & Poria, S. PuzzleVQA: Diagnosing Multimodal Reasoning Challenges of Language Models with Abstract Visual Patterns. *arXiv preprint arXiv:2403.13315* (2024).

13. Shao, H. *et al.* Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. *arXiv preprint arXiv:2403.16999* (2024).

14. Roberts, J. S. *et al.* Image2Struct: Benchmarking Structure Extraction for Vision-Language Models. *arXiv preprint arXiv:2410.22456* (2024).

15. Zhang, D. *et al.* MM-LLMs: Recent Advances in MultiModal Large Language Models. *arXiv preprint arXiv:2401.13601* (2024).

16. Cheng, K. *et al. Vision-Language Models Can Self-Improve Reasoning via Reflection* 2024.

17. Laurent, J. M. *et al.* LAB-Bench: Measuring Capabilities of Language Models for Biology Research. *arXiv preprint arXiv:2407.10362* (2024).

18. Zhang, D. *et al.* ChemLLM: A Chemical Large Language Model. *arXiv preprint arXiv:2402.06852* (2024).

19. Roberts, J., Han, K., Houlsby, N. & Albanie, S. SciFIBench: Benchmarking Large Multimodal Models for Scientific Figure Interpretation. *arXiv preprint arXiv:2405.08807* (2024).

20. Khalighinejad, G. *et al.* MatViX: Multimodal Information Extraction from Visually Rich Articles. *arXiv preprint arXiv:2410.20494* (2024).

21. Cui, J., Chiang, W.-L., Stoica, I. & Hsieh, C.-J. *OR-Bench: An Over-Refusal Benchmark for Large Language Models* 2024. arXiv: 2405.20947 [cs.CL].

22. Mirza, A. *et al.* Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475* (2024).

23. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials* **6,** 1–12 (2020).

Acronyms

AFM atomic force microscopy.

API application programming interface.

IUPAC International Union of Pure and Applied Chemistry.

LLM large language model.

MAE mean absolute error.

MCQ multiple-choice question.

MOF metal-organic framework.

MS mass spectrometry.

NMR nuclear magnetic resonance.

SMILES simplified molecular input line-entry system.

VLLM Vision Large Language Model.

XRD X-ray diffraction.