Benchmarking LLMs for atomic-level geometric manipulation in crystals

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advancements with video generators, language aligned robotics models and tool-augmented design frameworks suggest that large language models (LLMs) may soon no longer struggle with 3D spatial reasoning. To bring these developments into the material sciences, we present AtomWorld, a data generator and benchmark that evaluates LLMs on atomic-level operations (e.g. insert, move, rotate atoms) in CIF files. This benchmark was tested across major chat models, finding these models to generally take an algorithmic approach - which yielded successful completion of simple tasks such as adding and moving atoms, but struggled with more complex tasks such as rotating around an atom. LLM inaptitude with spatial reasoning limits their usefulness in crystallography - addressing this problem is a necessary first step towards enabling higher level tasks such as seeing motifs, symmetries, repairing or validating complex structures, and proposing novel structures.

1 Introduction

A Crystallographic Information File (CIF) [1] is the standard format for storing crystallographic structural data. Suppose that there are three stages for an LLM to reason with CIF files: motor skills, perceptual skills and cognitive skills. Motor skills are about the mechanics of geometry — being able to add, move, rotate, or insert atoms consistently within a structure. Perceptual skills are about recognising patterns — seeing motifs like octahedra, channels, or layered frameworks, and detecting symmetry or connectivity. Cognitive skills are about reasoning and creativity — engaging in hypothesis-driven modifications and proposing novel structures.

LLMs for crystallography would primarily benefit researchers at the cognitive stage, however challenges such as hypothesis-driven modification require LLMs to also be strong at the motor and perceptual stages. In current literature, perceptual skills have been tested through question-answer (QA) style benchmarks e.g. LLM4Mat-Bench [2], but less attention has been given to testing motor skills. To address this gap, our research question asks: how can we measure and improve LLM "crystallographic motor skills", i.e. ability to manipulate atoms in crystal structures? We present the following contributions:

1. AtomWorld Playground: A scalable data generator and benchmark that evaluates LLMs on atomic-level operations (e.g. add, move, rotate, insert atoms) in CIF files.
2. Obtained benchmark results across several frontier chat models. We found these models to generally take an algorithmic approach - which yielded successful completion of simple tasks such as adding and moving atoms, but struggled with more complex tasks such as rotating around an atom.

To the best of our knowledge, we are the first benchmark to evaluate LLM motor skills in crystallography. While these tasks are trivially solved via software or packages such as Ovito[3] and

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Atomic Simulation Environment(ASE)[4], installing this capability in LLMs can help unlock the more valuable downstream cognitive skills. Traditionally, LLMs have struggled with spacial reasoning tasks - but this may soon change with rapid advancements in tool-augmented design [5], diffusion LLMs [6, 7], and as language aligned video generation [8, 9] and robotics [10] models become increasingly capable. We hope that our AtomWorld playground can play a foundational role in testing the understanding of 3D CIF environments in tomorrow's LLMs.

2 Related Work

LLMs for crystallography. LLMs have been primarily explored for their capabilities in CIF generation and QA. LLMs have been demonstrated to hold an innate ability to generate crystal structures when pretrained on millions of CIF files [11]. This process may be further reinforced through evolutionary search frameworks [12]. However, as LLMs are pattern predictors, the search space is fundamentally limited by the scope of the pretraining data. LLMs can also be instruction fine-tuned to predict crystal properties or provide general QA responses from CIF, e.g. AlchemBERT[13], NatureLM[14], Darwin 1.5 [15], etc[16, 17]. Crystallography QA is well benchmarked, with the most comprehensive being LLM4Mat-Bench [2], consisting of approximately 2 million composition-structure-description pairs. Tool-augmented LLMs such as OSDA Agent [5] improve structure generation through coupling computational chemistry tools to LLMs. These tool-augmented design frameworks are able to address the lack of in-depth chemistry knowledge of LLMs without expensive (and not always effective) fine-tuning. LLMs may be able to reliably handle geometric CIF modification through tool-augmentation.

Multimodal reasoning. Approaches such as multimodal chain-of-thought (Multimodal-CoT) [18], visualization-of-thought (VoT) [19] add image modalities to the reasoning trace rather than pure textual chain-of-thought. In particular, Multimodal-CoT with under 1 billion parameters achieved state of the art in state-of-the-art performance on the ScienceQA benchmark, outperforming larger models like GPT-3.5. As CIF describes a 3D challenge, these results suggest that multimodal reasoning approaches can be highly applicable to improving LLM ability on CIF geometry tasks, as well as reasoning-intensive QA and structure generation/modification tasks. Approaches to multimodal representation may also be influenced from developments in video generation and robotics, where models such as Genie 3 [9] and V-JEPA 2 [10] are increasingly capable of understanding real-world physics and integrating this with natural language input/output. Finally, with the training objective of diffusion LLMs [6, 7] to be noise reversal, they have an advantage in understanding structural text compared to autoregressive LLMs - with LLaDA [6] surpassing GPT-4o in a reversal poem completion task. This also suggests diffusion LLMs may be inherently capable of differentiating between valid and invalid modifications to CIF - important for geometric modification tasks. Developments in multimodal reasoning and diffusion suggest that LLMs may be on the cusp of being able to grasp the 3D CIF environment, making it important to benchmark this progress.

3 Playground Design: AtomWorld

At its core, AtomWorld is designed as a scalable data generator which can be used for both benchmarking and training LLMs. The data generated follows a three-part structure: two CIF files of "before" and "after" states, and an action prompt describing the change - with the goal of the LLM to yield the "after" state, given the "before" state and action. A flowchart of the benchmarking workflow is presented in Figure 1. Detailed descriptions and examples of all supported actions prompts are found in Appendix A.1.

4 Experiments & Discussion

We benchmarked a selection of state-of-the-art LLMs, including variants from Gemini, GPT, Qwen, Deepseek, and LLaMA families. The results are summarized in Figure 2a and b. According to the evaluation metrics, it is evident that LLMs exhibit varying levels of performance across tasks. Simpler operations such as `add` are performed more consistently, whereas more spatially demanding manipulations, particularly `rotate around`, remain highly challenging. Overall, the relative task difficulty can be ordered as: `add` $<$ `move` $<$ `move_towards` $<$ `insert_between` $<$

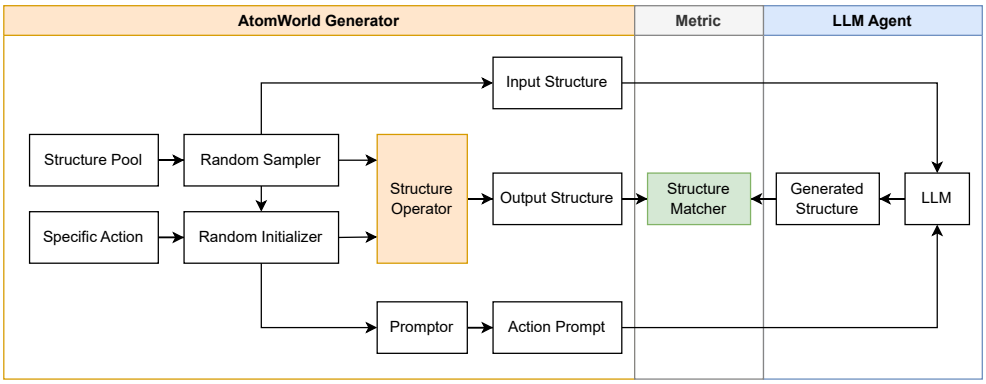

Figure 1: AtomWorld benchmark flowchart. The AtomWorld generator follows a structured data flow: the random sampler selects a structure from a predefined structure pool (in this work, a subset of CIF files from the Materials Project database[20]); the random initializer parametrizes the chosen action template by assigning atom indices and/or positions; the structure operator applies the instantiated action to the original structure to obtain the target structure; and the prompter generates a natural language description aligned with the action. The resulting (input structure, action prompt) pairs are then fed into the LLM agent system, whose generated structure is compared against the target structure using the `StructureMatcher` from *pymatgen*[21] to compute the evaluation metric (see Appendix A.2).

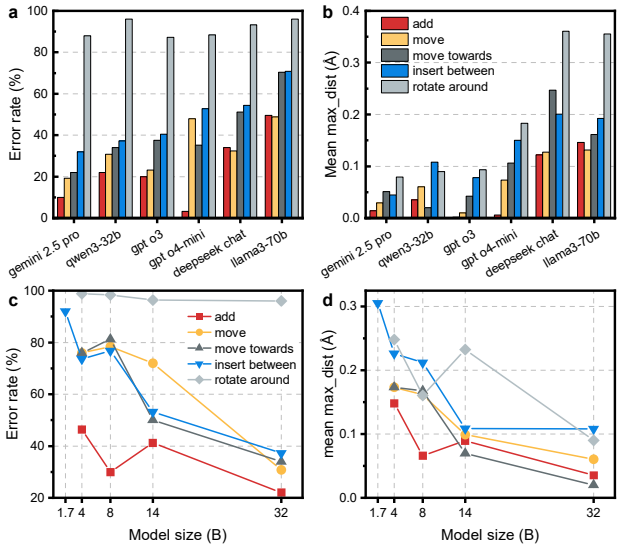

Figure 2: Evaluation results. **a** and **b** demonstrate the error rate and mean `max_dist` metrics for different actions. **c** and **d** demonstrate the change in performance with model sizes, tested using the Qwen3 series.

`rotate_around`. Gemini 2.5 Pro achieves the strongest performance across the evaluated tasks, showing particularly low error rates and displacement values in the `move`, `move_towards`, and `insert_between` tasks.

Geometric operation difficulty. To measure the inherent difficulty of each geometric operation, we tested Gemini 2.5 Pro and Deepseek V3-0324 on simplified point-based tasks, with results listed in Table 1. The models were given a set of points in three-dimensional space, expressed in raw coordinate format like "$[[x_1, y_1, z_1], [x_2, y_2, z_2]]$". The models were then asked to apply similar geometric operations directly on these points and return the transformed coordinates. This setting

removes the complexities of CIF files and serves as a controlled test of whether the LLM can handle spatial transformations at all. The results from this setting reflects the task difficulty found in AtomWorld benchmark results, observing that models perform well on simple actions like `move`, `move_towards`, and `insert_between`, but found the `rotate_around` action is significantly more difficult. The former could be solved with straightforward numerical calculations (e.g., addition or weighted averaging), which LLMs can handle reliably. In contrast, models often attempted to compute a rotation matrix for the `rotate_around` action and failed to apply it consistently, leading to high mean `max_dist`.

Table 1: Model performances on simplified point-based tasks. Error rate indicates the ratio of unreadable outputs from LLMs. Mean `max_dist` is calculated by the maximum distance between generated and target points after Hungarian sort.

	Gemini 2.5 Pro (50 frames)		Deepseek V3-0324 (250 frames)	
Action	Error rate (%)	mean `max_dist` (Å)	Error rate	mean `max_dist`
`move`	0.00	0.0000	0.00	0.0000
`move_towards`	2.00	0.0045	0.00	0.3172
`insert_between`	0.00	0.0051	21.2	0.0642
`rotate_around`	2.00	16.168	0.00	14.058

Paramater scaling. Qwen3-32B ranks second overall and is especially notable for its efficiency: despite having only 32B parameters, it outperforms or matches larger models (e.g., GPT o3, LLaMA3-70B) on several tasks. Figure 2c and d illustrate how parameter scaling of the Qwen3 series affects accuracy across tasks. In general, larger models tend to achieve lower error rates and smaller displacements, confirming that scaling improves spatial reasoning capabilities. This pattern is further supported by the Chemical Competence Score (CCS)[22], which increases with model size and highlights Qwen3-32B outperforming LLaMA3-70B. (See Appendix B.3) Nonetheless, the marginal benefits decrease with increasing model size, and for the `rotate_around` task, the improvements remain limited. These observations suggest that architectural design and training strategies play an equally important role as model scale in enabling atomic-level reasoning.

Solution approaches. Chat models generally approached these geometric challenges through generating the necessary linear algebra algorithms to solve. Failures across most CIF actions could be attributed to context-rot, as the chat models lost their train of thought across large reasoning traces. In Table 1, we found an interesting case where the Deepseek V3 model has an abnormally high error rate in the simplified `insert_between` tasks. A closer look at the wrong responses reveals that Deepseek often attempted to write a Python script to compute the coordinates, rather than directly performing the calculation.

5 Future Work & Conclusion

In this paper we presented AtomWorld as the first benchmark that evaluates LLM motor skills in crystallography. In general, we found that chat models took an algorithmic approach to solving the geometric tasks of our benchmark. With this approach, simpler operations such as add could be performed more consistently, whereas more spatially demanding manipulations, particularly rotations, remain highly challenging. These tasks are solved trivially via crystallography software, but for LLMs are an important first stage to enabling higher level tasks such as seeing motifs, symmetries, repairing or validating complex structures, and proposing novel structures.

In future work, we would like to increase the depth of our evaluation beyond frontier chat models. A stronger conclusion may be drawn about LLM capabilities through also evaluating specialised LLMs for material science, and tool-augmented LLMs. Future versions of the AtomWorld playground would likely see an expanded set of actions, prompt templates and evaluation metrics. A richer structure of modalities may also be included - e.g. graphs or visual depictions for input into multimodal LLMs.

LLMs have traditionally struggled with spacial reasoning tasks, however this may be soon to change with recent developments in tool-augmented design, diffusion, video generation and language aligned

robotics models [5, 7, 9, 10]. We hope that our AtomWorld playground can play a foundational role in helping researchers of tomorrow test LLM understanding of 3D CIF environments.

References

[1] S. R. Hall, F. H. Allen, and I. D. Brown. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47 (6):655–685, 1991. doi: https://doi.org/10.1107/S010876739101067X. URL https://onlinelibrary.wiley.com/doi/abs/10.1107/S010876739101067X. tex.eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1107/S010876739101067X.

[2] Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. LLM4Mat-bench: Benchmarking large language models for materials property prediction. *Machine Learning: Science and Technology*, 6(2):020501, May 2025. ISSN 2632-2153. doi: 10.1088/2632-2153/add3bb. Publisher: IOP Publishing.

[3] Alexander Stukowski. Visualization and analysis of atomistic simulation data with OVITO-the open visualization tool. *MODELLING AND SIMULATION IN MATERIALS SCIENCE AND ENGINEERING*, 18(1), January 2010. ISSN 0965-0393. doi: 10.1088/0965-0393/18/1/015012. Number: 015012 tex.eissn: 1361-651X tex.orcid-numbers: Stukowski, Alexander/0000-0001-6750-3401 tex.researcherid-numbers: Stukowski, Alexander/G-9695-2017 tex.unique-id: ISI:000272791800012.

[4] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, June 2017. ISSN 0953-8984. doi: 10.1088/1361-648X/aa680e. URL https://dx.doi.org/10.1088/1361-648X/aa680e. Publisher: IOP Publishing.

[5] Zhaolin Hu, Yixiao Zhou, Zhongan Wang, Xin Li, Weimin Yang, Hehe Fan, and Yi Yang. OSDA agent: Leveraging large language models for de novo design of organic structure directing agents. In *The thirteenth international conference on learning representations*, 2025. URL https://openreview.net/forum?id=9YNyiCJE3k.

[6] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large Language Diffusion Models, February 2025. URL http://arxiv.org/abs/2502.09992. arXiv:2502.09992 [cs].

[7] Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, Yuwei Fu, Jing Su, Ge Zhang, Wenhao Huang, Mingxuan Wang, Lin Yan, Xiaoying Jia, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Yonghui Wu, and Hao Zhou. Seed Diffusion: A Large-Scale Diffusion Language Model with High-Speed Inference, August 2025. URL http://arxiv.org/abs/2508.02193. arXiv:2508.02193 [cs].

[8] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing Efficient Video Production for All, December 2024. URL http://arxiv.org/abs/2412.20404. arXiv:2412.20404 [cs].

[9] Google DeepMind. Genie 3: A new frontier for world models, May 2025. URL https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models.

[10] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning, June 2025. URL http://arxiv.org/abs/2506.09985. arXiv:2506.09985 [cs].

[11] Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, December 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54639-7.

[12] Jingru Gan, Peichen Zhong, Yuanqi Du, Yanqiao Zhu, Chenru Duan, Haorui Wang, Carla P. Gomes, Kristin A. Persson, Daniel Schwalbe-Koda, and Wei Wang. Large Language Models Are Innate Crystal Structure Generators, February 2025. URL http://arxiv.org/abs/2502.20933. arXiv:2502.20933 [cond-mat].

[13] Xiaotong Liu, Yuhang Wang, Tao Yang, Xingchen Liu, and Xiaodong Wen. AlchemBERT: Exploring Lightweight Language Models for Materials Informatics, February 2025. URL https://chemrxiv.org/engage/chemrxiv/article-details/6781a6b481d2151a02a3212e.

[14] Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, Zekun Guo, Yeqi Bai, Pan Deng, Yaosen Min, Ziheng Lu, Hongxia Hao, Han Yang, Jielan Li, Chang Liu, Jia Zhang, Jianwei Zhu, Ran Bi, Kehan Wu, Wei Zhang, Kaiyuan Gao, Qizhi Pei, Qian Wang, Xixian Liu, Yanting Li, Houtian Zhu, Yeqing Lu, Mingqian Ma, Zun Wang, Tian Xie, Krzysztof Maziarz, Marwin Segler, Zhao Yang, Zilong Chen, Yu Shi, Shuxin Zheng, Lijun Wu, Chen Hu, Peggy Dai, Tie-Yan Liu, Haiguang Liu, and Tao Qin. Nature language model: Deciphering the language of nature for scientific discovery, 2025. URL https://arxiv.org/abs/2502.07527. arXiv: 2502.07527 [cs.AI].

[15] Tong Xie, Yuwei Wan, Yixuan Liu, Yuchen Zeng, Shaozhou Wang, Wenjie Zhang, Clara Grazian, Chunyu Kit, Wanli Ouyang, Dongzhan Zhou, and Bram Hoex. DARWIN 1.5: Large language models as materials science adapted learners, 2025. URL https://arxiv.org/abs/2412.11970. arXiv: 2412.11970 [cs.CL].

[16] Joren Van Herck, María Victoria Gil, Kevin Maik Jablonka, Alex Abrudan, Andy S. Anker, Mehrdad Asgari, Ben Blaiszik, Antonio Buffo, Leander Choudhury, Clemence Corminboeuf, Hilal Daglar, Amir Mohammad Elahi, Ian T. Foster, Susana Garcia, Matthew Garvin, Guillaume Godin, Lydia L. Good, Jianan Gu, Noémie Xiao Hu, Xin Jin, Tanja Junkers, Seda Keskin, Tuomas P. J. Knowles, Ruben Laplaza, Michele Lessona, Sauradeep Majumdar, Hossein Mashhadimoslem, Ruaraidh D. McIntosh, Seyed Mohamad Moosavi, Beatriz Mouriño, Francesca Nerli, Covadonga Pevida, Neda Poudineh, Mahyar Rajabi-Kochi, Kadi L. Saar, Fahimeh Hooriabad Saboor, Morteza Sagharichiha, K. J. Schmidt, Jiale Shi, Elena Simone, Dennis Svatunek, Marco Taddei, Igor Tetko, Domonkos Tolnai, Sahar Vahdatifar, Jonathan Whitmer, D. C. Florian Wieland, Regine Willumeit-Römer, Andreas Züttel, and Berend Smit. Assessment of fine-tuned large language models for real-world chemistry and material science applications. *Chemical Science*, 16(2):670–684, 2025. doi: 10.1039/D4SC04401K. Publisher: The Royal Society of Chemistry.

[17] Andrea Madotto Nate Gruver, Anuroop Sriram and Zachary Ward Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *International conference on learning representations 2024*, 2024.

[18] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models, May 2024. URL http://arxiv.org/abs/2302.00923. arXiv:2302.00923 [cs].

[19] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's Eye of LLMs: Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models, October 2024. URL http://arxiv.org/abs/2404.03622. arXiv:2404.03622 [cs].

[20] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. doi: 10.1063/1.4812323.

[21] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, February 2013. ISSN 0927-0256. doi: 10.1016/j.commatsci.2012.10.028.

[22] Andres M. Bran, Tong Xie, Shai Pranesh, Jeremy Goumaz, Xuan Vu Nguyen, David Ming Segura, Ruizhi Xu, Jeffrey Meng, Dongzhan Zhou, Wenjie Zhang, and Philippe Schwaller. MiST: Understanding the Role of Mid-Stage Scientific Training in Developing Chemical Reasoning Models. In *FM4LS 2025: Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences at ICML 2025*, July 2025.

[23] Alex M. Ganose and Anubhav Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, September 2019. ISSN 2159-6859, 2159-6867. doi: 10.1557/mrc.2019.94. URL `http://link.springer.com/10.1557/mrc.2019.94`.

A AtomWorld Setup Details

A.1 Supported action prompts

Table 2: Examples of actions and the corresponding action prompts for point-based tasks.

Action name	Action prompt
`move`	Move the point at index {index} by displacement {displacement}.
`move_towards`	Move the point at index {from_index} towards the point at index {to_index} by {distance}.
`insert_between`	Insert a new point between points at indices {index1} and {index2}, {distance} units away from point {index1}.
`rotate_around`	Rotate all points by {angle_deg} degrees around the axis {axis}, with the point at index {center_index} as the center of rotation. The rotation follows the right-hand rule.

Table 3: Examples of actions and the corresponding action prompts for AtomWorld.

Action name	Action prompt
`add`	Add one {symbol} atom at the Cartesian coordinate {position} to the cif file.
`move`	Move the atom at index {index} by {d_pos} angstrom in the cif file.
`move_towards`	Move the atom at index {index1} towards the atom at index {index2} by {distance} angstrom in the cif file.
`insert_between`	Insert a {symbol} atom in the line between atoms at indices {index1} and {index2}, and the inserted atom must be {distance:.2f} angstrom from atom at {index1} in the cif file.
`rotate_around`	Rotate all surrounding atoms within {radius} angstrom of the center atom at index {index} by {angle} degree around the axis {axis} in the cif file. The rotation should following the right-hand rule.

In addition to the actions listed above, we have also implemented several others, including `remove`, `swap`, `delete_around`, `move_selected`, etc. These actions are not presented here, as they have not yet undergone systematic evaluation.

A.2 Evaluation metrics

Two primary metrics are used to evaluate the correctness of the LLM-generated structures: the error rate and the mean maximum distance (`max_dist`).

The error rate is defined as the number of test cases exhibiting any of the following errors divided by the total number of test cases. These errors are categorized into three hierarchical levels:

1. **Wrong output format.** The LLM's response must enclose the generated structure within a predefined tag so that it can be correctly extracted from the textual output. Failure to do so constitutes an output format error.
2. **Wrong structure format.** Even if the structure is successfully extracted, its file format may still be invalid or incompatible with downstream processing tools. Such cases are counted as structure format errors.
3. **Mismatch of structures.** For structurally valid outputs, we compare them with the target structures using `StructureMatcher` with a site tolerance of 0.5. Any generated structure whose site matching exceeds this tolerance is considered a mismatch.

The second primary metric - mean `max_dist` - is computed only for structurally valid outputs that pass the tolerance check. For each matched pair of structures, we calculate the maximum pairwise atomic displacement after optimal alignment, and then average this value across all test cases. The `max_dist` metric is used because it is generally more significant than the RMSD value in our cases. This is because only a few or even a single atom is "moved" while others remain unchanged, making the maximum displacement a more representative indicator of the structural difference.

A.3 Full Prompt Templates

Listing 1: A prompt example for a specific task of AtomWorld

```
You are a CIF operation assistant. You will be given an input CIF
content and an action prompt. Your task is to apply the action
described in the action prompt to the initial CIF content. The
coordinates in the action are in Cartesian format. Return the modified
 CIF content in cif format within <cif> and </cif> tags.

Please ensure the output is a valid CIF file, with correct formula,
and atom positions.

Input CIF content:
{The specific CIF file is inserted here}

Action prompt: Insert Lu between atoms at indices 6 and 5 that is 4.03
 angstrom from atom 6.
```

A.4 Illustrative example of the framework

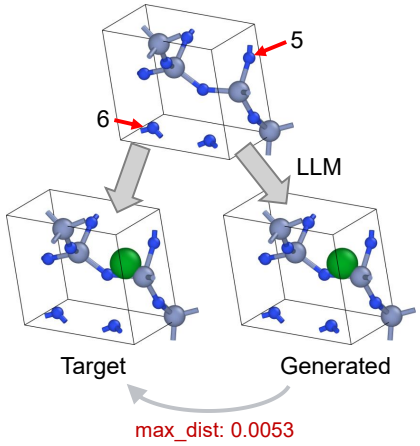

Figure 3: The workflow of a specific `insert_between` task.

To provide a concrete understanding of our proposed AtomWorld Bench, we present an illustrative example of its workflow. This case study focuses on a specific task: inserting a Lu atom between the fifth and the sixth atoms in the specific CIF structure. The prompt used here is listed in Appendix A.3. The workflow randomly selects the atom indices and determines the position of the atom to be inserted based on the selected atoms. Based on the initialized action, the framework gives out a target structure. The LLM will also generate a structure after processing the prompt, as shown in Figure 3. In this example, the two structures are nearly identical, with a `max_dist` of 0.0053 Å, indicating high accuracy.

B Full Evaluation Results

B.1 Tables for error rates and mean `max_dist`

Table 4: Model performances of `add` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	10.0	0.0140
Qwen3 32b	22.0	0.0352
ChatGPT o3	20.0	**0.0015**
ChatGPT o4-mini	**3.20**	0.0060
DeepSeek V3-0324	34.0	0.1221
Llama3 70b	49.6	0.1315

Table 5: Model performances of `move` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	**19.2**	0.0293
Qwen3 32b	30.8	0.0605
ChatGPT o3	23.2	**0.0102**
ChatGPT o4-mini	48.0	0.0734
DeepSeek V3-0324	32.4	0.1274
Llama3 70b	48.8	0.1315

Table 6: Model performances of `move_towards` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	**22.0**	0.0513
Qwen3 32b	34.0	**0.0201**
ChatGPT o3	37.6	0.0425
ChatGPT o4-mini	35.2	0.1063
DeepSeek V3-0324	51.2	0.2467
Llama3 70b	70.4	0.1613

Table 7: Model performances of `insert_between` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	**32.0**	**0.0444**
Qwen3 32b	37.2	0.1082
ChatGPT o3	40.4	0.0778
ChatGPT o4-mini	52.8	0.1501
DeepSeek V3-0324	54.4	0.2004
Llama3 70b	70.8	0.1921

Table 8: Model performances of `rotate_around` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	88.0	0.0790
Qwen3 32b	96.0	0.0900
ChatGPT o3	**87.2**	0.0933
ChatGPT o4-mini	88.4	0.1832
DeepSeek V3-0324	93.2	0.3607
Llama3 70b	96.0	0.3557

B.2 The `max_dist` violin plots

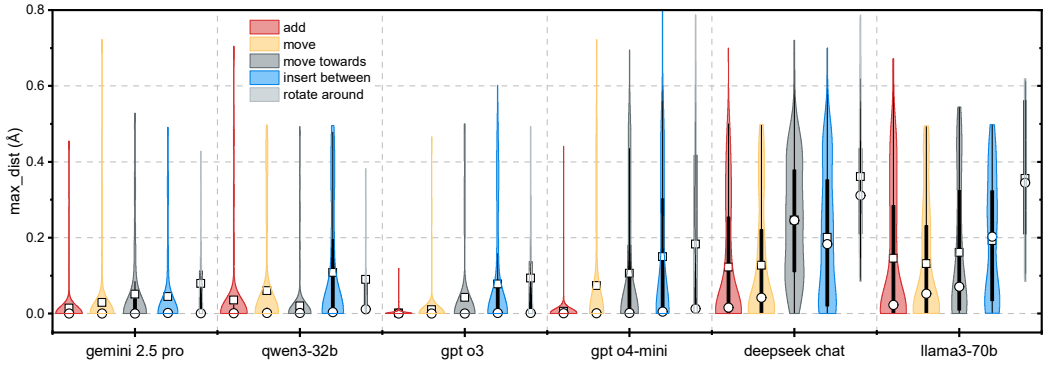

Figure 4: The violin plots of `max_dist` of evaluation results. The hollow squares indicate the mean values, and the hollow circles indicate the medians.

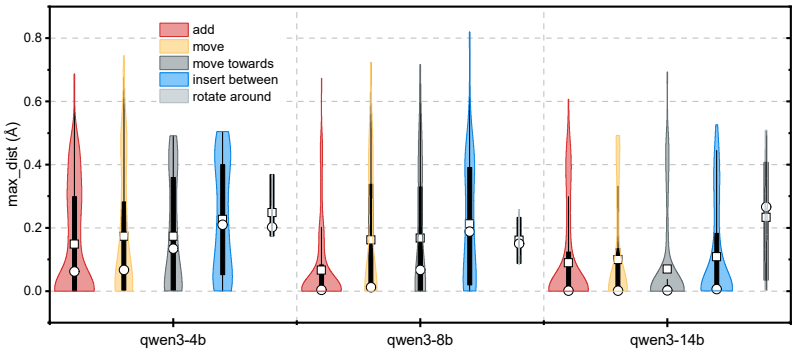

Figure 5: The violin plots of `max_dist` of evaluation results.

B.3 Chemical Competence Score

The Chemical Competence Score (CCS) is designed to assess a model's latent chemical knowledge by evaluating its precision in distinguishing chemically accurate from inaccurate descriptions of crystal structures. Following the methodology of Bran *et al.*[22], the dataset was constructed by sampling 600 unique crystal structures from the Materials Project, with corresponding descriptions generated using Robocrystallographer[23]. An inaccurate dataset was then created by replacing one sentence in each original description with a sentence describing a different crystal. Because the CCS is computed from the token log-likelihoods at the model's final layer, access to these probabilities is required; consequently, the score was calculated only for the open-source models benchmarked in this study. The resulting scores are reported in Table 9 and visualised in Fig. 6.

Table 9: CCS score of open-source models

Model	CCS
Qwen3 4B	0.768
Qwen3 8B	0.829
Qwen3 14B	1.061
Qwen3 32B	**1.141**
Llama3 70b	0.987

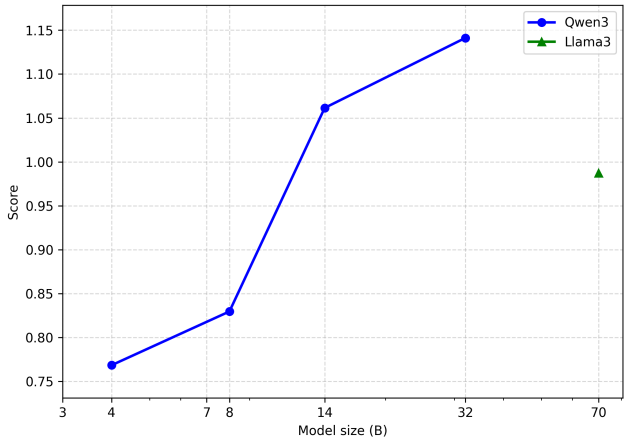

Figure 6: Line plot illustrating the relationship between CCS and model size for open-source models

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim that we have provided a new benchmark and playground for testing LLMs' spatial understanding of materials.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have mentioned the limitations that will be further solved in the discussion and conclusion parts.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The current paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the full benchmark code along with a detailed README. The experiments can be reproduced directly by running the provided scripts, without requiring manual parameter tuning.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the benchmark code and data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The benchmark details are included in the methodology, as well as the readme file inside the code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to skewed and task-dependent distributions, mean $\pm$ SD may be misleading. Instead, raw data and distribution figures are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments are mainly based on API calls to external LLM providers, but we do not provide detailed statistics on runtime or compute usage.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms fully with the NeurIPS Code of Ethics, and no deviations are present.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper mainly focuses on the usage of LLMs in materials science, which does not directly cause the societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We rely on the Materials Project database, pymatgen, etc., which are properly cited and used under their respective licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new benchmark implementation, including code, dataset, and prompt templates. These will be released in a GitHub repository.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The benchmark is designed to test LLMs, and we have also used LLMs for revising the manuscript and some of the coding. But we did not develop the core method with LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.