AtomWorld: A Benchmark for Evaluating Spatial Reasoning in Large Language Models on Crystalline Materials

Taoyuze Lv [1] **Alexander Chen** [2] **Fengyu Xie** [1] **Chu Wu** [1] **Jeffrey Meng** [2] **Dongzhan Zhou** [1]
Yingheng Wang [3] **Bram Hoex** [2] **Zhicheng Zhong** [1] **Tong Xie** [2]
[1]*Suzhou Institute for Advanced Research, University of Science and Technology of China* [2]*University of New South Wales*
[3]*Cornell University*. Correspondence to: Fengyu Xie , Tong Xie tong.xie@unsw.edu.au.

1. Introduction and Motivation

Large Language Models (LLMs) have demonstrated exceptional proficiency in textual reasoning, yet their application to complex, domain-specific tasks involving spatial understanding remains an open challenge. This capability is particularly critical in materials science, where a deep understanding of 3D atomic structures is fundamental. While previous studies have explored LLMs for crystal generation [1, 2, 3] and materials property prediction [4], there has been a notable absence of standardized benchmarks to systematically evaluate the core reasoning abilities required to manipulate these structures.

To bridge this gap, this paper introduces AtomWorld, a benchmark designed to evaluate LLMs on tasks based on Crystallographic Information Files (CIFs) [5], the standard format for storing structural data. The authors propose a theoretical framework dividing LLM capabilities into three stages: motor skills (mechanics of geometry), perceptual skills (pattern recognition, property prediction), and cognitive skills (new hypotheses, structure generation). The hypothesis driving this work is that for LLMs to succeed in "cognitive" material discovery, they must first master "motor skills" - the ability to add, move, rotate, or insert atoms consistently within a structure.

2. The AtomWorld Benchmark Design

The core of the study is AtomMotor-1K, a test set comprising 1,500 questions designed to benchmark reasoning LLMs on CIF motor skills.

2.1 Dataset Generation

The dataset is generated via the AtomWorld data generator, which creates pairs of "before" and "after" CIF states alongside natural language action prompts. The benchmark uses standard CIF representations from *pymatgen* [6] and the Materials Project (MP) [7] to minimise formatting uncertainty.

2.2 Action Types

The benchmark evaluates 10 distinct action types that simulate real-world structural modifications researchers perform, categorising them into:

- Point Defect and Doping: change, remove, add, insert_between, swap.
- Structure Perturbation: move, move_towards, rotate_around.
- Surface/Supercell Generation: delete_below, super_cell.

2.3 Complementary Tests

To isolate specific LLM weaknesses, the authors introduced several complementary evaluation modules:

- PointWorld: A simplified test using raw 3D coordinates (stripping away CIF syntax) to measure inherent geometric difficulty.
- CIF-Repair: Tests the model's robustness by asking it to fix corrupted or incomplete CIF files.
- CIF-Gen: Evaluates the generation of syntactically valid CIFs for prototype crystals (e.g., perovskite, diamond).
- Chemical Competence Score (CCS) [8]: Assesses latent chemical knowledge by evaluating the model's ability to distinguish accurate from inaccurate crystal descriptions.

3. Experimental Setup and Models

The study evaluated a range of frontier models, including Gemini 2.5 Pro, GPT-o3, GPT-4-mini, DeepSeek Chat, Llama-3 70B, and the Qwen-3 series (ranging from 4B to 32B parameters). Performance was measured primarily by Success Rate (correct format, and within maximum site tolerance) and Mean Maximum Distance (max_dist) (maximum pairwise atomic displacement).

4. Key Results and Analysis

The main results of AtomMotor-1K, alongside complementary tests are presented in Figure 1.

4.1 Performance Hierarchy

The results reveal some separation of task difficulty depending on number of atoms to be operated on. LLMs performed well on single-atom tasks - which could be categorised as "easy"; such as change, remove, and add, but struggled significantly with "moderate" (involving 2 atoms) tasks like move and insert_between, and failed consistently at "hard" tasks like rotate_around which involved 3+ atoms. While basic operations had high success rates, the rotate_around action proved highly challenging, with models often failing to apply rotation matrices consistently.

4.2 Scaling and Architecture

Parameter scaling within the Qwen-3 series showed that larger models achieved higher success rates and smaller spatial displacements. However, the Qwen3-32B model did outperform the larger Llama3-70B across most tasks.

4.3 PointWorld vs. CIF competency tests

In the simplified PointWorld tests, models like DeepSeek V3 achieved near-perfect success rates on "moderate" difficulty tasks (move, insert_between), suggesting that the difficulty in AtomWorld stems from the complexity of CIF syntax rather than pure calculation. Yet the CIF competency tests (CIF-Repair, CIF-Gen, CSS) demonstrated that models (particularly Gemini 2.5 Pro, GPT-o3, GPT-4-mini) were familiar with CIF syntax. Because models demonstrate high success rates when tested for spatial reasoning and CIF syntax following in isolation, the authors found the difficulty of AtomWorld tasks comes from when both requirements are combined.

4.4 Memorization vs. Understanding

In the CIF-Gen tasks, models generated standard prototype chemical compositions (e.g., NaCl) more accurately than non-standard compounds with the same crystal structure (e.g., MgSe). This asymmetry suggests that current LLMs rely heavily on memorizing specific training examples rather than understanding underlying structural principles.

4.5 Tool-Augmented LLMs

The authors explored a tool-augmented framework using Retrieval-Augmented Generation (RAG) over the pymatgen library.

- Impact: Providing models with tools significantly improved performance. For instance, DeepSeek-chat's success rate on insert_between jumped from 45.6% to 83.0% when equipped with tools.

- Limitations: Even with tools, complex spatial reasoning remained an obstacle. The rotate_around task only saw an improvement to 18% success, indicating that code tools alone are insufficient without better inherent spatial reasoning or task-specific fine-tuning.

5. Conclusion

AtomWorld establishes that while current LLMs possess promising baselines, they consistently fail in robust structural understanding and spatial reasoning required for crystallography. The models tend to approach geometric tasks algorithmically, succeeding at simple arithmetic operations but failing at complex spatial transformations like rotation.

The paper concludes that mastering these "motor skills" is a prerequisite for high-value "cognitive" tasks in material discovery, making it vital to benchmark LLM progress on this capability. Future progress will likely depend on developments in tool-augmented design and multimodal reasoning to bridge the gap between textual proficiency and 3D spatial intelligence.

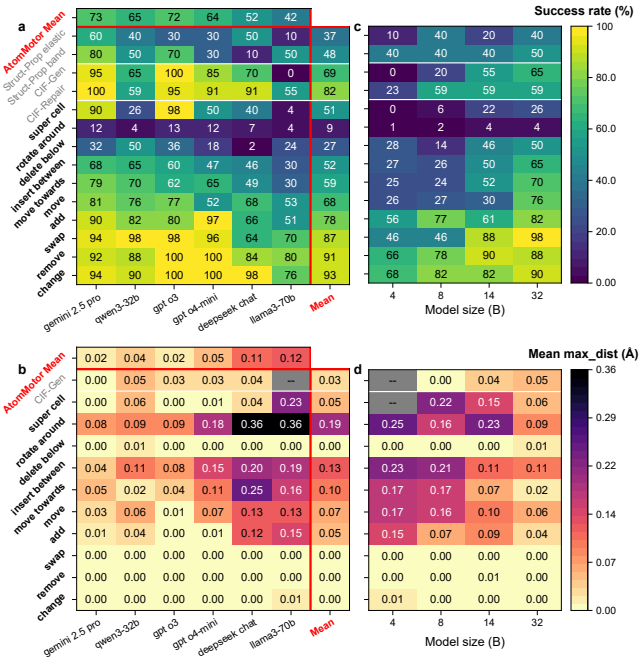

Fig. 1: **a.** Success rate metric across AtomMotor-1K, CIF-Repair, CIF-Gen and StructProp datasets. **b.** Mean `max_dist` metric across AtomMotor-1K and CIF-Gen datasets. **c**, **d.** Parameter scaling results on Qwen3 series. The right side are some randomly sampled structures from the tested data.

References

[1] 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.

[2] Yan Chen, Xueru Wang, Xiaobin Deng, Yilun Liu, Xi Chen, Yunwei Zhang, Lei Wang, and Hang Xiao. MatterGPT: A Generative Transformer for Multi-Property Inverse Design of Solid-State Materials, August 2024. arXiv:2408.07608 [cond-mat].

[3] Kamal Choudhary. AtomGPT: Atomistic Generative Pretrained Transformer for Forward and Inverse Materials Design. *The Journal of Physical Chemistry Letters*, 15(27):6909–6917, July 2024.

[4] 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. Publisher: IOP Publishing.

[5] 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. tex.eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1107/S010876739101067X.

[6] 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.

[7] 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.

[8] 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.