MULTIPHASEWORLD: A MULTIMODAL BENCHMARK FOR STRUCTURAL EVOLUTION IN MULTIPHASE SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

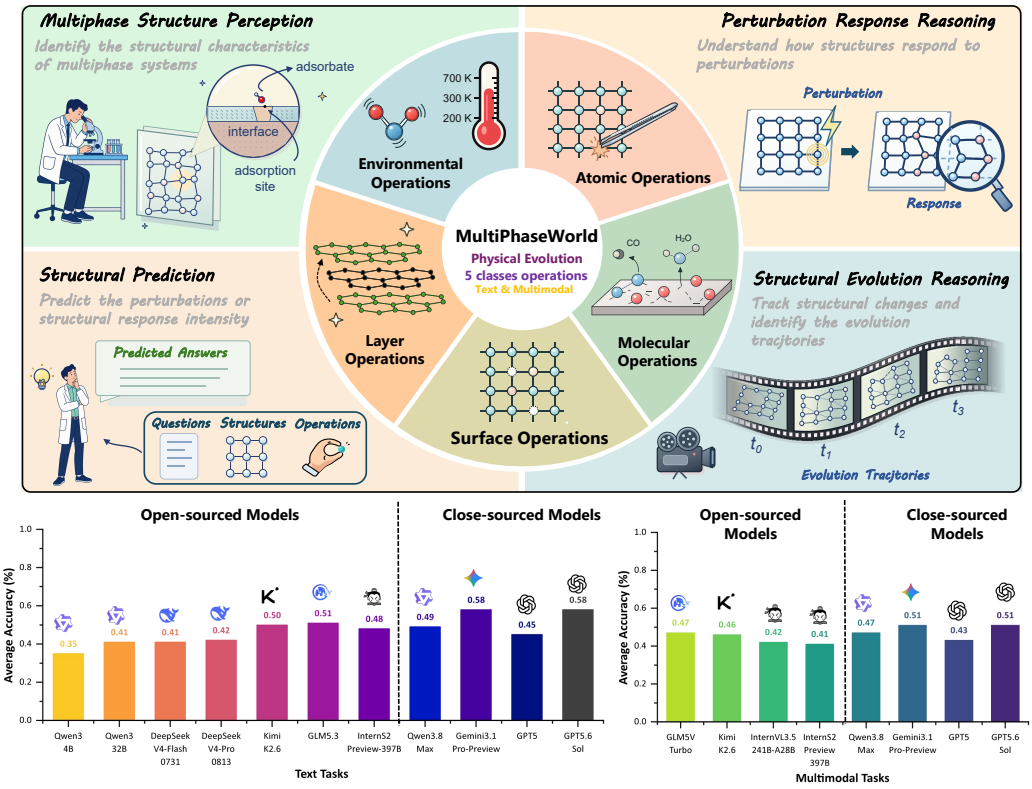
Large language models (LLMs) and multimodal large language models (MLLMs) are increasingly applied to scientific problems. However, existing scientific benchmarks primarily evaluate factual knowledge and static structure understanding, leaving evolution reasoning in diverse chemical environments underexplored. We introduce **MultiPhaseWorld**, a multimodal benchmark for evaluating scientific reasoning over structural evolution in multiphase systems. Multiphase systems involve complex interactions across phases and diverse chemical environments, making structural evolution challenging to understand. At the same time, this complexity provides a rich setting for evaluating the reasoning capabilities of LLMs and MLLMs across multiple scales, including adsorbate, surface, interface, layer, and bulk environment. MultiPhaseWorld contains 11,075 text and multimodal questions across 28 tasks and 14 capability dimensions, organized into four progressive levels: *Multiphase Structure Perception*, *Perturbation Response Reasoning*, *Structural Evolution Reasoning*, and *Structural Prediction*. The benchmark is built from 336 solid-solid and 400 solid-gas systems. Controlled structural perturbations, structural optimization, and molecular dynamics simulations are used to generate structural responses, evolved structures, and evolution trajectories. The strongest models achieve only 58% average accuracy on text tasks and 51% on multimodal tasks, remaining more than 20 percentage points below the expert performance. Fine-grained results further reveal limitations in integrating physicochemical knowledge, reasoning about evolution relationships, and resolving subtle structural changes. Overall, current models show partial understanding of chemical information, but remain limited in linking explicit perturbations to structural responses, tracking state evolution, and predicting perturbation effects.

1 INTRODUCTION

Large language models (LLMs) and multimodal large language models (MLLMs) have shown rapidly growing capabilities in scientific domains, including chemistry and materials science (OpenAI, 2026; Google DeepMind, 2026; Qwen Team, 2026; ScienceOne AI & Wenge AI, 2026; Bai et al., 2026). These models are increasingly used to interpret molecules, crystals, interfaces, and other structures (Birhane et al., 2023; Alampara et al., 2026; Liang et al., 2026), creating a growing need for systematic benchmarks that characterize their scientific capabilities.

Scientific benchmarks have expanded from factual knowledge and question answering to multimodal scientific understanding and structural reasoning (Cui et al., 2025; Wang et al., 2026; Rubungo et al., 2025; Zhang et al., 2026a; Alampara et al., 2025a; Liu et al., 2025b). Given the importance of structural evolution for understanding material behavior (Fairhurst et al., 2025; Hollingsworth & Dror, 2018), increasing attention has recently been directed toward structural perturbations and dynamic evolution. AtomWorld (Lv et al., 2026) evaluates LLMs' structural manipulation capabilities through atomic-level operations, but does not consider the structural responses induced by these perturbations. EvoMD-LLM (Tang et al., 2026b) evaluates temporal prediction over reactive molecular trajectories, but does not explicitly model the effects of structural perturbations on the ensuing evolution. As a result, existing benchmarks do not fully capture reasoning from perturbations to structural evolution, leaving a critical gap in the evaluation of **dynamic scientific reasoning**.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Benchmark	Capability	Multi.	Traj.	Perturb.	Evol.
SciEvalKit (Wang et al., 2026)	Science QA	×	×	×	×
MaCBench (Alampara et al., 2025a)	Science Interpretation	✓	×	×	×
MatSciBench (Zhang et al., 2026a)	Materials Reasoning	✓	×	×	×
MatTools (Liu et al., 2025b)	Tool Use	✓	×	×	×
LLM4Mat-Bench (Rubungo et al., 2025)	Structure-Property Reasoning	×	×	×	×
MatterChat (Tang et al., 2026a)	Materials Understanding	✓	×	×	×
EvoMD-LLM (Tang et al., 2026b)	Molecular Evolution	×	✓	×	✓
AtomWorld (Lv et al., 2026)	Structure Editing	✓	×	✓	×
Ours	**Structural Evolution Reasoning**	✓	✓	✓	✓

Figure 1: Overview and comparison of ***MultiPhaseWorld***. Top: MultiPhaseWorld overview. Middle: model performance on text-based and multimodal tasks. Bottom: comparison with recent scientific LLM benchmarks.

Addressing this gap requires capabilities beyond structural perception and basic chemical knowledge, including *linking perturbations to structural responses*, *tracking state evolution*, and *predicting perturbation effects*. Perturbations of varying scales and types induce diverse structural responses, generating a broad range of dynamic trajectories across different chemical environments. Under settings where observations are restricted to 2D images, the task becomes a partially observable multimodal reasoning problem, requiring models to infer structural changes and evolution from limited visual evidence (Alampara et al., 2025b; Lai et al., 2025). These capabilities could advance LLMs and MLLMs toward deeper scientific reasoning, enabling them to understand perturbations, track evolution, and predict structural responses, while supporting experimental design (Mandal et al., 2025), characterization and computational analysis (Du et al., 2026; Huang et al., 2026), and materials design (Liu et al., 2025a) toward more autonomous end–to–end scientific research (Kusne & McDannald, 2026; Alfred & Sundararaghavan, 2026).

To evaluate these capabilities, we introduce **MultiPhaseWorld**, a simulation–grounded benchmark for general LLMs and MLLMs. We accumulated **10,298** structural evolution trajectories and ultimately constructed **11,075** questions organized into four capability levels. The benchmark is built from 336 solid-solid and 400 solid-gas systems, which encompass surface, interfacial, and bulk-like chemical environments. As shown at the top of Figure 1, it targets **four** capabilities in multiphase systems: *Multiphase Structure Perception*, *Perturbation Response Reasoning*, *Structural Evolution Reasoning*, and *Structural Prediction*. The structural evolution trajectories are generated by applying molecular dynamics simulations (MD) and structural optimization to the configurations under **five** classes of perturbations: *atomic*, *molecular*, *surface*, *layer*, and *environmental*.

Our benchmark evaluates 13 open-weight and proprietary LLMs and MLLMs. As shown in the middle panel of Fig. 1, current models leave substantial room for improvement: the highest average accuracy reaches only 58 % on text tasks and 51 % on multimodal tasks. Our results reveal persistent limitations in current LLMs when addressing scientific problems about structural evolution, particularly in integrating and applying physicochemical knowledge, reasoning about evolution relationships, and resolving subtle structural changes. Several models evaluated in both settings show accuracy drops of 3–7 percentage points on multimodal tasks, consistent with difficulty in extracting and grounding sufficient structural information from limited visual evidence. In the model families examined, parameter scaling yields modest gains, while newer model generations show larger improvements. Nevertheless, the strongest models still lag behind the expert baseline by more than 20 percentage points. Overall, current LLMs and MLLMs remain limited in robust reasoning over structural evolution, particularly in physicochemical integration, evolution inference, and visual-geometric grounding.

2 Related Work

Scientific LLMs and MLLMs LLMs and MLLMs are increasingly being adapted for scientific applications, where they support tasks beyond conventional question answering and knowledge retrieval (Ramos et al., 2025; Fang et al., 2024; M. Bran et al., 2024; Jablonka et al., 2024). Chemistry has become an important application area for scientific LLMs. ChemMiner (Chen et al., 2025) focuses on chemical literature understanding and information extraction, while ChemCrow (M. Bran et al., 2024) and Wang et al. (2025a) enable reaction reasoning and synthesis planning. Mol-Instructions (Fang et al., 2024) and LLaMo (Park et al., 2024) further advance molecular understanding, generation, and property prediction. Their capabilities have expanded beyond text reasoning toward multimodal scientific understanding over diverse representations. Recent models such as Intern-S2 (Bai et al., 2026) and S1-Omni (ScienceOne AI & Wenge AI, 2026) aim to unify heterogeneous scientific modalities for broader scientific understanding and reasoning.

Scientific Benchmarks Recent scientific benchmarks have expanded from knowledge evaluation to more diverse scientific capabilities. SciEvalKit (Wang et al., 2026) focuses on scientific question answering, while MaCBench (Alampara et al., 2025a), MatSciBench (Zhang et al., 2026a), and MatTools (Liu et al., 2025b) extend evaluation toward multimodal scientific understanding, materials reasoning, and tool use. LLM4Mat-Bench (Rubungo et al., 2025) and MatterChat (Tang et al., 2026a) further explore structure–property understanding and analysis. AtomWorld (Lv et al., 2026) introduces atomic structure editing, and EvoMD-LLM (Tang et al., 2026b) incorporates molecular trajectory reasoning. The key differences among these benchmarks are summarized at the bottom of Figure 1. However, existing benchmarks leave perturbation-induced structural responses and subsequent structural evolution largely unexplored, motivating **MultiPhaseWorld** for multiphase systems.

3 MultiPhaseWorld

MultiPhaseWorld is a benchmark for evaluating LLMs and MLLMs on understanding and reasoning about dynamic structural behaviors across diverse chemical environments in multiphase systems. To define a tractable scope, we use heterostructures to represent solid-solid systems and surface-adsorbate structures as representative solid-gas systems. As shown in Figure 2, the benchmark covers four capability levels and 14 capability dimensions. The benchmark contains 11,075 questions,

including 6,824 text-based and 4,251 multimodal instances, enabling a comprehensive evaluation of LLMs' understanding of dynamic behaviors.

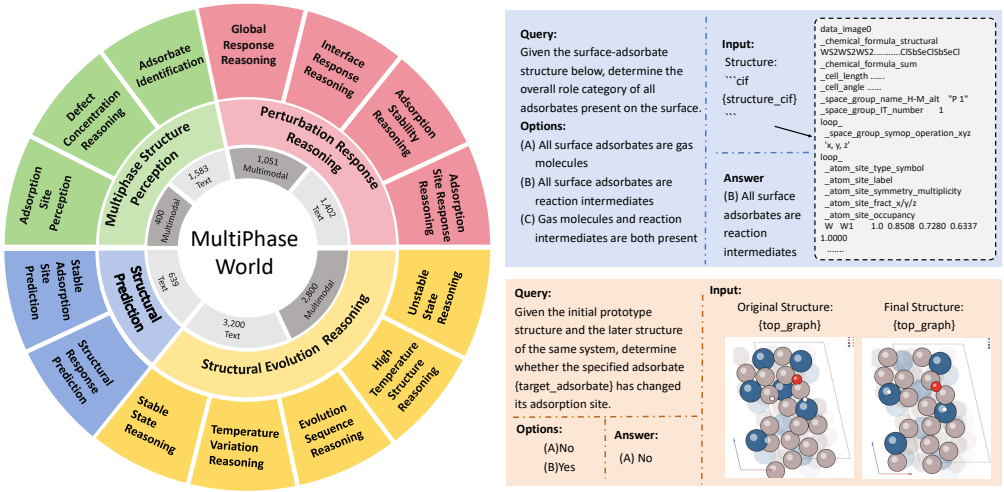


Figure 2: Capability taxonomy and representative evaluation examples in ***MultiPhaseWorld***. Left: 14 capability dimensions grouped into four progressive levels, from multiphase structure perception to structural prediction. Right: representative text and multimodal evaluation examples based on CIF structures and visual structural comparisons.

3.1 Task Taxonomy

The tasks are organized into four categories: *multiphase structure perception*, *perturbation response reasoning*, *structural evolution reasoning*, and *structural prediction*. Structural Prediction evaluates whether models can infer stable structural preferences or identify perturbation effects. In total, the benchmark contains 28 task types covering 14 capability dimensions, with the complete task list provided in the Appendix A.

Multiphase Structure Perception Basic structural perception is a prerequisite for LLMs and MLLMs to reason about subsequent structural responses and evolution. This level evaluates whether models can recognize the basic structural characteristics in different chemical environments and relate them to chemical knowledge, including four tasks spanning three capability dimensions: *recognizing adsorbate roles*, *defect concentration reasoning*, and *identifying local adsorption configurations*.

Perturbation Response Reasoning Before reasoning about dynamic structural evolution, models must first link explicit perturbations to their structural responses. At this level, models must compare structures before and after perturbation and identify their relationship. For solid-solid systems, four tasks evaluate reasoning about *global* and *interface structural reconstruction*. In solid-gas systems, three tasks examine *adsorption site changes reasoning* and *adsorption stability reasoning*.

Structural Evolution Reasoning This level extends evaluation to the ***evolution process***. Models are given multiple states sampled from evolution trajectories after perturbation and required to reason about structural evolution across steps. There are 15 tasks, including *identifying the most or least stable state*, *ordering shuffled frames from evolution trajectories*, and *high-temperature structure reasoning*. Overall, this level evaluates whether models can integrate physicochemical knowledge to track the structural changes across trajectory steps induced by perturbations.

Structural Prediction The final level evaluates whether models can make structural predictions by integrating structural geometry, chemical environments, and more physicochemical knowledge. We design two tasks: for solid-solid systems, models are asked to *predict the perturbation that*

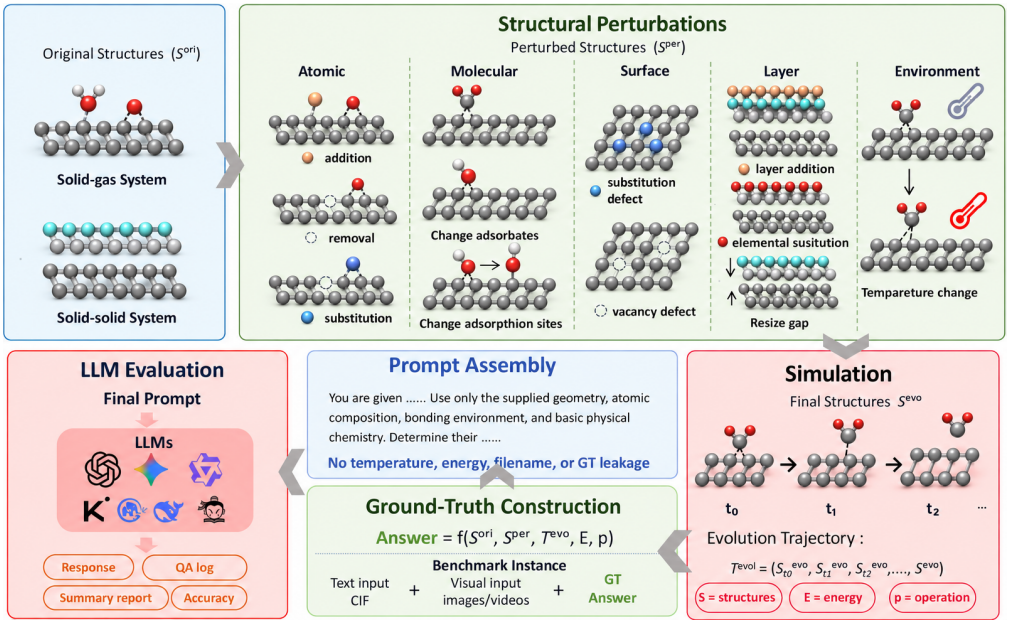


Figure 3: Overview of the MultiPhaseWorld benchmark construction and evaluation pipeline. Original structures S^{ori} are transformed through atomic, molecular, surface, layer, and environmental perturbations to obtain perturbed structures S^{per}. Physical simulations then are employed to obtain evolved structures S^{evo} and evolution trajectories T^{evol}. The resulting structural information is summarized for constructing ground truth answers. Finally, text and visual inputs are assembled into prompts for evaluating LLMs across the benchmark tasks.

induces the strongest reconstruction, testing their capabilities in interfacial environments; for solid-gas systems, models are asked to *predict the preferred adsorption site of the adsorbate*, testing their capabilities in surface environments. This level requires forward reasoning toward an unseen evolved state.

3.2 Benchmark Generation Pipeline

The overall benchmark pipeline is illustrated in Figure 3. We first collect original structures S^{ori} and apply structural operations to obtain perturbed structures S^{per}. These structures then undergo physical simulation, yielding final structures S^{evo} and evolution trajectories T^{evol}. We construct the ground truth answers and assemble benchmark prompts with CIF-based text or visual inputs. The resulting prompts are finally used to evaluate LLMs and MLLMs across the four capability levels.

Data Collection MultiPhaseWorld considers two representative classes of multiphase systems as the original structures S^{ori}: solid-solid heterostructures and solid-gas catalyst–adsorbate systems (Zhang et al., 2026b; Ahmad et al., 2026; Fernández-Vidal & Koper, 2026; Ye et al., 2026). We use all 336 heterostructures from HetDB (Sauer et al., 2025) and sample 400 surface adsorbate structures from OC20 (Chanussot et al., 2021) based on element and adsorbate distributions (Appendix B).

Structural Perturbations and Simulation In this section, we transform original structures S^{ori} into perturbed structures S^{per} and compute their evolved structures S^{evo} and trajectories T^{evol}. As summarized in the table in Table 1, perturbations are grouped into five categories: *atomic*, *molecular*, *surface*, *layer*, and *environmental* operations. Each HetDB structure undergoes six structural perturbations: *atom addition*, *removal*, and *substitution*, together with *monolayer addition*, *elemental substitution*, and *vacuum resizing*. Each OC20 structure is assigned 21 structural perturbations, comprising *three atomic*, *nine molecular*, *three environmental*, and *six surface operations* detailed in Figure 5. Detailed implementation procedures for these operations are provided in Appendix C.

Table 1: Structural operations at different levels.

Level	Operations
Atomic	Atom addition, removal, and substitution
Molecular	Gas adsorbate substitution (CO_2, H_2O, CO) Intermediate adsorbate substitution (O, OH, OOH) Adsorption-site relocation (top, bridge, hollow)
Surface	Surface vacancy defects (5%, 10%, 15%) Surface substitutional defects (5%, 10%, 15%)
Layer	Layer element replacement, layer addition, and vacuum resizing
Environmental	Temperature variation (200, 300, 700 K)

Finally, we collect a total of **10,298** evolution trajectories. Our simulations are based on structural optimization and MD driven by MACE-MP-0, which has been reported to perform well on surface and multiphase systems (Batatia et al., 2025). These simulations are intended to capture the trends of evolution rather than to provide quantitatively exact properties. The simulation settings are provided in Appendix D.

Ground Truth Construction and Evaluation All reference answers in MultiPhaseWorld are constructed automatically from dataset annotations, perturbation metadata, evolved structures, energies, and simulation trajectories, without LLM involvement. Static structural labels are derived from composition and geometry, while reconstruction and adsorption site references are determined from structural changes and local coordination. Stability, trajectory, and prediction tasks are grounded in simulated energies, evolution indices, and post-evolution structures. Invalid operations, failed simulations, and ambiguous records are excluded to ensure reliable and physically grounded references. Detailed generation procedures and validation criteria are described in the Appendix E. The answer builders then combine individual or multiple structural states into text and multimodal questions according to the task templates. After task-eligibility filtering and stratified sampling, the final benchmark contains ***11,075*** questions (Appendix A.1).

Quality Control We apply a two-stage quality-control procedure before question construction. First, geometrically invalid or physically implausible perturbed structures S^{per} are removed based on geometric and physical validity criteria (Appendix D). Second, non-converged or inconsistent simulation trajectories are excluded. Only samples passing both stages are used for question construction. To ensure fair comparison across models, all benchmark questions follow standardized prompt templates with consistent task descriptions, input formats, and answer constraints. For paired text and multimodal questions, the task formulation and answer space remain identical, differing only in structural input: CIF files or rendered images.

Statistics and Metrics MultiPhaseWorld contains 11,075 questions, comprising 6,824 text-based and 4,251 multimodal tasks across 28 task types and 14 capability dimensions in four levels: *Multiphase Structure Perception*, *Perturbation Response Reasoning*, *Structural Evolution Reasoning*, and *Structural Prediction*. Table 2 summarizes the dataset statistics. We evaluate predictions using task accuracy, gain over random baseline (Δ*Rand.*), and the number of task wins (N_{win}). We report both aggregate and fine-grained results across task levels. Detailed metric definitions are provided in Appendix F.

Table 2: Statistics of MultiPhaseWorld.

Statistic	Number
Structures	HetDB (336); OC20 (400)
Questions	11,075 (62% text; 38% multimodal)
Levels	4
Dimensions	14
Tasks	28
Level 1	1,983 (18%)
Level 2	2,453 (22%)
Level 3	6,000 (54%)
Level 4	639 (6%)

Table 3: Text accuracy across different reasoning levels. Δ*Rand.* denotes the gain of the overall average accuracy over the random baseline. N_{win} denotes the number of tasks on which a model achieves the highest accuracy; tied best results are counted as wins for all tied models. The best result in each column is highlighted in bold.

Model	Level-1 Structure Perception	Level-2 Perturbation Response	Level-3 Structural Evolution	Level-4 Structural Prediction	Avg.	Δ*Rand.*	N_{win}
Random	0.31	0.44	0.28	0.29	0.32	–	–
Expert	–	–	–	–	0.80	–	–
Open-sourced Models							
Qwen3-32B	0.36	0.52	0.40	0.26	0.41	0.09	0
Qwen3-4B	0.35	0.45	0.33	0.24	0.35	0.03	1
DeepSeek-V4-Flash-0731	0.40	0.51	0.39	0.32	0.41	0.09	1
DeepSeek-V4-Pro-0813	0.40	0.50	0.41	0.37	0.42	0.10	0
GLM-5.3	0.66	0.54	0.44	0.52	0.51	0.19	0
Intern-S2-Preview-397B	0.42	0.54	0.46	0.54	0.48	0.16	1
Kimi-K2.6	0.48	0.56	0.47	0.49	0.50	0.18	2
Close-sourced Models							
GPT-5	0.45	0.54	0.43	0.34	0.45	0.13	1
Gemini-3.1-Pro-Preview	**0.72**	0.58	**0.52**	**0.63**	**0.58**	**0.26**	5
Qwen3.8-Max	0.47	0.57	0.43	0.61	0.49	0.17	1
GPT-5.6-Sol	0.64	**0.62**	**0.52**	0.62	**0.58**	**0.26**	**7**

Table 4: Multimodal accuracy across different reasoning levels. Δ*Rand.* denotes the gain of the overall average accuracy over the random baseline. N_{win} denotes the number of tasks on which a model achieves the highest accuracy; tied best results are counted as wins for all tied models. The best result in each column is highlighted in bold.

Model	Level-1 Structure Perception	Level-2 Perturbation Response	Level-3 Structural Evolution	Avg.	Δ*Rand.*	N_{win}
Random	0.25	0.44	0.27	0.32	–	–
Expert	–	–	–	0.73	–	–
Open-sourced Models						
InternVL3.5-241B-A28B	0.48	**0.60**	0.33	0.42	0.10	0
Intern-S2-Preview-397B	0.34	0.56	0.35	0.41	0.09	0
Kimi-K2.6	0.39	0.59	0.41	0.46	0.14	2
Close-sourced Models						
GPT-5	0.47	0.55	0.37	0.43	0.11	1
Gemini-3.1-Pro-Preview	**0.55**	0.58	**0.48**	**0.51**	**0.19**	**5**
Qwen3.8-Max	0.53	0.57	0.42	0.47	0.15	2
GPT-5.6-Sol	0.53	**0.60**	0.47	**0.51**	**0.19**	1
GLM-5V-Turbo	0.45	0.59	0.42	0.47	0.15	3

4 EXPERIMENTS

We conduct a comprehensive evaluation of advanced LLMs and MLLMs on MultiPhaseWorld. For open-weight models, we evaluate Qwen3-4B/32B (Yang et al., 2025), DeepSeek-V4-Flash and DeepSeek-V4-Pro-0813 (DeepSeek-AI, 2026), GLM-5.3 (Z.AI, 2026), InternVL3.5-241B-A28B (Wang et al., 2025b), Intern-S2-Preview-397B (Bai et al., 2026), and Kimi-K2.6 (Moonshot AI, 2026). For proprietary models, we evaluate Qwen3.8-Max (Qwen Team, 2026), GPT-5 (OpenAI, 2025), Gemini-3.1-Pro-Preview (Google DeepMind, 2026), GLM-5V-Turbo (GLM-V Team et al., 2026), and GPT-5.6 Sol (OpenAI, 2026). All models are evaluated using a unified zero-shot QA pipeline. We set the temperature to 0 to reduce randomness, except for Kimi-K2.6 with a temperature of 0.6. All models used the lowest reasoning setting or reasoning disabled. Models are required to return only the final multiple-choice answer in a structured format. The expert evaluation was conducted by researchers with expertise in computational chemistry, standard simulation methods, and the interpretation of characterization data (Appendix A.2).

4.1 MAIN RESULTS

The evaluation results are summarized in Tables 3 and 4. Overall, current LLMs and MLLMs show limited performance on **MultiPhaseWorld**. Even the strongest models achieve only 58% average accuracy in the text tasks and 51% in the multimodal tasks, more than 20 percentage points below the expert baseline. Intern-S2-Preview also shows no clear advantage, suggesting that scientific specialization alone is insufficient for this benchmark. Across capability levels, performance drops substantially on *Structural Evolution Reasoning*. This indicates that stronger performance on static reasoning does not consistently transfer to structural evolution reasoning, highlighting that the two require distinct capabilities and that current models still face significant challenges in reasoning about structural evolution.

To further characterize these limitations, the fine-grained results in Appendix F reveal complementary strengths across models. GPT-5.6 Sol attains $N_{\text{win}} = 8$ (7 text and 1 multimodal), while Gemini-3.1-Pro-Preview achieves $N_{\text{win}} = 10$ (5 text and 5 multimodal). GPT-5.6 Sol performs better on fine-grained structural reasoning (e.g., T4, T5, T6, MT4), whereas Gemini-3.1-Pro-Preview shows stronger performance on knowledge-driven structure perception (T1 and T3) and dynamic reasoning tasks (e.g., T10, T11, MT5). These results suggest that Gemini-3.1-Pro-Preview might show stronger capabilities in dynamic reasoning, structural perception, and multimodal understanding, while GPT-5.6 Sol performs better on reasoning over structural details in text-based tasks. We therefore analyze representative errors made by GPT-5.6 Sol, which achieves high accuracy on Level-2 and Level-3 tasks. These errors reveal limitations in understanding structural relationships and dynamic evolution (Appendix G). Overall, these fine-grained results further reveal several limitations across models, particularly in *integrating scientific criteria*, *dynamic reasoning*, and *detecting subtle structural changes*.

Insufficient Integration of Scientific Criteria Across both text and multimodal settings, models show a clear gap between stable and unstable state reasoning. Stable state reasoning remains near or below the random baseline, with average accuracy across all evaluated models of 39% (T9, solid-gas) and 20% (T8, solid-solid) in text tasks, and 35% (MT6, solid-gas) and 26% (MT5, solid-solid) in multimodal tasks. In contrast, unstable state reasoning reaches 72% (T11, solid-gas) and 23% (T10, solid-solid) in text, and 54% (MT8, solid-gas) and 37% (MT7, solid-solid) in multimodal tasks. When assessing stability, obvious local abnormalities can directly indicate an unstable structure, whereas identifying the more stable structure requires considering multiple factors. This makes stable state reasoning more demanding. The results therefore indicate a limited ability to integrate scientific knowledge for reasoning about structural stability across evolution frames.

Difficulty Detecting Subtle Structural Changes As mentioned above, solid-gas systems consistently outperform solid-solid systems in both stable state reasoning (T9 vs. T8; MT6 vs. MT5) and unstable state reasoning (T11 vs. T10; MT8 vs. MT7). This pronounced difference may arise because adsorbates are more spatially distinct from the host structure, providing localized and clearer structural cues than the distributed changes across solid-solid interfaces. Meanwhile, models struggle with both adsorption-site change reasoning and adsorption-site identification, achieving average accuracies of only 52% on T6 and 53% on MT4 for the former, and 37% on T2 and 47% on MT1 for the latter. These results together suggest that current models remain limited in recognizing subtle structural changes.

Gaps in Dynamic Reasoning Models perform substantially better at identifying individual structural states than at reasoning about their evolution order. For solid-gas systems, average accuracy drops from 72% on unstable-state identification (T11) to 47% on trajectory ordering (T13) in text tasks, and from 54% (MT8) to 42% (MT10) in multimodal tasks. Thus, recognizing individual states does not guarantee reliable inference of their evolution relationships.

4.2 ANALYSIS

Scaling Effect Scaling Qwen3 from 4B to 32B parameters produces consistent but moderate improvements across the four capability levels (Figure 4a). The largest gains occur at Levels 2 and 3, while improvements at Level 4 remain limited. This suggests that scaling improves structural comparison and stability reasoning but does not substantially enhance prediction of atomistic evolution.

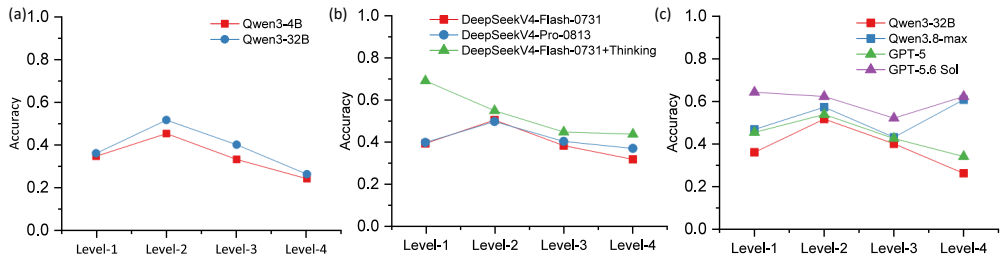


Figure 4: **Effects of model scale, inference configuration, and model generation.** Mean accuracy across four task levels for (a) Qwen3-4B vs. Qwen3-32B, (b) DeepSeek-V4-Flash, DeepSeek-V4-Pro-0813, and DeepSeek-V4-Flash with thinking, and (c) Qwen and GPT model generations.

Model Variants and Inference Settings Effect As shown in Figure 4b, DeepSeek-V4-Pro-0813 provides only limited improvements over DeepSeek-V4-Flash. For DeepSeek-V4-Flash, high reasoning effort boosts performance on simple Level 1 multiphase perception questions, from 0.40 to 0.69, but yields only marginal gains on the harder Level 2–4 tasks. Although inference settings differ across models, with Gemini-3.1-Pro-Preview and GPT-5.6 Sol using low-effort reasoning, all models still exhibit clear limitations. These results suggest that stronger model variants and chain-of-thought reasoning offer only partial benefits and remain insufficient for solving the dynamic reasoning challenges of MultiPhaseWorld.

Generational Effect Across the model families examined (Figure 4c), generational upgrades produce larger performance gains than parameter scaling alone. GPT-5.6 Sol consistently outperforms GPT-5 across all four levels, with improvements spanning structural perception, perturbation response, structural evolution, and prediction. Qwen3.8-Max shows a particularly pronounced gain at Level-4, suggesting that newer model generations can yield stronger improvements on more demanding predictive reasoning tasks. Overall, these results indicate that generational advances provide broader benefits than simply increasing model size within the same family.

5 Limitations and Future Work

MultiPhaseWorld currently focuses on representative solid-solid and solid-gas settings rather than the full diversity of multiphase systems. Its simulation-derived references are designed to capture structural responses and evolutionary trends, but are constrained by finite simulation horizons due to computational cost, which may affect quantitative estimates and limit the observation of slow or rare structural transitions. Future work will broaden the range of multiphase systems, accumulate higher-fidelity and longer-timescale simulation data, and further support the development of LLMs/MLLMs for structural perception and scientific reasoning in dynamic multiphase worlds.

6 Conclusion

We introduce **MultiPhaseWorld**, a benchmark for evaluating dynamic scientific reasoning in LLMs and MLLMs through multiphase structural perturbations and evolution. It tests whether models can connect perturbations to structural responses, reason over evolution trajectories, and predict structural states. For MLLMs, we further evaluate these capabilities under limited visual observations.

Our results show that current models remain limited in dynamic reasoning, particularly in integrating physicochemical criteria, inferring evolution relationships, and detecting subtle structural changes. Multimodal inputs do not consistently improve performance, indicating limitations in extracting and grounding structural information from visual observations. Overall, strong static understanding does not directly translate into reliable dynamic reasoning, highlighting the need for models that can better connect perturbation, response, evolution, and prediction.

AI USE STATEMENT

We used OpenAI Codex to assist with experimental and evaluation code, including the implementation of ground-truth construction. The authors designed the ground-truth generation logic; reference answers were computed programmatically from dataset records and simulation outputs, rather than generated through LLM inference. We used ChatGPT to assist with language polishing and editing of the manuscript. The authors reviewed all AI-assisted code and manuscript content and take responsibility for the final text, claims, code, and artifacts presented in this work.

REFERENCES

Waqas Ahmad, Jamal Kazmi, Muhammad Zubair Nawaz, Amine El Moutaouakil, Jinying Zhang, Yury Illarionov, and Zhiming Wang. Interface engineering in van der waals heterostructures: enhancing photodetector efficiency through structural and functional modifications. *Advanced Functional Materials*, 36(24):e16893, 2026.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 5(10):952–961, Oct 2025a. ISSN 2662-8457. doi: 10.1038/s43588-025-00836-3. URL https://doi.org/10.1038/s43588-025-00836-3.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 5(10):952–961, Oct 2025b. ISSN 2662-8457. doi: 10.1038/s43588-025-00836-3. URL https://doi.org/10.1038/s43588-025-00836-3.

Nawaf Alampara, Anagha Aneesh, Martiño Ríos-García, Adrian Mirza, Mara Schilling-Wilhelmi, Ali Asghar Aghajani, Meiling Sun, Gordan Prastalo, and Kevin Maik Jablonka. General-purpose models for the chemical sciences: Llms and beyond. *Chemical Reviews*, 126(4):2484–2549, 2026. doi: 10.1021/acs.chemrev.5c00583.

Samuel Onimpa Alfred and Veera Sundararaghavan. From data to theory: Autonomous large language model agents for materials science. *Computational Materials Science*, 272:114862, 2026. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2026.114862. URL https://www.sciencedirect.com/science/article/pii/S0927025626003812.

Lei Bai, Jiaqi Cao, Chiyu Chen, Guanzhou Chen, Kai Chen, Guangran Cheng, Erfei Cui, Xuanlang Dai, Shengyuan Ding, Shangheng Du, et al. Intern-s2-preview: Scientific agentic foundation model. *arXiv preprint arXiv:2608.13505*, 2026. URL https://arxiv.org/abs/2608.13505.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf,

Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, 11 2025. ISSN 0021-9606. doi: 10.1063/5.0297006. URL https://doi.org/10.1063/5.0297006.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280, May 2023. ISSN 2522-5820. doi: 10.1038/s42254-023-00581-4. URL https://doi.org/10.1038/s42254-023-00581-4.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 05 2021. ISSN 2155-5435. doi: 10.1021/acscatal.0c04525. URL https://doi.org/10.1021/acscatal.0c04525.

Kexin Chen, Yuyang Du, Junyou Li, Hanqun Cao, Menghao Guo, Xilin Dang, Lanqing Li, Jiezhong Qiu, Guangyong Chen, and Pheng Ann Heng. Chemminer: A large language model agent system for chemical literature data mining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 7654–7662, October 2025.

Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, Yasaman Bahri, Victor V. Albert, Pranesh Srinivasan, Haining Pan, Philippe Faist, Brian Rohr, Michael Statt, Dan Morris, Drew Purves, Elise Kleeman, Ruth Alcantara, Matthew Abraham, Muqthar Mohammad, Ean VanLee, Chenfei Jiang, Elizabeth Dorfman, Eun-Ah Kim, Michael Brenner, Sameera Ponda, and Subhashini Venugopalan. Curie: Evaluating llms on multitask scientific long-context understanding and reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 9859–9906, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1ae4999aefb509d75d8608e07280922c-Paper-Conference.pdf.

DeepSeek-AI. Deepseek-v4 preview: Entering the era of affordable million-token context. DeepSeek, April 2026. URL https://www.deepseek.com/en/news/v4-preview/.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, Sep 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00716-3. URL https://doi.org/10.1038/s42256-023-00716-3.

Ming Du, Yanqi Luo, Srutarshi Banerjee, Michael Wojcik, Jelena Popovic, and Mathew J. Cherukara. Experiment automation agents: automating materials characterization with vision language model agents. *npj Computational Materials*, Jul 2026. ISSN 2057-3960. doi: 10.1038/s41524-026-02213-8. URL https://doi.org/10.1038/s41524-026-02213-8.

Alasdair R. Fairhurst, Joshua Snyder, Chao Wang, Dusan Strmcnik, and Vojislav R. Stamenkovic. Electrocatalysis: From planar surfaces to nanostructured interfaces. *Chemical Reviews*, 125(3):1332–1419, 01 2025. ISSN 0009-2665. doi: 10.1021/acs.chemrev.4c00133. URL https://doi.org/10.1021/acs.chemrev.4c00133.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 48221–48251, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d346d91999074dd8d6073d4c3b13733b-Paper-Conference.pdf.

Julia Fernández-Vidal and Marc TM Koper. The role of surface science in electrocatalysis. *ACS catalysis*, 16(4):2925, 2026.

GLM-V Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, Xijun Liu, Wenmeng Yu, Weihan Wang, Wei Li, Shuaiqi Duan, Sheng Yang, Ruiliang Lv, Mingdao Liu, Lihang Pan, Ke Ning, Junhui Ji, Jinjiang Wang, Jing Chen, Jiazheng Xu, Jiale Zhu, Jiale Cheng, Ji Qi, Guobing Gan, Guo Wang, Cong Yao, et al. Glm-5v-turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arXiv:2604.26752*, 2026. URL https://arxiv.org/abs/2604.26752.

Google DeepMind. Gemini 3.1 pro model card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

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, jun 2017. doi: 10.1088/1361-648X/aa680e. URL https://doi.org/10.1088/1361-648X/aa680e.

Scott A. Hollingsworth and Ron O. Dror. Molecular dynamics simulation for all. *Neuron*, 99 (6):1129–1143, Sep 2018. ISSN 0896-6273. URL https://www.sciencedirect.com/science/article/pii/S0896627318306846.

Ziyang Huang, Yi Cao, Ali K. Shargh, Jing Luo, Ruidong Mei, Mohd Zaki, Zhan Liu, Wyatt Bunstine, William Jurayj, Somdatta Goswami, Tyrel McQueen, Michael Shields, Jaafar El-Awady, Paulette Clancy, Benjamin Van Durme, Nicholas Andrews, William Walden, and Daniel Khashabi. Can coding agents reproduce findings in computational materials science?, 2026. URL https://arxiv.org/abs/2605.00803.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, Feb 2024. ISSN 2522-5839. doi: 10.1038/s42256-023-00788-1. URL https://doi.org/10.1038/s42256-023-00788-1.

A. Gilad Kusne and Austin McDannald. Managing autonomous materials labs with multi-agent ai and its implications for the science of science. *Communications Materials*, 7(1):173, Jul 2026. ISSN 2662-4443. doi: 10.1038/s43246-026-01219-5. URL https://doi.org/10.1038/s43246-026-01219-5.

Zhengzhao Lai, Youbin Zheng, Zhenyang Cai, Haonan Lyu, Jingpu Yang, Hong-Qing Liang, Yan Hu, and Benyou Wang. Can multimodal LLMs see materials clearly? a multimodal benchmark on materials characterization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4384–4404, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.235. URL https://aclanthology.org/2025.findings-emnlp.235/.

Yueqing Liang, Liangwei Yang, Chen Wang, Congying Xia, Rui Meng, Xiongxiao Xu, Haoran Wang, Ali Payani, and Kai Shu. Benchmarking llms for political science: A united nations perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 738–745, 2026.

Quanliang Liu, Maciej P. Polak, So Yeon Kim, MD Al Amin Shuvo, Hrishikesh Shridhar Deodhar, Jeongsoo Han, Dane Morgan, and Hyunseok Oh. Beyond designer's knowledge: Generating materials design hypotheses via a large language model. *Acta Materialia*, 297:121307, 2025a. ISSN 1359-6454. doi: https://doi.org/10.1016/j.actamat.2025.121307. URL https://www.sciencedirect.com/science/article/pii/S1359645425005932.

Siyu Liu, Bo Hu, Beilin Ye, Jiamin Xu, David J. Srolovitz, and Tongqi Wen. Mattools: Benchmarking large language models for materials science tools, 2025b. URL https://arxiv.org/abs/2505.10852.

Taoyuze Lv, Alexander Chen, Fengyu Xie, Chu Wu, Jeffrey Meng, Dongzhan Zhou, Yingheng Wang, Bram Hoex, Zhicheng Zhong, and Tong Xie. Atomworld: A benchmark for evaluating spatial reasoning in large language models on crystalline materials, 2026. URL https://arxiv.org/abs/2510.04704.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00832-8. URL https://doi.org/10.1038/s42256-024-00832-8.

Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M. Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and N. M. Anoop Krishnan. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16(1):9104, Oct 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64105-7. URL https://doi.org/10.1038/s41467-025-64105-7.

Koichi Momma and Fujio Izumi. *VESTA3* for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of Applied Crystallography*, 44(6):1272–1276, Dec 2011. doi: 10.1107/S0021889811038970. URL https://doi.org/10.1107/S0021889811038970.

Moonshot AI. Kimi k2.6: From code to creation, from one to many. Moonshot AI, April 2026. URL https://www.kimi.ai/ai-models/kimi-k2-6.

OpenAI. Gpt-5 system card. OpenAI, August 2025. URL https://openai.com/index/gpt-5-system-card/.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. OpenAI, July 2026. URL https://openai.com/index/gpt-5-6/.

Jinyoung Park, Minseong Bae, Dohwan Ko, and Hyunwoo J. Kim. Llamo: Large language model-based molecular graph assistant. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 131972–132000. Curran Associates, Inc., 2024. doi: 10.52202/079017-4195. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ee46288ab2aaf5c6e53aebebe719712c-Paper-Conference.pdf.

Qwen Team. Qwen3.8-max. Qwen Model Documentation, 2026. URL https://chat.qwen.ai/legal-agreement/models. Accessed September 2026.

Mayk Caldas Ramos, Christopher J. Collison, and Andrew D. White. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 16(6):2514–2572, 02 2025. ISSN 2041-6520. doi: 10.1039/d4sc03921a. URL https://doi.org/10.1039/d4sc03921a.

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. doi: 10.1088/2632-2153/add3bb. URL https://doi.org/10.1088/2632-2153/add3bb.

Mikkel Ohm Sauer, Peder Meisner Lyngby, and Kristian Sommer Thygesen. Dispersion-corrected machine learning potentials for 2d van der waals materials. *Phys. Rev. Mater.*, 9:074007, Jul 2025. doi: 10.1103/cl8c-8f1f. URL https://link.aps.org/doi/10.1103/cl8c-8f1f.

ScienceOne AI and Wenge AI. S1-omni: A unified multimodal reasoning model for scientific understanding, prediction, and generation, 2026. URL https://arxiv.org/abs/2607.15686.

Yingheng Tang, Wenbin Xu, Jie Cao, Weilu Gao, Steven Farrell, Benjamin Erichson, Michael W. Mahoney, Andy Nonaka, and Zhi Jackie Yao. A multimodal large language model for materials science. *Nature Machine Intelligence*, 8(4):588–601, Apr 2026a. ISSN 2522-5839. doi: 10.1038/s42256-026-01214-y. URL https://doi.org/10.1038/s42256-026-01214-y.

Zhichen Tang, Zhengzheng Dang, Yulin Chen, Jixin Wu, Haiwen Li, and Yanming Wang. Evomd-llm: Learning the language of species evolution in reactive molecular dynamics, 2026b. URL https://arxiv.org/abs/2605.29394.

Haorui Wang, Jeff Guo, Lingkai Kong, Rampi Ramprasad, Philippe Schwaller, Yuanqi Du, and Chao Zhang. LLM-augmented chemical synthesis and design decision programs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62980–63001. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/wang25ag.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. URL https://arxiv.org/abs/2508.18265.

Yiheng Wang, Yixin Chen, Shuo Li, Yifan Zhou, Bo Liu, Hengjian Gao, Jiakang Yuan, Jia Bu, Wanghan Xu, Yuhao Zhou, Xiangyu Zhao, Zhiwang Zhou, Fengxiang Wang, Haodong Duan, Songyang Zhang, Jun Yao, Han Deng, Yizhou Wang, Jiabei Xiao, Jiaqi Liu, Encheng Su, Yujie Liu, Weida Wang, Junchi Yao, Shenghe Zheng, Haoran Sun, Runmin Ma, Xiangchao Yan, Bo Zhang, Dongzhan Zhou, Shufei Zhang, Peng Ye, Xiaosong Wang, Shixiang Tang, Wenlong Zhang, and Lei Bai. Scievalkit: An open-source evaluation toolkit for scientific general intelligence, 2026. URL https://arxiv.org/abs/2512.22334.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Yike Ye, ChenCheng Dai, Kamal Elouarzaki, Gang Wu, and Zhichuan J Xu. Catalysts for electrochemical oxidation of ammonia: A comprehensive review of fundamentals and optimization strategies. *Advanced Materials*, 38(4):e09364, 2026.

Z.AI. Glm-5.3. Z.AI Model Documentation, 2026. URL https://huggingface.co/zai-org/GLM-5.3. Accessed September 2026.

Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, and Wei Wang. Matscibench: Benchmarking the reasoning ability of large language models in materials science. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 12868–12879, New York, NY, USA, 2026a. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3818888. URL https://doi.org/10.1145/3770855.3818888.

Shuchen Zhang, Yuan Lu, Linghai Zhang, Hongsheng Shi, Xiao Wang, Zhaoming Wang, Jiaxin Li, Zehua He, Kang Wang, Ke Ma, et al. Mosaic lateral heterostructures in two-dimensional perovskite. *Nature*, 649(8097):612–620, 2026b.

A Detailed Task Descriptions

A.1 Stratified sampling and final size.

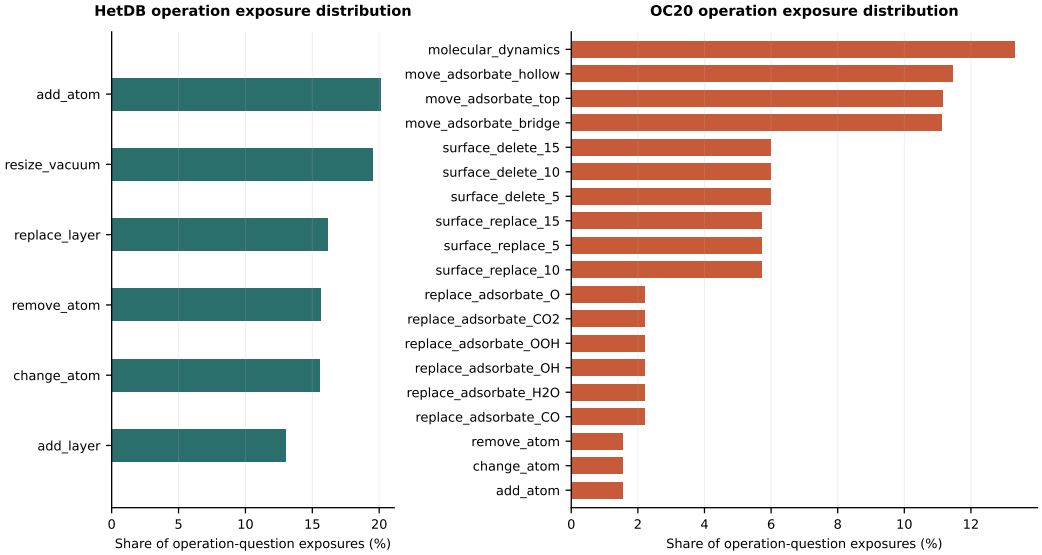


Figure 5: Operation exposure distributions for HetDB and OC20.

Multimodal tasks retain the reference answers and option mappings of their text counterparts and require complete visual inputs. For each text task, all eligible instances are retained when their number does not exceed 1,000; otherwise, 400 instances are selected through stratified sampling by operation and structural composition. The operation distributions cover atomic, molecular, surface, layer, and environmental perturbations across the two systems (Fig. 5). Multimodal counterparts of subsampled tasks inherit the selection status of the corresponding text instances, while other tasks retain all available multimodal instances. The final benchmark contains 11,075 questions: 6,824 text and 4,251 multimodal instances. By system, HetDB contributes 4,002 instances and OC20 contributes 7,073. Counts refer to question instances; multiple structures, views, or trajectory frames within a question are not counted separately. The detailed distribution is provided in Table 5.

Table 5: Detailed task taxonomy of **MultiPhaseWorld**. S-S and S-G denote solid-solid and solid-gas systems, respectively.

Task	Level	System	Evaluation Objective	N
			Text-based Tasks	
T1	L1	S-G	Classify surface adsorbates as gas molecules, reaction intermediates, or mixed.	400
T2	L1	S-G	Identify the adsorption site of the target adsorbate.	400
T3	L1	S-G	Identify the structure with the highest surface-defect concentration.	783
T4	L2	S-S	Determine whether perturbation induces significant global reconstruction.	400
T5	L2	S-S	Select the structure exhibiting the most interface-concentrated reconstruction after the specified perturbation.	251
T6	L2	S-G	Determine whether the adsorption site changes after perturbation.	400
T7	L2	S-G	Identify the most stable adsorption site.	351
T8	L3	S-S	Identify the most stable state from an evolution trajectory.	400

Continued on next page

Table 5 – continued from previous page

Task	Level	System	Evaluation Objective	N
T9	L3	S-G	Identify the most stable state from an evolution trajectory.	400
T10	L3	S-S	Identify the least stable state from an evolution trajectory.	400
T11	L3	S-G	Identify the least stable state from an evolution trajectory.	400
T12	L3	S-S	Recover the structural evolution sequence following perturbation from shuffled frames.	400
T13	L3	S-G	Recover the structural evolution sequence following perturbation from shuffled frames.	400
T14	L3	S-G	Identify the structure corresponding to the highest temperature.	400
T15	L3	S-G	Recover the environmental evolution sequence following perturbation from shuffled frames.	400
T16	L4	S-G	Predict the preferred adsorption site for target adsorbate.	339
T17	L4	S-S	Predict which perturbation produces the strongest global reconstruction.	300
Multimodal Tasks				
MT1	L1	S-G	Identify the adsorption site of the target adsorbate from structural images.	400
MT2	L2	S-S	Determine whether perturbation induces significant global reconstruction from structural images.	400
MT3	L2	S-S	Select the structure exhibiting the most interface-concentrated reconstruction after the specified perturbation from structural images.	251
MT4	L2	S-G	Determine whether the adsorption site changes after perturbation from structural images.	400
MT5	L3	S-S	Identify the most stable state along a post-perturbation evolution trajectory from structural images.	400
MT6	L3	S-G	Identify the most stable state along a post-perturbation evolution trajectory from structural images.	400
MT7	L3	S-S	Identify the least stable state along a post-perturbation evolution trajectory from structural images.	400
MT8	L3	S-G	Identify the least stable state along a post-perturbation evolution trajectory from structural images.	400
MT9	L3	S-S	Recover the structural evolution sequence following perturbation from shuffled structural images.	400
MT10	L3	S-G	Recover the structural evolution sequence following perturbation from shuffled structural images.	400
MT11	L3	S-G	Order evolution trajectories by temperature from multi-frame structural sequences.	400

A.2 Tasks for expert evaluation.

The human evaluators are seven PhD-level experts specializing in computational physical chemistry, familiar with standard simulation methods and the interpretation of characterization data. The evaluation set comprises 70 questions: 40 text-based and 30 multimodal questions. Each evaluator is asked to complete 10 questions per modality for L1-L3, and 10 text-based questions for L4. Questions were sampled without replacement from benchmark instances marked as available and selected, using a fixed random seed of 20260915. Sampling was performed independently within each level–modality stratum, and model correctness was not used as a sampling criterion. The sampled instance identifiers and their task mappings are provided in the supplementary materials for reproducibility. Table 6 summarizes the task assignments, modalities, and physical systems.

Table 6: Sample-level task assignments for expert performance evaluation. Task identifiers follow Table 5.

Sample ID	Level	Task	Modality	System
L1_text_01	L1	T3	Text	Solid-gas
L1_text_02	L1	T1	Text	Solid-gas
L1_text_03	L1	T1	Text	Solid-gas
L1_text_04	L1	T2	Text	Solid-gas
L1_text_05	L1	T3	Text	Solid-gas
L1_text_06	L1	T1	Text	Solid-gas
L1_text_07	L1	T3	Text	Solid-gas
L1_text_08	L1	T2	Text	Solid-gas
L1_text_09	L1	T1	Text	Solid-gas
L1_text_10	L1	T3	Text	Solid-gas
L1_multimodal_01	L1	MT1	Multimodal	Solid-gas
L1_multimodal_02	L1	MT1	Multimodal	Solid-gas
L1_multimodal_03	L1	MT1	Multimodal	Solid-gas
L1_multimodal_04	L1	MT1	Multimodal	Solid-gas
L1_multimodal_05	L1	MT1	Multimodal	Solid-gas
L1_multimodal_06	L1	MT1	Multimodal	Solid-gas
L1_multimodal_07	L1	MT1	Multimodal	Solid-gas
L1_multimodal_08	L1	MT1	Multimodal	Solid-gas
L1_multimodal_09	L1	MT1	Multimodal	Solid-gas
L1_multimodal_10	L1	MT1	Multimodal	Solid-gas
L2_text_01	L2	T5	Text	Solid-solid
L2_text_02	L2	T5	Text	Solid-solid
L2_text_03	L2	T4	Text	Solid-solid
L2_text_04	L2	T4	Text	Solid-solid
L2_text_05	L2	T6	Text	Solid-gas
L2_text_06	L2	T5	Text	Solid-solid
L2_text_07	L2	T4	Text	Solid-solid
L2_text_08	L2	T6	Text	Solid-gas
L2_text_09	L2	T6	Text	Solid-gas
L2_text_10	L2	T6	Text	Solid-gas
L2_multimodal_01	L2	MT2	Multimodal	Solid-solid
L2_multimodal_02	L2	MT4	Multimodal	Solid-gas
L2_multimodal_03	L2	MT2	Multimodal	Solid-solid
L2_multimodal_04	L2	MT4	Multimodal	Solid-gas
L2_multimodal_05	L2	MT2	Multimodal	Solid-solid
L2_multimodal_06	L2	MT2	Multimodal	Solid-solid
L2_multimodal_07	L2	MT2	Multimodal	Solid-solid
L2_multimodal_08	L2	MT2	Multimodal	Solid-solid
L2_multimodal_09	L2	MT2	Multimodal	Solid-solid
L2_multimodal_10	L2	MT3	Multimodal	Solid-solid
L3_text_01	L3	T12	Text	Solid-solid
L3_text_02	L3	T9	Text	Solid-gas
L3_text_03	L3	T9	Text	Solid-gas
L3_text_04	L3	T10	Text	Solid-solid
L3_text_05	L3	T9	Text	Solid-gas
L3_text_06	L3	T12	Text	Solid-solid
L3_text_07	L3	T10	Text	Solid-solid
L3_text_08	L3	T8	Text	Solid-solid
L3_text_09	L3	T9	Text	Solid-gas
L3_text_10	L3	T8	Text	Solid-solid
L3_multimodal_01	L3	MT7	Multimodal	Solid-solid

Continued on next page

Table 6 – continued

Sample ID	Level	Task	Modality	System
L3_multimodal_02	L3	MT8	Multimodal	Solid-gas
L3_multimodal_03	L3	MT5	Multimodal	Solid-solid
L3_multimodal_04	L3	MT8	Multimodal	Solid-gas
L3_multimodal_05	L3	MT11	Multimodal	Solid-gas
L3_multimodal_06	L3	MT8	Multimodal	Solid-gas
L3_multimodal_07	L3	MT9	Multimodal	Solid-solid
L3_multimodal_08	L3	MT11	Multimodal	Solid-gas
L3_multimodal_09	L3	MT10	Multimodal	Solid-gas
L3_multimodal_10	L3	MT8	Multimodal	Solid-gas
L4_text_01	L4	T17	Text	Solid-solid
L4_text_02	L4	T17	Text	Solid-solid
L4_text_03	L4	T17	Text	Solid-solid
L4_text_04	L4	T16	Text	Solid-gas
L4_text_05	L4	T17	Text	Solid-solid
L4_text_06	L4	T16	Text	Solid-gas
L4_text_07	L4	T16	Text	Solid-gas
L4_text_08	L4	T16	Text	Solid-gas
L4_text_09	L4	T16	Text	Solid-gas
L4_text_10	L4	T17	Text	Solid-solid

B SOURCE-DATA AND OPERATION DISTRIBUTION

OC20 sampling retains all 46 adsorbates while reducing the dominance of frequent classes (Fig. 8). The datasets provide complementary coverage: HetDB spans diverse structure sizes and 28 elements, whereas selected OC20 structures mainly contain 40-49 atoms and cover all 55 elements (Figs. 6 and 7).

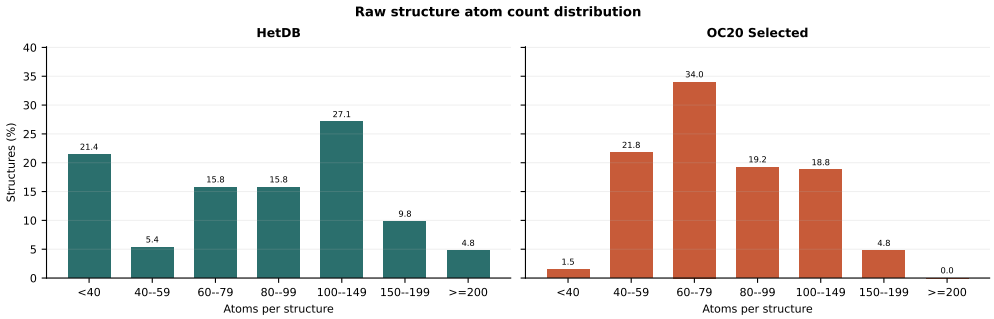


Figure 6: Raw structure atom-count distributions for HetDB (left) and OC20 Selected (right).

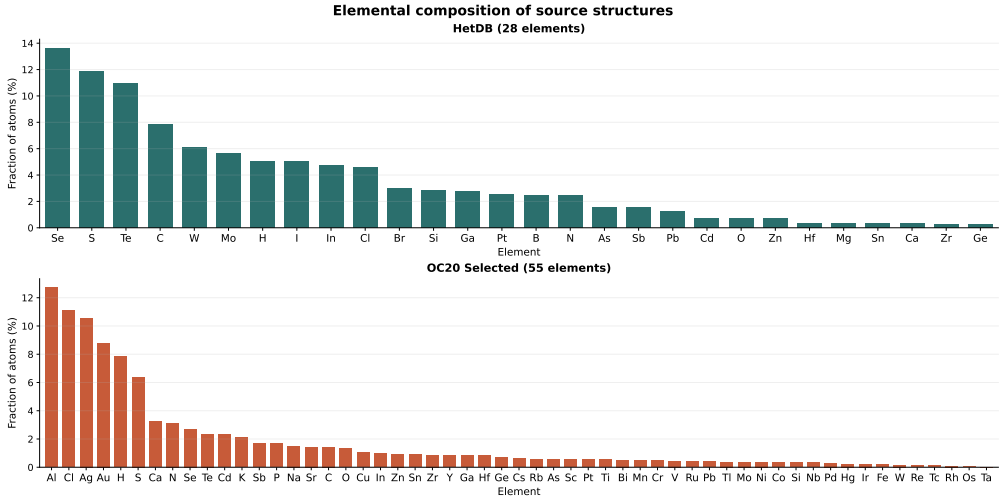


Figure 7: Elemental composition of the HetDB (top) and OC20 Selected (bottom) source structures.

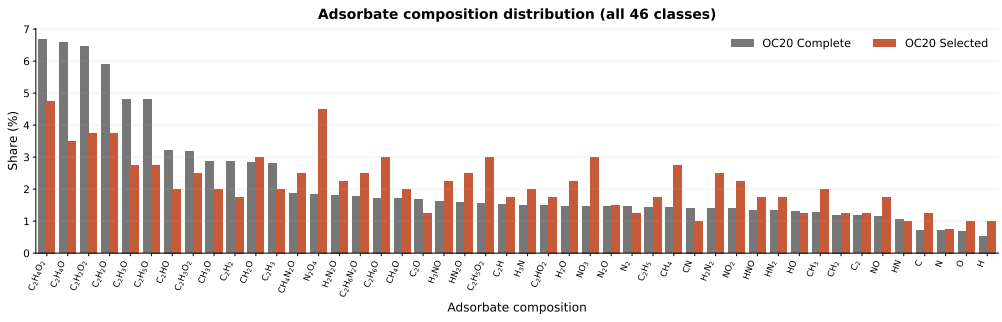


Figure 8: Adsorbate-composition distributions for OC20 Complete and OC20 Selected. All 46 composition classes are shown.

C Operation Generation

MultiPhaseWorld constructs perturbed structures through five physically motivated operation levels: atomic, molecular, surface, layer, and environmental perturbations. All stochastic choices in the perturbation pipeline are reproducible. For each structure, operation, and retry attempt, a deterministic seed is generated as

$$s = \mathrm{SHA256}\left(s_{\mathrm{master}}, \text{structure ID}, \text{operation}, \text{attempt}\right),$$

and all subsequent random choices use the corresponding local pseudorandom number generator. The basic supported element set is defined as the elements supported by the MACE-MP-0 potential, excluding noble gases and actinide elements by default:

$$\mathcal{E}_{\mathrm{sup}} = \{1, \ldots, 83\} \setminus \{2, 10, 18, 36, 54\},$$

where each integer denotes an atomic number. All operations that introduce new atoms draw candidate elements exclusively from $\mathcal{E}_{\mathrm{sup}}$. For all operations, operated structures are accepted only when they pass the *global pair–distance validity*:

$$d_{ij} \geq \max\left(0.50\ \text{Å}, 0.55\left(r_i^{\mathrm{cov}} + r_j^{\mathrm{cov}}\right)\right),$$

where r_i^{cov} is the ASE covalent radius. If multiple candidates pass this validity check, one candidate is selected with the deterministic local seed.

C.1 Atomic Operations.

Atomic perturbations are applied to both HetDB and OC20 and include atom *addition*, *deletion*, and *substitution*. Each structure–operation pair is tried at most three times with deterministic retry seeds. In each retry, *add atom* samples up to 500 candidate insertions, *remove atom* enumerates all possible deletion candidates, and *change atom* tests up to 800 shuffled elemental substitutions in HetDB or all radius–compatible substitutions in OC20. Candidates that fail the *global pair-distance validity* are discarded. If no valid candidate is found after all retries, the operation is recorded as failed.

Atom addition An existing atom is randomly selected as an anchor atom i, and a new element $e \in \mathcal{E}_{\mathrm{sup}}$ is sampled. The inserted atom is placed along a random three-dimensional unit vector $\mathbf{u}$ at

$$\mathbf{r}_{\mathrm{new}} = \mathbf{r}_i + \lambda\left(r_i^{\mathrm{cov}} + r_e^{\mathrm{cov}}\right)\mathbf{u}, \qquad \lambda \sim \mathcal{U}(1.0, 1.5).$$

Atom deletion All atoms are enumerated as deletion candidates. A candidate is valid only if the resulting structure contains at least two atoms. One valid deletion candidate is then selected using the deterministic local seed.

Atom substitution Substitution candidates are generated by replacing one atom with a different element from $\mathcal{E}_{\mathrm{sup}}$. For OC20, the replacement element is additionally required to have a covalent radius within 0.25 Å of the original element, avoiding overly disruptive substitutions in surface systems.

C.2 Molecular Level Operations

Molecular perturbations are applied to OC20 adsorbate–surface systems and include adsorbates substitution (with reaction intermediates or gas molecules) and adsorption site relocation. Adsorbate atoms are identified from OC20 atomic tags ($\mathrm{tag} \geq 2$) and grouped into molecular components using periodic covalent connectivity,

$$d_{ij}^{\mathrm{3D}} \leq 1.25\left(r_i^{\mathrm{cov}} + r_j^{\mathrm{cov}}\right).$$

The benchmark defines the adsorbate anchor as the target adsorbate atom closest to the slab, and constructs site candidates from non-adsorbate surface atoms. Many operations require **adsorption site detection**. Surface atoms are inferred from the uppermost slab region. Surface neighbor edges are defined by the periodic lateral-distance criterion

$$d_{ij}^{xy} \leq \alpha_{\mathrm{nn}}\left(r_i^{\mathrm{cov}} + r_j^{\mathrm{cov}}\right),$$

where α_{nn} is the neighbor-radius factor. A **top site** is defined by a single surface atom and uses that atom's in-plane coordinate as the site center. A **bridge site** is defined by a neighboring pair of surface atoms and uses their periodic in-plane midpoint as the site center; bridge candidates that are within 0.50 Å of a hollow center are discarded so that hollow sites have priority. A **hollow site** is defined by a three- or four-atom surface pocket in the surface-neighbor graph and uses the periodic in-plane centroid of the pocket as the site center; for corrugated or defective surfaces, compact non-hydrogen three-atom pockets are additionally considered as fallback hollow candidates. For a candidate site $C = \{p_1, \ldots, p_n\}$, where p_i denotes a surface atom defining the site, we define the number of effective adsorbate–surface interactions as

$$N_{\mathrm{int}}(C) = \sum_{p \in C} \mathbf{1}\left[d_{ap}^{\mathrm{3D}} \leq 1.5\left(r_a^{\mathrm{cov}} + r_p^{\mathrm{cov}}\right)\right],$$

where a denotes the adsorbate anchor atom. A candidate site is considered valid if $N_{\mathrm{int}}(C) = 1$ for a top site, $N_{\mathrm{int}}(C) = 2$ for a bridge site, or $N_{\mathrm{int}}(C) \geq 2$ for a hollow site. Among all valid candidates, the assigned site label is the type whose site center is nearest to the anchor projection in periodic in-plane distance. If the anchor has no effective interaction with the surface, or if no valid candidate can be identified, the structure is labeled as $\mathrm{detached}$.

Adsorbate replacement The target adsorbate is chosen from CO_2, H_2O, CO, O, OH, and OOH. Adsorbates already matching the target formula are excluded. One remaining component is sampled uniformly at random and replaced by a template anchored at the original adsorption position. The anchor is the non-hydrogen atom closest to the surface, with all atoms considered for hydrogen-only adsorbates. Candidate orientations are generated by rotation around the surface normal, and the first geometry-valid structure is retained.

Adsorption site relocation The primary adsorbate is selected as the largest adsorbate component, with surface distance used as a tie-breaker, and its closest atom to the slab is used as the anchor. Target sites (top, bridge, or hollow) are generated using the definitions above. The adsorbate is translated to the target site while adjusting its height to satisfy geometric and interaction constraints. Strict placements with interaction factors 1.35 and 1.5 are attempted first, followed by a collision-free geometric fallback if needed. If the original site already matches the target, no translation is applied and the operation is recorded as an identity perturbation.

C.3 Surface Level Operations

Surface perturbations are applied to OC20 slabs and include creating *vacancy and substitution defects* through surface atom deletion and substitution at nominal concentrations of 5%, 10%, and 15%. Surface atoms are identified using OC20 surface tags when available; otherwise, they are inferred from the uppermost slab region. For a target concentration $c \in \{0.05, 0.10, 0.15\}$, the number of modified atoms is

$$N_{\mathrm{mod}} = \max\left(1, \mathrm{round}\left(cN_{\mathrm{surf}}\right)\right),$$

where N_{surf} is the number of detected surface atoms. Deleted or substituted atoms are selected with a minimum mutual spacing constraint. For surface substitution, each selected atom is replaced by a supported element with a similar covalent radius, using a 0.25 Å radius tolerance when available.

C.4 Layer Level Operations

Layer perturbations are applied to HetDB solid–solid structures and include *monolayer element replacement*, *monolayer addition*, and *vacuum resizing*. Atomic layers are identified along the surface-normal direction. The largest layer gap is treated as the vacuum gap, and the remaining slab is split into lower and upper monolayer domains at the internal adjacent-layer gap with the largest compositional dissimilarity. Layer replacement substitutes all atoms of a selected element within one selected monolayer domain. Layer addition copies one monolayer domain and places it above or below a selected monolayer, using a random lateral shift and the closest valid vertical separation. Vacuum resizing changes only the surface-normal cell length while preserving all Cartesian atomic positions.

C.5 Environmental Operations

Environmental perturbations are applied to OC20 through temperature-conditioned molecular dynamics simulations. For each selected OC20 structure, NVT trajectories are generated at 200, 300, and 700 K, corresponding to a low-temperature condition below room temperature, a near-room-temperature condition, and a high-temperature condition above room temperature, respectively. Text-based temperature-reasoning tasks use the final saved structure from each temperature trajectory, whereas multimodal video-style tasks use the last five consecutive saved frames from each trajectory.

D SIMULATION SETTINGS

Interatomic potential and structural relaxation The benchmark configuration uses the medium `MACE-MP-0` (Batatia et al., 2025), a machine-learned interatomic potential (MLP) trained on the Materials Project Trajectory (MPtrj) dataset (Deng et al., 2023), to evaluate energies and forces, with double-precision arithmetic and no additional dispersion correction. Perturbed HetDB and OC20 structures are relaxed using the BFGS optimizer in the Atomic Simulation Environment (ASE) (Hjorth Larsen et al., 2017). The force convergence threshold is $0.03\ \mathrm{eV\ \AA^{-1}}$, with a maximum of 1,000 optimization steps. Atomic positions are optimized at fixed cell vectors; the relaxation routine does not perform cell optimization. These trajectories represent optimization paths that gradually converge to a stable state following perturbation. If the relaxation fails to converge, the attempt is terminated and the status and reasons are logged. Four structures are retained at relative trajectory indices 0, $1/3$, $1/2$, and 1, rounded to the nearest available index. The configured operation sampling seeds are 20260812 for HetDB and 20260813 for OC20, with up to three attempts per operation.

Finite-temperature molecular dynamics We simulate OC20 structures at 200, 300, and 700 K using LAMMPS with the double-precision `MACE-MP-0` medium potential. Each run uses an NVT ensemble with fixed cell vectors, fully periodic boundaries, and the `metal` unit system. Initial velocities are drawn from a Gaussian distribution at the target temperature, and net translational and rotational momentum is removed. A single NVT thermostat is applied to all atoms; no pre-minimization or separate equilibration stage is used. The base velocity seed is `20260815`, incremented per source structure index, and thermodynamics plus atomic snapshots are output every 100 steps.

MD outputs The MD simulation pipeline retains the complete simulation trajectory, accompanied by thermodynamic data; the benchmark utilizes the final five consecutive frames as input candidates. The per-run summary log records the following parameters: integration time step, number of steps, thermostat damping time, output frequency, velocity random seed, potential model path, and execution status. Additionally, the structural relaxation pipeline logs additional system configuration details and software versions in its run manifest.

Table 7: Molecular dynamics simulation settings.

Parameter	Setting
Engine	LAMMPS
Potential	MACE-MP-0 medium
Precision	Float64
Ensemble	NVT
Target temperatures	200, 300, 700 K
Boundary conditions	Periodic in x, y, and z
Integration time step Δt	0.001 ps
Integration steps N	5000
Trajectory duration $N\Delta t$	5 ps
Thermostat damping time	0.1 ps
Thermodynamic output interval	100 steps (configured default)
Snapshot interval	100 steps (configured default)

E GROUNDED REFERENCE CONSTRUCTION

All reference answers in **MultiPhaseWorld** are generated automatically from dataset labels, operation metadata, optimized structures, simulated energies, and stored trajectories. No LLM is involved in ground-truth construction, and questions are generated only when the required records pass the corresponding validity checks.

Structure and composition references Adsorbates are classified as gas molecules or reaction intermediates from their recorded formulas. The adsorbates in S^{evo} are identified using atomic tags and partitioned into individual molecular components based on covalent radius distance criteria under periodic boundary conditions with a covalent radius factor of 1.25. Surface-defect concentration is computed as

$$c_{\mathrm{defect}} = \frac{N_{\mathrm{defect}}}{N_{\mathrm{surface}}}, \tag{1}$$

where N_{defect} is the number of defect atoms and N_{surface} is the number of original surface atoms. Adsorption sites are determined from optimized geometries using the benchmark site detection procedure (next paragraph) and classified as top, bridge, hollow, or detached.

Adsorption site references. For adsorption site tasks, all labels are based on the actual site detected from the optimized geometry rather than the nominal operation name. The adsorption site is determined from the optimized geometry by **adsorption site detection** in Appendix C.2. For each target adsorbate, the optimized structure is classified as *top*, *bridge*, or *hollow*. *Site change tasks* compare the initial and final detected site labels of the target adsorbate; identical labels are counted as no change, including identity relocation operations where the operated structure is unchanged, while different labels are counted as a site change. *Stable site tasks* compare relaxed candidates by their final detected adsorption sites and energies. Because relaxation may move an adsorbate away from its nominal target, the benchmark assigns each candidate to its final non–detached site before comparison. When multiple candidates relax to the same site, those preserving the nominal target are preferred; otherwise, the lowest–energy candidate is used. The answer is the lowest–energy final adsorption site, with configurations within $0.05\ \mathrm{eV}$ treated as nearly degenerate.

Structural reconstruction references For solid–solid systems, each structure is represented as an atom-mapped graph $G = (V, E)$. Two atoms are connected when

$$d_{ij} \le 1.2\left(r_i^{\mathrm{cov}} + r_j^{\mathrm{cov}}\right). \tag{2}$$

Let E^{ori} and E^{evo} denote the edge sets before and after evolution. The factor of 1.2 defines graph connectivity for reconstruction scoring, whereas the factor of 1.5 used in adsorption-site detection defines geometric neighborhood and interaction criteria.

Connectivity changes are measured by

$$\Delta E = E^{\mathrm{ori}} \triangle E^{\mathrm{evo}}, \tag{3}$$

with

$$|\Delta E| = |E^{\mathrm{ori}} \cup E^{\mathrm{evo}}| - |E^{\mathrm{ori}} \cap E^{\mathrm{evo}}|. \tag{4}$$

The global reconstruction score is defined as

$$R_{\mathrm{edit}} = \frac{|\Delta E| + N_{\mathrm{elem}}^{\mathrm{chg}} + N_{\mathrm{atom}}^{\mathrm{del}}}{|E^{\mathrm{ori}}| + |V^{\mathrm{ori}}|}, \tag{5}$$

where $N_{\mathrm{elem}}^{\mathrm{chg}}$ counts element changes on mapped atoms and $N_{\mathrm{atom}}^{\mathrm{del}}$ counts deleted prototype atoms. Added atoms are not directly counted. Global reconstruction labels are obtained by comparing R_{edit} with the median over valid HetDB perturbation records.

The interface region is defined on the original prototype structure as the two atomic layers adjacent to the internal solid–solid interface gap. Using the same atom-mapped graph, we count all changed edges incident to atoms in this region, together with element changes and deletions of mapped interface atoms:

$$R_{\mathrm{int}} = \frac{|\Delta E_{\mathrm{int}}| + N_{\mathrm{elem,int}}^{\mathrm{chg}} + N_{\mathrm{atom,int}}^{\mathrm{del}}}{|E_{\mathrm{int}}^{\mathrm{ori}}| + |V_{\mathrm{int}}^{\mathrm{ori}}|}. \tag{6}$$

Here, E_{int} contains every edge with at least one endpoint in the original interface region, so the score captures changes in the local coordination of interface atoms rather than only bonds fully contained within the interface layers. Interface-local reconstruction is evaluated only when the original interface atoms can be mapped to the evolved structure and retain the same element inventory, preventing direct element substitutions from being treated as purely structural reconstruction. The *concentration of reconstruction* at the original interface is measured by

$$D_{\mathrm{int}} = \frac{R_{\mathrm{int}}}{R_{\mathrm{edit}}}. \tag{7}$$

A reconstruction is labeled as interface-concentrated when $D_{\mathrm{int}} \geq 1.3$. For comparison tasks, candidates are ranked by R_{int}, and the reference answer is the candidate with the largest interface reconstruction score. Operations that create an additional interface, such as layer addition, are excluded from this comparison.

Energy-based stability references For relaxation trajectories, the least stable frame is defined as the one with the highest energy, while the most stable frame is determined by the lowest energy. Structures within 0.05 eV of the minimum are accepted as equivalent stable answers. The same tolerance is used for adsorption-site stability comparisons. Failed relaxations are excluded.

Trajectory references Temporal-order questions use the original frame indices as ground truth, while the presented frames are randomly relabeled. For finite-temperature MD tasks, references are constructed from successful trajectories at 200, 300, and 700 K. The 700 K trajectory defines the high-temperature reference, and the multimodal sequence task uses the final five frames of each MD trajectory.

Structural prediction references For solid-solid systems, candidate perturbations are ranked by R_{edit}, and the perturbation with the largest score is used as the strongest reconstruction reference. For solid-gas systems, adsorption site prediction is based on relaxed top, bridge, and hollow trials. If the final adsorption sites remain distinct, the lowest-energy configuration defines the preferred site. When multiple trials converge to the same final site, candidates that preserve their nominal target site are preferred before selecting the lowest energy representative. Detached configurations are excluded.

F METRICS AND EXPERIMENTAL RESULTS

Metric Definitions and Computation We evaluate model performance using task accuracy, the gain over the random baseline ($\Delta Rand.$), and the number of task wins (N_{win}). If multiple models achieve the same highest accuracy on a task, each of them is counted as one task win.

- **Accuracy.** For each task t, accuracy is computed as

$$\text{Acc}_t = \frac{N_t^{\text{correct}}}{N_t}, \tag{8}$$

  where N_t is the number of sampled evaluation instances that pass the quality control (Table 5) and N_t^{correct} is the number of correct predictions. The overall accuracy is the macro-average over tasks:

$$\text{AvgAcc} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{Acc}_t, \tag{9}$$

  where $\mathcal{T}$ denotes the corresponding set of text or multimodal tasks.

- **Gain Over Random Baseline.** The above-random gain $\Delta Rand.$ is defined as

$$\Delta\text{Rand.} = \text{AvgAcc} - \text{AvgAcc}_{\text{random}}, \tag{10}$$

  where $\text{AvgAcc}_{\text{random}}$ is the macro-average accuracy of the random-choice baseline.

- **The Number of Task Wins.** N_{win} counts the number of tasks on which a model achieves the highest accuracy among all evaluated models:

$$N_{\text{win}} = \sum_{t \in \mathcal{T}} \mathbb{I}\left[\text{Acc}_t = \max_m \text{Acc}_{m,t}\right]. \tag{11}$$

Detailed results Tables 8 and 9 report detailed task-level results for the text and multimodal settings, respectively. For each task, Avg. denotes the arithmetic mean accuracy across all evaluated models (11 for text and 8 for multimodal tasks), excluding the random baseline. These results complement the aggregate evaluation by showing how model performance varies across reasoning levels, task types, and material systems.

Table 8: Fine-grained text-based evaluation results on **MultiPhaseWorld**. Best-performing model(s) for each task are highlighted in bold. Avg. denotes the arithmetic mean accuracy across all evaluated models for each task, excluding the random baseline.

			Baseline	Open-sourced Models							Closed-sourced Models				
No.	**Task**	**Level**	**Random**	**Qwen3 4B**	**Qwen3 32B**	**DeepSeek-V4 Flash-0731**	**DeepSeek-V4 Pro-0813**	**GLM-5.3**	**Intern-S2-Preview 397B**	**Kimi-K2.6**	**Qwen3.8-Max**	**Gemini-3.1-Pro Preview**	**GPT-5**	**GPT-5.6 Sol**	**Avg.**
T16	Site Prediction (S-G)	4	0.33	0.19	0.18	0.34	0.32	0.47	0.51	0.58	**0.64**	0.63	0.30	0.58	0.43
T17	Reconstruction Prediction (S-S)	4	0.25	0.30	0.35	0.31	0.43	0.57	0.56	0.40	0.57	0.62	0.39	**0.67**	0.47
T8	Stable State (S-S)	3	0.32	**0.27**	0.19	0.22	0.22	0.17	0.18	0.20	0.19	0.18	0.18	0.24	0.20
T9	Stable State (S-G)	3	0.35	0.32	0.36	**0.42**	0.39	0.38	0.41	0.39	0.41	0.38	0.40	0.40	0.39
T10	Unstable State (S-S)	3	0.25	0.31	0.38	0.20	0.18	0.20	0.22	0.13	0.13	**0.42**	0.15	0.21	0.23
T11	Unstable State (S-G)	3	0.25	0.48	0.63	0.76	0.69	0.71	0.73	0.76	0.76	**0.80**	**0.80**	0.76	0.72
T12	Relaxation Order (S-S)	3	0.25	0.33	0.43	0.49	0.52	0.25	0.49	**0.57**	0.32	0.37	0.44	0.32	0.41
T13	Relaxation Order (S-G)	3	0.25	0.27	0.26	0.36	0.41	0.60	0.57	0.42	0.50	**0.76**	0.37	0.63	0.47
T14	High-Temp. Identification (S-G)	3	0.33	0.44	0.65	0.33	0.54	0.72	0.69	0.66	0.71	0.83	0.71	**0.93**	0.66
T15	Temperature Ordering (S-G)	3	0.25	0.27	0.32	0.32	0.31	0.52	0.39	**0.66**	0.45	0.43	0.36	**0.66**	0.43
T4	Global Reconstruction (S-S)	2	0.50	0.55	0.60	0.63	0.47	0.46	0.50	0.69	0.67	0.68	0.60	**0.72**	0.60
T5	Interface Reconstruction (S-S)	2	0.33	0.31	0.50	0.46	0.58	0.70	0.65	0.63	0.64	0.67	0.55	**0.73**	0.58
T6	Site Change (S-G)	2	0.50	0.55	0.49	0.51	0.54	0.53	0.53	0.50	0.51	0.49	0.54	**0.56**	0.52
T7	Site Stability (S-G)	2	0.42	0.41	0.48	0.43	0.40	0.47	**0.51**	0.42	0.46	0.47	0.47	0.48	0.45
T1	Adsorbate Type (S-G)	1	0.33	0.49	0.46	0.48	0.48	0.78	0.52	0.60	0.49	**0.81**	0.46	0.67	0.57
T2	Site Identification (S-G)	1	0.25	0.24	0.20	0.32	0.38	0.45	0.28	0.34	0.42	0.50	0.41	**0.51**	0.37
T3	Defect Concentration (S-G)	1	0.33	0.32	0.43	0.39	0.34	0.77	0.46	0.52	0.49	**0.86**	0.50	0.75	0.53

Table 9: Fine-grained multimodal evaluation results on **MultiPhaseWorld**. Best-performing model(s) for each task are highlighted in bold. Avg. denotes the arithmetic mean accuracy across all evaluated models for each task, excluding the random baseline.

			Baseline	Open-sourced Models			Closed-sourced Models					
No.	**Task**	**Level**	**Random**	**InternVL3.5 241B-A28B**	**Intern-S2-Preview 397B**	**Kimi-K2.6**	**Qwen3.8-Max**	**Gemini-3.1-Pro Preview**	**GPT-5**	**GLM-5V-Turbo**	**GPT-5.6 Sol**	**Avg.**
MT5	Stable State (S-S)	3	0.32	0.26	0.26	0.26	0.26	**0.28**	**0.28**	0.25	0.20	0.26
MT6	Stable State (S-G)	3	0.35	0.35	0.35	0.33	0.33	**0.37**	0.34	0.35	0.35	0.35
MT7	Unstable State (S-S)	3	0.25	0.38	0.22	0.35	0.39	**0.43**	0.38	**0.43**	0.40	0.37
MT8	Unstable State (S-G)	3	0.25	0.54	0.43	**0.59**	0.53	0.57	0.53	**0.59**	0.54	0.54
MT9	Relaxation Order (S-S)	3	0.25	0.21	0.48	0.48	0.35	0.41	0.43	**0.68**	0.55	0.45
MT10	Relaxation Order (S-G)	3	0.25	0.26	0.33	0.41	0.38	**0.76**	0.31	0.27	0.63	0.42
MT11	Temperature Ordering (S-G)	3	0.25	0.34	0.39	0.45	**0.67**	0.56	0.30	0.40	0.60	0.46
MT2	Global Reconstruction (S-S)	2	0.50	0.58	0.48	**0.63**	0.49	0.53	0.50	0.59	0.53	0.54
MT3	Interface Reconstruction (S-S)	2	0.33	0.69	0.69	0.62	**0.71**	0.68	0.63	0.66	0.70	0.67
MT4	Site Change (S-G)	2	0.50	0.53	0.52	0.50	0.52	0.52	0.52	0.53	**0.56**	0.53
MT1	Site Identification (S–G)	1	0.25	0.48	0.34	0.39	0.53	**0.55**	0.47	0.45	0.53	0.47

G Detailed Error Analysis

We present four representative errors from the responses of GPT-5.6 Sol, covering main capability levels. These examples illustrate observable discrepancies between model predictions and the provided references. The CIF structures used in the text-based tasks are visualized using VESTA (Momma & Izumi, 2011).

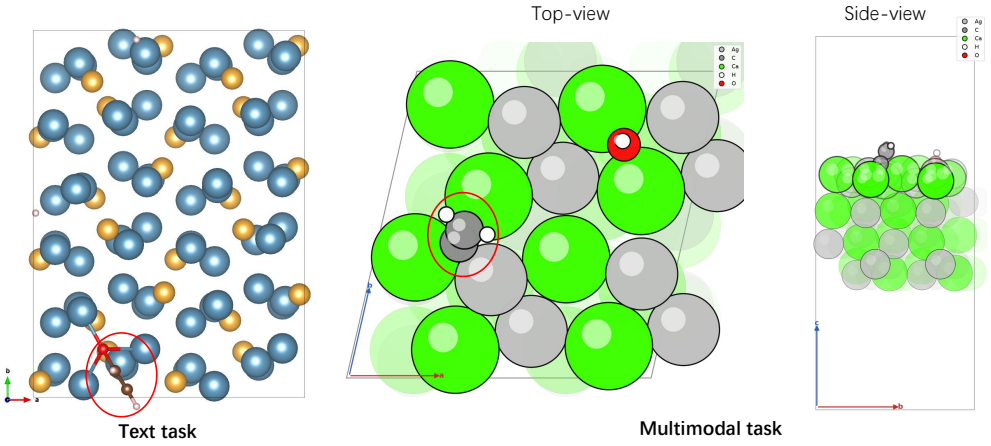


Figure 9: Adsorption site misclassification. CIF visualization examples for the text-based task (left) and the multimodal task (right). GPT-5.6 Sol predicts wrong sites, while the reference identifies hollow sites.

Adsorption site identification As shown in Figure 9, both structures correspond to hollow sites, yet GPT-5.6 Sol makes incorrect predictions (top sites) in both cases. In the left example, the oxygen atom (red) serves as the adsorption anchor and lies near the geometric center of three blue surface atoms, forming a typical hollow site configuration. The right example differs in that the adsorption site is coordinated by two different atomic species (Ca and Ag). These examples indicate that current LLMs still have difficulty reliably recognizing local adsorption geometries.

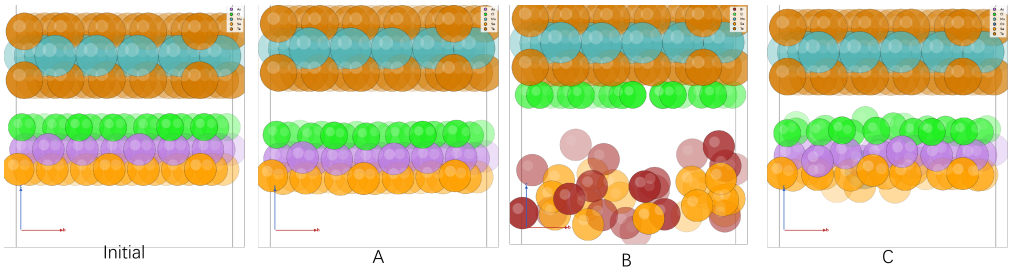


Figure 10: Mislocalization of the Reconstruction Region. Initial denotes the structure before perturbation. A, B, and C show the corresponding optimized stable structures after different operations, viewed along the crystallographic a axis.

Interface reconstruction detection As shown in Figure 10, GPT-5.6 Sol selects structure B, while the correct answer is C. After atomic substitution, structure B undergoes severe and widespread atomic rearrangement, with the structural disruption extending across the entire lower domain. In contrast, structure C largely preserves the layered framework of the initial configuration, and the major structural changes remain localized near the surface or interface region. This error suggests that GPT-5.6 Sol may not reliably distinguish global reconstruction from localized surface or interface reconstruction.

Temperature ordering As shown in Figure 11, the three MD trajectories correspond to different temperatures and are each represented by their final five consecutive frames. The reference order from lowest to highest temperature is $A \rightarrow B \rightarrow C$ (option B). In the displayed views, A exhibits a relatively compact atomic arrangement, whereas C shows a more dispersed local structure, with a potassium atom (purple) on the left protruding beyond the surface. These visible differences provide structural cues for comparing temperature conditions. However, the model selects $C \rightarrow B \rightarrow A$ (option D), completely reversing the reference order. This example illustrates a failure in temperature ordering despite discernible structural differences among the candidate sequences.

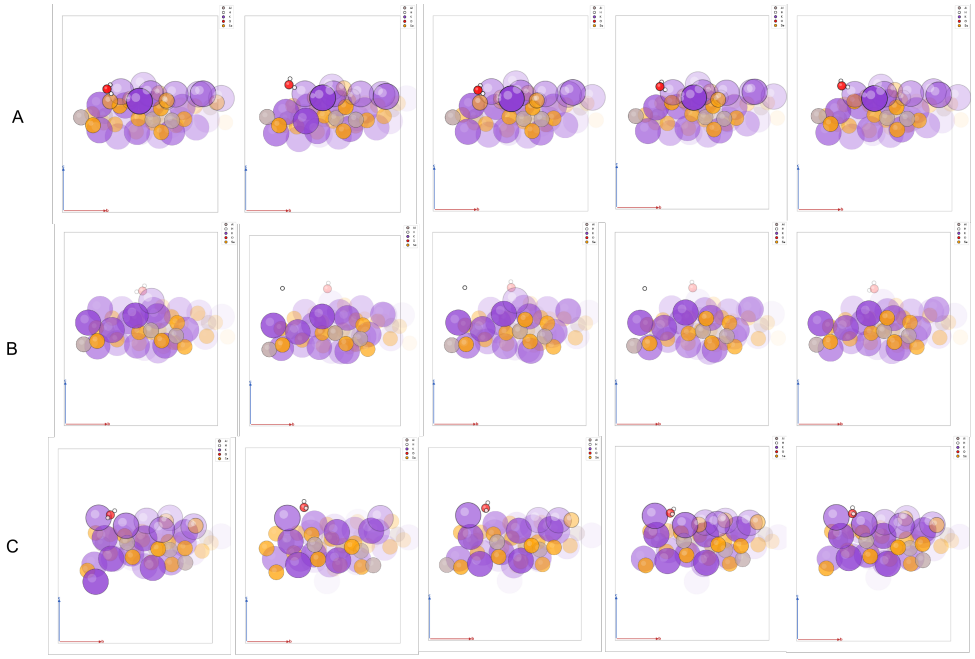


Figure 11: Reversed temperature ordering. The reference order from the lowest to the highest temperature is $A \rightarrow B \rightarrow C$, while GPT-5.6 Sol predicts $C \rightarrow B \rightarrow A$. Frames within each row are presented in their original chronological order.

Evolution ordering. As illustrated in Figure 12, the optimization trajectory shows a boron atom (B) in the upper domain being replaced by an indium atom (In) with a larger atomic radius. After substitution, the surrounding atoms progressively reorganize, causing local bending of the upper domain and driving the In atom away from its initial in-plane position as the structure relaxes toward a more chemically reasonable configuration. The correct temporal order is therefore $A \rightarrow B \rightarrow C \rightarrow D$. However, GPT-5.6 Sol predicts $D \rightarrow C \rightarrow B \rightarrow A$, completely reversing the relaxation sequence. This error indicates difficulty in inferring the direction of structural evolution from successive relaxation states.

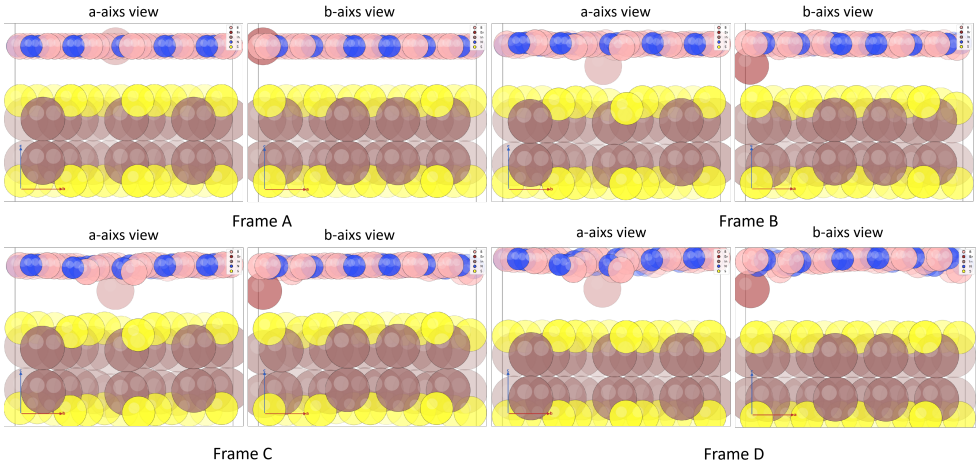


Figure 12: Errors in evolution ordering. Frames A to D show four states sampled from the structural optimization trajectory. For each frame, the structure is visualized from both the crystallographic a-axis and b-axis views.