Towards Generating Stable Materials via Large Language Models with Reinforcement Learning Finetuning

Abstract

Designing stable crystal structures is central to accelerating the discovery of new materials, yet most generative approaches remain limited to reproducing known patterns rather than exploring novel possibilities. We present a method that trains large language models with reinforcement learning guided by verifiable energy-based rewards, optimizing toward physically grounded stability objectives. Compared to supervised finetuning and base models, our reinforcement learning–trained model generates crystals with higher predicted stability and a greater diversity. These results suggest that combining verifiable energy rewards and reinforcement learning provides a powerful path toward automated discovery of novel, stable materials.

1 Introduction

The discovery of new materials drives technological innovation, enabling advances in fields ranging from energy storage [1] to electronics [2] and medicine [3]. Traditionally, the process of identifying and synthesizing novel materials has been slow and resource-intensive, relying on iterative experimentation and domain expertise. Machine learning (ML) offers an alternative, data-driven approach that can accelerate this process by guiding exploration within vast chemical and structural spaces. While predictive models can identify promising candidates from existing materials [4], generative models provide the additional capability of proposing entirely new compositions and structures, thereby expanding the search space beyond existing materials and opening opportunities for breakthrough discoveries.

Modeling materials is particularly challenging because it requires capturing joint distributions over variables of different types: atomic species, which are discrete, and lattice parameters and atomic positions, which are continuous. A further challenge in generative materials design is ensuring that the proposed candidates are **thermodynamically stable**, meaning they can exist without spontaneously breaking down into other material phases. Among the vast number of ways atoms can be arranged, only a tiny fraction corresponds to such stable structures. Despite progress in generative modeling, reliably producing stable candidates under high-fidelity quantum-mechanical (QM) evaluation remains a difficult task [5–8].

Large language models (LLMs) offer a promising avenue for generative materials design due to their ability to incorporate natural language prompting, enabling straightforward and flexible conditioning on desired properties or constraints. Pretrained on broad corpora that include chemistry and materials science knowledge from the scientific literature, these models possess strong prior knowledge of chemical rules and patterns. Fine-tuning an LLM for crystallographic information file (CIF) generation further aligns it with the structural and compositional distributions of known materials, improving its capacity to generate valid candidates.

In this work, we explicitly steer generation toward thermodynamically stable materials by integrating a reinforcement learning (RL) framework with an LLM-based generative model. The RL component

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Science.

provides feedback based on stability evaluations, enabling the model to iteratively refine its outputs and improve the likelihood of producing candidates with low E_{hull}. This integration bridges the flexibility of language-based conditioning with targeted optimization for stability.

2 Related Work

Early work on *Crystal structure Prediction*, the task of generating a crystal structure given a chemical composition, relied on producing candidate materials using atomistic simulations, followed by high-throughput quantum-mechanical calculations [9] to estimate their energies and identify stable structures [10–12]. This screening process can be accelerated using ML interatomic potentials (MLIPs), which are used to predict energy and relax crystals via potential energy minimization, such as CHGNet [13], M3GNet [14], and UMA [15].

More recent efforts have focused on generative modeling to accelerate the discovery of stable materials. Trained on large databases of QM-verified stable structures, these models aim to generate new materials that follow similar distributions. Approaches include combining variational autoencoders (VAEs) with diffusion decoders [7, 16], using diffusion or flow-matching models that jointly model lattices and atomic positions [17, 18, 6], and applying classifier-free guidance to enable conditional generation and improved alignment with target properties [5]. Other directions include language models with domain-specific tokenization schemes [19], fine-tuned large language models (LLMs) [20], and hybrid methods that combine LLMs with flow matching [21].

Recent work has explored RL as a way to improve large language models by providing more reliable training signals. In mathematics, verifiable rewards—such as correctness of intermediate steps or final answers—have proven especially effective for guiding models toward consistent reasoning [22]. In parallel, separate efforts on preference alignment leverage human or proxy feedback to better match model outputs with user expectations [23].

3 Preliminaries

Crystal Representation A crystal structure can be described mathematically as $\mathcal{C}(L, A, X)$, where $L \in \mathbb{R}^{3\times3}$ is the lattice matrix defining the periodic unit cell, $Ai \in \mathbb{R}$ are the atomic species, and X_i are their fractional coordinates $\mathbf{X}_i \in [0, 1)^3$ within the unit cell. This representation uniquely specifies a periodic arrangement of atoms in three dimensions. Crystal structures can be stored in Crystallographic Information Files (CIFs), which encode the lattice parameters, atomic species, and atomic coordinates in a standardized text format.

Stability of Materials Stability is commonly evaluated using the *convex hull* of formation energies, which defines the phase that is energetically favorable relative to all competing phases. A material's stability is quantified by its energy above the hull, E_{hull}. Materials with $E_{hull} = 0$ eV/atom lie exactly on the hull and are considered stable, those with small positive values are metastable and may be synthesizable, while larger values indicate a stronger tendency to decompose.

4 Method

Tokenization. Tokenization of the CIF string is done by *byte pair encoding* (BPE) as suggested in [20], a compression method that assigns tokens to common substrings, making overall sequence lengths shorter [24].

Model. We build on the Qwen family of large language models (LLMs) [25–27], using Qwen2.5-7B-Instruct as our base model. Qwen2.5 is a general-purpose transformer trained on web-scale corpora for natural language and programming tasks, with the instruct variant further tuned to follow natural language prompts. In our setup, the model is prompted with a chemical formula (see A) and generates candidate crystal structures in text form.

Reinforcement Learning Finetuning. To bias the model towards generating stable materials, we apply RL finetuning. We use Group Relative Policy Optimization (GRPO) [28], a variant of PPO [29]. We define the state (i.e., model inputs) x as the prompt containing a chemical formula of atom

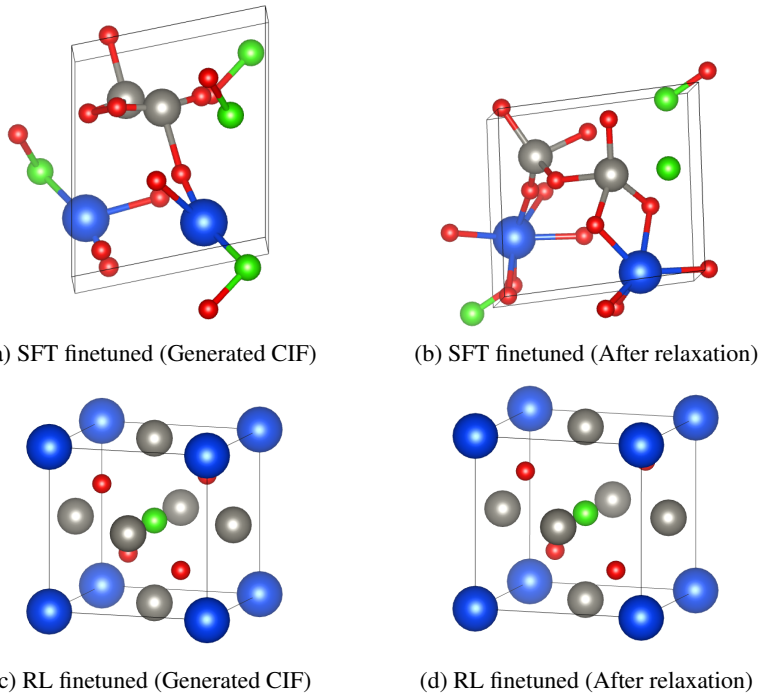
composition, and then the output y as the bulk representation of the crystal structure of that formula. RL learns a policy π that take in x and generate y that maximize a given reward R. We will explain the setting of reward in Section 5. PPO maximizes a clipped objective:

$$L_{PPO} = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)], \tag{1}$$

where $r(\theta) = \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)}$ is the policy ratio that measures how likely the new policy π_θ is to take the same action compared to the old policy $\pi_{\theta_{\text{old}}}$, and A_t is the advantage estimated by a value function. The clipping ensures $r_t(\theta)$ does not stray far from 1, thereby constraining updates so that the policy improves steadily without collapsing. GRPO eliminates the need for a value function by defining relative advantages within a group of candidates:

$$A_i = R_i - \frac{1}{n}\sum_{j=1}^{n} R_j \tag{2}$$

where R_i is the reward of candidate i and the baseline is the group mean. This formulation is simpler and more computationally efficient while maintaining stability. In our setting, Qwen2.5-7B-Instruct acts as the policy, candidate crystal structures are sampled, and CHGNet [13] provides the reward signal based on predicted potential energies, directly aligning the LLM with the goal of producing low-energy stable structures. Also, we give the model a penalty of -0.1 when it generates an invalid crystal structure (e.g., parsing error).



(a) SFT finetuned (Generated CIF)

(b) SFT finetuned (After relaxation)

(c) RL finetuned (Generated CIF)

(d) RL finetuned (After relaxation)

Figure 1: Comparison of generated CIF structures (left) and their UMA relaxed counterparts (right) for the models finetuned via SFT and via RL.

5 Experiments

We compare our RL-finetuned model against three baselines. First, the supervised fine-tuned (SFT) Qwen2.5-7B-Instruct, which reflects the performance of a general-purpose instruction-tuned LLM with supervised domain-specific alignment. Second, CDVAE [7], a generative model explicitly designed for crystal structure generation using a VAE with a diffusion-based decoder. Together, these baselines enable us to disentangle the contributions of RL finetuning LLMs, alternative LLM architectures, and domain-specific generative models.

5.1 Stability

Metric. To assess thermodynamic stability, we first relax the generated structures using UMA [15], a large pretrained MLIP, and obtain their total energies. The **energy above the convex hull**, E_{hull}, is then calculated by comparing these energies to reference values from the Materials Project database [30]. Based on this metric, we define the **Stability Rate** as the fraction of generated (and subsequently relaxed) structures with $E_{\text{hull}} < 0.1$ eV/atom. Additionally, we report the **Match Rate**, which quantifies how often a generated CIF preserves its structure after relaxation, and the **RMSD**, which measures the root-mean-square deviation in atomic positions between each generated structure and its relaxed counterpart. Stability metrics rely on structural relaxation. While QM simulations would provide the most accurate relaxations, they are very expensive; we therefore use UMA as a computationally efficient proxy.

Results. Table 1 reports the stability of generated materials across methods. Surprisingly, RL alone does not substantially improve stability, raising the rate only from 5.03 to 7.06 out of 7.6. The SFT model trained on ground-truth crystal structures achieves higher stability than RL, though it performs worse on average CHGNet energy predictions. This suggests that RL exploits weaknesses in the CHGNet reward model—maximizing its scores without producing stable materials when evaluated by UMA, a newer energy predictor. Interestingly, RL achieves a high match rate, indicating its generated structures closely resemble relaxed structures. This implies RL tends to produce local minima in the energy landscape. Since deeper minima correspond to more stable states, generating structures already near stable or metastable configurations both better captures the data distribution and reduces the cost of relaxation. Combining SFT and RL—first training with SFT, then applying RL—improves both stability and match rate relative to the base and SFT models. This shows that SFT can regularize reward hacking, suggesting future work should integrate both ground-truth data and learned energy predictors.

Table 1: Stability Evaluation

Model	E_{hull} [eV/atom] ↓	Stability Rate (%)† ↑	Match Rate (%) ↑	RMSD (Å) ↓
CDVAE [7]	0.376 ± 0.997	8	38.68	0.083
Qwen2.5	3.429 ± 10.958	5.03	31.02	0.052
Qwen2.5 SFT	0.374 ± 1.792	26.5	56.92	0.044
Qwen2.5 RL	0.629 ± 0.493	7.6	**84.40**	**0.025**
Qwen2.5 SFT+RL	0.3589 ± 1.2630	**27.29**	77.59	0.037

† $E_{hull} < 0.1$ [eV/atom].

5.2 Novelty and diversity

Metric. We evaluate generated crystal structures using complementary metrics. Validity has two components: structural validity (no overlapping atoms; interatomic distances must exceed half the sum of covalent radii) and compositional validity (net charge must be zero). Diversity measures variability across the set via pairwise distances in Matminer feature space. Novelty quantifies distinctness from the training data using the nearest-neighbor distance. Coverage is assessed with precision and recall, capturing how well the generated set reproduces the test distribution. Distances for novelty and coverage are computed using Matminer features [31]. All metrics are calculated directly on raw model outputs, without relaxation or post-processing.

Results. Table 2 summarizes the basic evaluation of generated structures. All methods achieve near-perfect structural validity, while compositional validity varies more widely: SFT ensures 100% validity, whereas RL tends to reduce compositional validity. In terms of coverage, CDVAE and SFT perform best, while RL alone shows weaker recall. RL, however, improves diversity—particularly in composition space—surpassing both CDVAE and SFT. Novelty shows a trade-off: CDVAE produces the most distinct structures, while SFT reduces novelty, and RL partially recovers it. Overall, SFT excels in validity and coverage, whereas RL enhances diversity and novelty at the cost of compositional validity.

Table 2: Validity and Novelty Evaluation

Method	Validity (%) ↑		Coverage ↑		Diversity ↑		Novelty ↑	
	Structure	Composition	Recall	Precision	Structure	Composition	Structure	Composition
CDVAE [7]	0.999	0.840	0.992	0.992	0.690	14.543	0.926	0.763
Qwen2.5 [26]	0.994	0.848	0.849	0.955	0.922	15.659	0.737	0.495
Qwen2.5 SFT	1.000	1.000	0.992	0.995	0.912	15.611	0.563	0.338
Qwen2.5 RL	1.000	0.740	0.683	0.992	1.203	16.250	0.401	0.609
Qwen2.5 SFT+RL	1.000	0.754	0.691	0.979	0.968	16.680	0.568	0.670

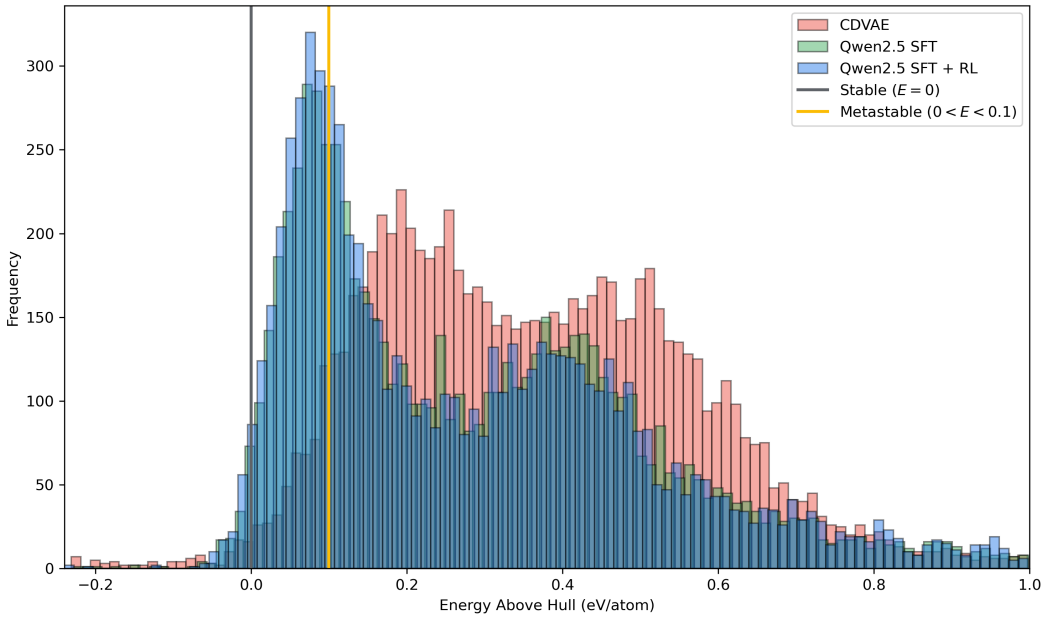


Figure 2: Energy above hull distribution of samples from CDVAE, Qwen2.5-7b-SFT finetuned, and Qwen2.5-7b-SFT finetuned and then RL finetuned for 40 steps.

6 Discussion

Table 1 highlights the differences between supervised finetuning (SFT) and reinforcement learning (RL) finetuning. SFT yields generated structures with a lower match rate to their relaxed states, and need to undergo substantial relaxation before reaching an energy minimum. In contrast, RL finetuning leads to a higher match rate, as structures are generated closer to relaxed geometries. Examples in Figures 1,3 illustrate this contrast, as RL-finetuned models produce symmetric and energetically favorable structures that change little upon relaxation, while SFT-generated structures deform significantly.

Despite this advantage, the stability rate of RL alone remains low, indicating that these "locally stable" configurations do not always correspond to the most thermodynamically favorable phases. Figure 2 shows that finetuned LLMs outperform CDVAE, a domain-specific baseline, and that combining SFT with RL produces the best results: modest improvement in stability rate and substantial gains in match rate compared with SFT alone. These findings suggest that RL finetuning effectively steers generation toward structurally consistent and energetically plausible candidates. However, relying solely on potential energy as the reward seems to be insufficient, and possibly enables hacking. The model may exploit this signal by narrowing the range of structures it generates, as reflected in the slightly reduced novelty observed for longer RL runs.

7 Future Work

A promising direction is to redefine the reward function. Instead of optimizing only for potential energy, incorporating the energy above hull directly would better align training with thermodynamic stability. To avoid reduced novelty, future work would also explore mechanisms that encourage structural diversity during finetuning, balancing stability with exploration and thereby enhancing the likelihood of discovering novel stable materials.

References

[1] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

[2] Martin A Green, Anita Ho-Baillie, and Henry J Snaith. The emergence of perovskite solar cells. *Nature photonics*, 8(7):506–514, 2014.

[3] Usman Shareef, Aisha Altaf, Madiha Ahmed, Nosheen Akhtar, Mohammed S Almuhayawi, Soad K Al Jaouni, Samy Selim, Mohamed A Abdelgawad, and Mohammed K Nagshabandi. A comprehensive review of discovery and development of drugs discovered from 2020–2022. *Saudi Pharmaceutical Journal*, 32(1):101913, 2024.

[4] Nofit Segal, Aviv Netanyahu, Kevin P Greenman, Pulkit Agrawal, and Rafael Gomez-Bombarelli. Known unknowns: Out-of-distribution property prediction in materials and molecules. *arXiv preprint arXiv:2502.05970*, 2025.

[5] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

[6] Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. *arXiv preprint arXiv:2406.04713*, 2024.

[7] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

[8] Yong Zhao, Edirisuriya M Dilanga Siriwardane, Zhenyao Wu, Nihang Fu, Mohammed Al-Fahdi, Ming Hu, and Jianjun Hu. Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Computational Materials*, 9(1):38, 2023.

[9] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

[10] Anubhav Jain, Yongwoo Shin, and Kristin A Persson. Computational predictions of energy materials using density functional theory. *Nature Reviews Materials*, 1(1):1–13, 2016.

[11] Mercedes Boronat, Antonio Leyva-Perez, and Avelino Corma. Theoretical and experimental insights into the origin of the catalytic activity of subnanometric gold clusters: attempts to predict reactivity with clusters and nanoparticles of gold. *Accounts of chemical research*, 47(3): 834–844, 2014.

[12] James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65(11):1501–1509, 2013.

[13] 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, 2023.

[14] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

[15] Brandon M Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R Kitchin, Daniel S Levine, et al. Uma: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

[16] Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965*, 2025.

[17] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36:17464–17497, 2023.

[18] Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. *arXiv preprint arXiv:2311.09235*, 2023.

[19] Daniel Flam-Shepherd and Alán Aspuru-Guzik. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv preprint arXiv:2305.05708*, 2023.

[20] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.

[21] Anuroop Sriram, Benjamin Miller, Ricky TQ Chen, and Brandon Wood. Flowllm: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.

[22] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[24] Philip Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

[25] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[26] A Yang Qwen, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, et al. Qwen2. 5 technical report. *arXiv preprint*, 2024.

[27] 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.

[28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[29] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[30] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[31] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

A Supplementary Material

In all our experiments, LLM prompts are in the following form:

Generation Prompt

You are a material scientist expert in crystal structure prediction. Your task is to predict the stable structure of a given chemical formula {formula}. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice. Format your answer as lattice lengths, lattice angles, then element symbols with coordinates:

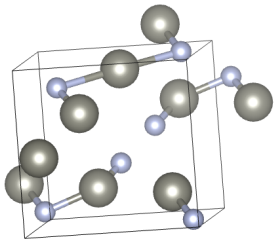
Example 1:

```
10.3 6.0 4.7
90 90 90
Li
0.25 0.50 0.75
Fe
0.75 0.50 0.25
P
0.50 0.00 0.50
O
0.10 0.60 0.40
O
0.90 0.40 0.60
O
0.40 0.90 0.10
O
0.60 0.10 0.90
```

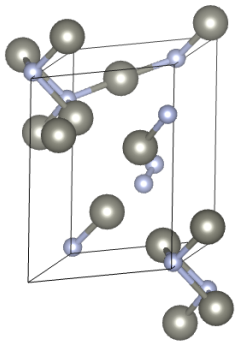
Example 2:

```
5.2 5.2 11.8
90 90 120
Mg
0.33 0.67 0.25
Al
0.00 0.00 0.00
O
0.31 0.69 0.38
O
0.69 0.31 0.62
```

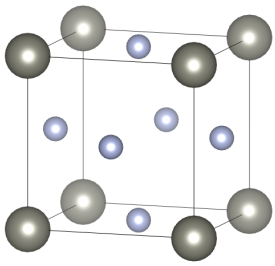
Provide ONLY the bulk representation like the example with no additional text.



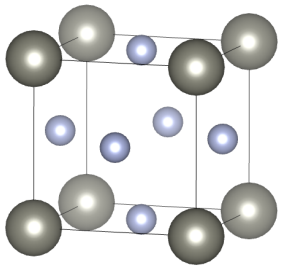
(a) SFT finetuned (Generated CIF)



(b) SFT finetuned (After relaxation)



(c) RL finetuned (Generated CIF)



(d) RL finetuned (After relaxation)

Figure 3: Comparison of generated CIF structures (left) and their UMA relaxed counterparts (right) for SFT and RL finetuned models.