Latent Crystallographic Microscope: Probing the Emergent Crystallographic Knowledge in Large Language Models

Jingru Gan[*]
UCLA

Yanqiao Zhu
UCLA

Wei Wang
UCLA

Abstract

Recent works are exploring the application of large language models to materials discovery, from property prediction to structure generation. However, the internal mechanisms through which LLMs perform crystallographic understanding and reasoning tasks remain unexplored. This lack of mechanistic understanding prevents the development of principled approaches for reliable materials discovery. We introduce the Latent Crystallography Microscope (LCM), a mechanistic interpretability framework for reverse-engineering crystallographic reasoning in large language models. We conduct three experiments mapping the progression from mechanistic understanding to controlled intervention. First, format recognition and property extraction tasks reveal that LLMs excel at direct metadata retrieval but struggle with geometric computations, indicating reliance on pattern matching over true geometric reasoning. Second, activation patching identifies task-specific neural circuits where attention heads mediate information routing while MLP blocks encode abstract crystallographic rules, with computational onset progressing to later layers as task complexity increases. Third, onset layer interventions during structure generation demonstrate that these mechanistic insights enable targeted neural modifications, though intervention effectiveness remains material-system dependent. Our analysis locates crystallographic computations to specific neural circuits, providing intervention targets for future work. This work maps the computational mechanisms underlying crystallographic tasks while demonstrating current limitations in leveraging these insights for reliable materials generation.[2]

1 Introduction

The emergence of Large Language Models (LLMs) as powerful tools for scientific discovery has revolutionized computational materials science, with recent breakthroughs demonstrating their ability to generate thermodynamically stable crystal structures [2, 10, 12]. These models can predict material properties with accuracy comparable to specialized graph neural networks [5, 31] and generate novel crystalline structures through autoregressive text modeling, functioning as "innate crystal structure generators" without explicit domain training [10]. However, despite these impressive capabilities, the internal mechanisms by which LLMs encode, process, and manipulate crystallographic knowledge remain fundamentally unclear, which represent a critical gap that limits both scientific understanding and practical control over materials discovery applications.

This gap constitutes a fundamental barrier to trustworthy AI-driven materials discovery. When LLMs generate crystal structures or demonstrate sophisticated understanding of space group symmetries

[*] Corresponding author: jrgan@cs.ucla.edu

[2] Code available at https://github.com/JingruG/LCM

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

and thermodynamic properties, we lack insight into the computational pathways enabling these behaviors. The lack of mechanistic interpretability prevents systematic debugging when models generate thermodynamically unstable structures, limits targeted interventions to guide generation toward specific material properties, and hinders the development of more controllable and reliable materials discovery systems.

Modern materials discovery faces a persistent stability-novelty dilemma, as comprehensively documented in recent benchmarking studies [24]. Traditional techniques like data-driven ion exchange achieve high thermodynamic stability rates (9.2%) by performing substitutions on known stable compounds, but exhibit zero prototype novelty because they rely exclusively on existing structural templates. Conversely, advanced generative models like Crystal Diffusion Variational Autoencoder (CDVAE) [32] and MatterGen [35] excel at structural innovation with prototype novelty rates up to 8.2%, but suffer from poor stability rates as most novel prototypes are thermodynamically unstable [24]. This trade-off presents researchers with a limiting choice: pursue stability at the expense of innovation, or chase novelty at the risk of instability.

Recent advances in mechanistic interpretability offer a promising path forward by developing sophisticated techniques for reverse-engineering neural network computational mechanisms into human-understandable components [8, 22]. Linear probing methods reveal how different types of knowledge are encoded across transformer layers [1, 7], while causal intervention techniques like activation patching enable direct manipulation of model behaviors [14, 18]. Sparse autoencoders decompose neural representations into interpretable features [4, 25], providing unprecedented insight into complex concept representation.

Within this broader agenda, we adopt and extend the concept of frequency-conditioned onset [13] where the model shifts from passively reading inputs and guessing to actively computing task-relevant structure. This transition is causally identifiable via intervention and its location depends on task complexity and the kind of computation involved. Prior work has shown hierarchical, layer-wise specialization [8, 26], causally responsible layers for factual recall and circuit behavior [18, 23, 29], and a division of labor where attention heads implement sequence operations while MLP blocks serve as key-value memories for stored knowledge [11]. We leverage these insights to measuring onset in crystallographic reasoning tasks, connecting attention-based geometric consistency to earlier onsets and MLP-driven parametric knowledge to later onsets.

However, these interpretability techniques have focused primarily on linguistic tasks, leaving scientific reasoning domains largely unexplored. Materials science presents unique interpretability challenges, requiring models to integrate discrete categorical variables (space groups), continuous geometric parameters (lattice constants), and emergent physical properties (formation energies). Understanding how LLMs process these diverse crystallographic concepts could unlock principled approaches to the stability-novelty dilemma through targeted neural interventions.

In this work, we introduce the **Latent Crystallography Microscope (LCM)** framework to understand crystallographic conception and reasoning in LLMs through a systematic progression from observation to control. Our investigation follows three key experiments that map the journey from understanding computational mechanisms to attempting practical control over materials discovery.

First, we evaluate format recognition and property extraction capabilities to understand what crystallographic information LLMs can reliably process. We test models on CIF and POSCAR formats across properties of varying computational complexity, revealing reliance on pattern matching rather than true geometric reasoning. Second, we employ activation patching to causally identify which neural circuits are responsible for different crystallographic computations, examining how attention heads and MLP blocks divide labor across coordinate parsing, stability assessment, and valence verification tasks. Third, we test whether mechanistic insights can enable targeted interventions by injecting stability vectors at identified onset layers during crystal structure generation.

Our analysis of Llama 3.1-70B reveals that crystallographic computations occur through task-specific neural circuits with distinct onset patterns. Format recognition experiments demonstrate that models excel at metadata retrieval from CIF files but fail at geometric computations from POSCAR coordinates. Activation patching identifies specialized circuits where attention heads handle information routing while MLP blocks show more tendency to encode abstract rules, with onset layers often appearing in the middle layers and progressing later as task complexity increases. Onset layer interventions during generation achieve limited success, with effectiveness varying significantly across material systems, highlighting the gap between mechanistic understanding and practical control.

This work represents the first comprehensive mechanistic analysis of crystallographic reasoning in LLMs. Our findings demonstrate that while we can identify and characterize the neural circuits underlying crystallographic tasks, translating these insights into reliable control remains challenging. The framework establishes a foundation for developing more interpretable AI systems in scientific domains while highlighting fundamental limitations in current intervention approaches.

2 Related Work

2.1 Mechanistic Interpretability of Large Language Models

Mechanistic interpretability seeks to reverse-engineer the computational mechanisms learned by neural networks into human-understandable algorithms and concepts [8, 22]. This field has developed sophisticated methodologies for understanding transformer architectures, moving beyond black-box analysis to provide granular, causal understanding of model behaviors. Linear probing has emerged as a fundamental technique for assessing what information is encoded within neural representations [1], with Conneau et al. [7] demonstrating that sentence embeddings contain rich linguistic structure accessible through linear classifiers. Subsequent work revealed hierarchical organization of syntactic knowledge across transformer layers [26], while recent advances have extended these insights to reveal linear structure in how models represent truth [17] and sentiment [27], supporting the hypothesis that many concepts are encoded as linear directions in activation space.

Understanding correlation between representations and concepts requires establishing causal relationships through direct intervention. Vig et al. [28] pioneered causal mediation analysis for language models, while Meng et al. [18] developed techniques for locating and editing specific factual associations in GPT models. Activation patching has emerged as a particularly powerful intervention technique [14], with recent work establishing best practices for applying these methods systematically [36]. A major breakthrough has been the development of sparse autoencoders (SAEs) for decomposing neural representations into interpretable features [4]. Building on theoretical foundations in superposition [9], SAEs address the fundamental challenge of polysemanticity by extracting monosemantic features from entangled neural activations. Recent scaling work has demonstrated that these techniques can be applied to production-scale models like Claude 3 Sonnet [25], revealing highly abstract features that capture sophisticated reasoning patterns. Understanding how features interact requires mapping the computational circuits that connect inputs to outputs. Olsson et al. [23] identified induction heads as a key mechanism for in-context learning, while Conmy et al. [6] developed automated techniques for discovering computational circuits, enabling detailed understanding of specific reasoning tasks [29].

2.2 Large Language Models in Materials Science

The application of LLMs to materials science has emerged as a transformative approach that challenges traditional computational methods. Unlike specialized architectures that require explicit geometric encodings, LLMs process crystallographic information through unified textual representations, offering unprecedented flexibility and generality. Before the emergence of LLMs, materials property prediction was dominated by graph neural network approaches. Crystal Graph Convolutional Neural Networks (CGCNNs) [31] provided the first systematic framework for learning from crystal structures, while MEGNet [5] established graph networks as a universal framework for molecular and crystalline systems. Recent scaling efforts have pushed these approaches to handle massive datasets [19], setting performance benchmarks for materials discovery applications.

The breakthrough application of LLMs to crystal generation began with CrystaLLM [2], which demonstrated that autoregressive modeling of Crystallographic Information Files (CIF) could generate thermodynamically plausible structures. Gruver et al. [12] demonstrated that fine-tuning pre-trained LLMs could generate stable inorganic materials with higher success rates than specialized diffusion models, achieving 49% vs 28% metastable generation rates compared to CDVAE baselines. Recently, MatLLMSearchGan et al. [10] has revealed the capabilities of pre-trained LLMs functioning as crystal structure generators without additional training, achieving high metastable rates through evolutionary search. Beyond direct structure generation, sophisticated frameworks have emerged that combine multiple AI components. GenMS [34] integrates language models with diffusion models and graph neural networks to enable hierarchical materials search from natural language descriptions. Technical innovations include invariant tokenization approaches [33] that ensure SE(3) and periodic invariance in crystal representations, and text-conditioned generation methods [20] that enable targeted design through natural language prompts.

2.3 Interpretability Challenges in Scientific Domains

While mechanistic interpretability has made significant progress in linguistic tasks, its application to scientific reasoning presents unique challenges that remain largely unaddressed. Recent systematic investigations have revealed fundamental limitations when interpretability methods are applied to scientific domains. In biology-inspired deep learning, systematic analysis has demonstrated that interpretations lack robustness upon repeated training and are systematically influenced by biases in knowledge graphs, with interpretation variability increasing with network depth and knowledge incompleteness creating spurious feature attributions [3]. Comprehensive reviews of biologically-informed models have found that interpretation reliability decreases significantly when models encounter data distributions outside their training domain, highlighting fundamental generalization challenges in scientific interpretability [30].

Existing analysis of explainable AI methods applied to regulatory genomics has revealed that commonly used attribution methods produce inconsistent explanations across model architectures and training procedures [21]. The evaluations show that gradient-based interpretations often fail to capture true causal relationships in biological systems, with explanation fidelity varying dramatically based on input representation format choices. Recent work in computational biology has identified specific failure modes where attention-based explanations highlight irrelevant regions due to models' reliance on long-range dependencies that attention mechanisms cannot reliably capture [3], while gradient-based interpretations systematically conflate input relevance with gradient magnitude [16].

The application of interpretability to scientific domains faces the fundamental challenge of validation against domain-specific ground truth [15]. Many interpretability techniques designed for natural data assume clear categorical distinctions that apply to continuous scientific phenomena. In materials science, this leads to the challenge of validating interpretations against quantum mechanical principles, crystallographic theory, and thermodynamic constraints. In these domains, ground truth may be computationally intractable or experimentally inaccessible. These challenges motivate the development of specialized interpretability frameworks that can handle the unique requirements of scientific reasoning domains while providing reliable insights into model behavior.

3 Experiment 1: Format and Property Conceptual Understanding

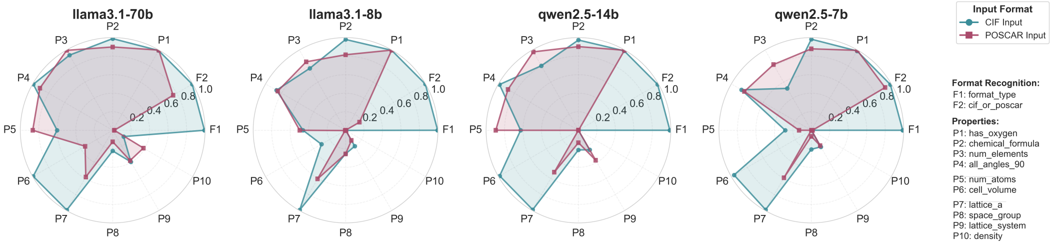


Figure 1: Performance across crystallographic format recognition tasks (F_n) and tiered property extraction tasks (P_n). Each task is evaluated with input format in CIF (blue) or POSCAR (red).

We evaluate whether LLMs can extract crystallographic information from two common formats: CIF (comprehensive metadata) and POSCAR (geometric coordinates). We test each LLM on 1000 crystal structures from Matbench v0.1 band gap prediction test dataset. We assess format recognition, tiered property extraction (10 properties across 3 complexity tiers), and cross-format consistency across four models: Llama 3.1-70B, Llama 3.1-8B, Qwen2.5-7B, and Qwen2.5-14B. A holistic assessment for format recognition and property extraction performance is provided in Figure 1, detailed task performance provided in Figure 5 (Appendix).

Format Recognition. We evaluate format recognition using two task formulations: F1 (open-ended format identification) and F2 (binary choice of CIF or POSCAR). Smaller models consistently default to CIF predictions regardless of input, indicating CIF dominance in pre-training data and reliance on general heuristics rather than actual format syntax analysis. This asymmetry suggests that format recognition depends on explicit label space constraints rather than learned format representations.

Property Extraction. We evaluate LLMs' conceptual understanding through extraction of ten crystallographic properties organized into three tiers of increasing computational complexity. **Tier 1** (simple lookup) includes directly accessible properties (`has_oxygen`, `chemical_formula`, `num_elements`, `all_angles_90`); **Tier 2** (basic computation) requires simple calculations (`num_atoms`, `cell_volume`); and **Tier 3** (domain knowledge) demands crystallographic expertise or format-dependent reasoning (`lattice_a`, `space_group`, `lattice_system`, `density`). The sharp accuracy decline from Tier 2 to Tier 3 reveals a fundamental transition from pattern matching to computational reasoning. Evaluation employs property-specific tolerances: $\pm1\%$ for lattice parameters, $\pm5\%$ for cell volume, and $\pm10\%$ for density.

LLMs consistently achieve higher accuracy with CIF than POSCAR inputs across all tiers, suggesting CIF's explicit metadata enables direct retrieval while POSCAR demands implicit geometric reasoning. Analysis of `lattice_a` extraction failures reveals models extract vector components rather than computing norms, confirming reliance on pattern matching over geometric computation. This finding motivates our subsequent focus on POSCAR, as it better isolates the computational mechanisms underlying geometric reasoning.

Mechanistic Implications. While we observe declining performance with increasing task complexity, the underlying neural mechanisms remain unclear. The performance gap of the two input formats raises questions about whether models truly compute geometric properties or merely retrieve memorized patterns. The consistent failure to compute lattice norms from POSCAR vectors suggests the latter, yet the mechanism behind even successful property extraction remains opaque. To investigate these mechanisms, we employ causal analysis to identify which model components are responsible for different types of crystallographic reasoning. We focus subsequent experiments on Llama 3.1-70B with POSCAR input, as POSCAR's minimal explicit metadata forces the model to engage computational pathways that would otherwise be bypassed through direct retrieval from CIF's rich annotations.

4 Experiment 2: Causal Analysis of Crystallographic Reasoning Circuits

Building on our finding that crystallographic tasks exhibit hierarchical computational complexity, we then investigate where and how these computations occur within the LLM. We propose two hypotheses on the emergence of crystallographic reasoning and employ activation patching to causally identify the underlying computational mechanisms.

First, we investigate the *onset layer* hypothesis: LLMs perform frequency-conditioned onset [13], where early layers generate statistical guesses that later layers refine into knowledge-informed answers. For crystallographic tasks, we hypothesize that the onset layer, where the model transitions from passive input processing to active property computation, will occur later for computationally complex tasks than for simple retrieval tasks.

In addition, we test the *circuit specialization* hypothesis: different computational tasks rely on distinct neural circuits within the transformer architecture. Specifically, we investigate whether attention heads primarily handle information routing (e.g., locating relevant atoms or coordinates) while MLPs perform abstract computations (e.g., calculating distances or applying symmetry rules). This distinction is crucial for understanding how models process geometric information versus symbolic metadata.

Our causal analysis follows the causal tracing and activation patching methodology established by Meng et al. [18], and aligns with earlier causal mediation perspectives on information flow in transformers [28]. Our goal is to identify the earliest layer range where restoring clean activations recovers correct crystallographic predictions.

4.1 Task Design and Activation Patching

We designed three tasks from concrete information retrieval to abstract reasoning, to isolate different types of crystallographic computation:

Coordinate Patching: Tests the LLM's ability to identify correct atomic positions from multiple geometric options. The task requires parsing spatial coordinates and comparing geometric relationships within the prompt context.

Stability Judge: Evaluates the LLM's capacity to distinguish between thermodynamically stable and strained crystal structures. This task demands integration of multiple structural features through cross-referencing and comparison.

Valence Verifier: Assesses whether the LLM can apply charge neutrality principles to determine if given ions can form stable compounds. This task minimizes information routing demands while maximizing reliance on parametric chemical knowledge.

For each task, we construct minimal contrast prompt pairs using same structural data and answer options. The clean prompts (e.g., "Which option is the correct, stable atomic position?") lead to correct responses and corrupted prompts (e.g., "Which of the following is the misplaced position?") produce incorrect answers. We evaluate on 1000 test cases per task: Coordinate Patching uses coordinate perturbation pairs from crystal structures, Stability Judge uses stable-strained structure pairs, and Valence Verifier uses charge-balanced and imbalanced ion combinations. Before patching, we validate that the clean prompt produces the correct answer and the corrupted prompt produces the expected incorrect answer; only validated cases proceed to activation patching analysis. We employ activation patching following the causal tracing framework [18]: we first cache internal activations at the decision token (final token position before answer generation) from the clean prompt for every layer, storing three types of activations (full residual stream h_l, attention output attn_l, and MLP output mlp_l). We then run the corrupted prompt, interrupting the forward pass at each layer l to overwrite the decision token activation with the corresponding cached clean activation, creating three patching conditions: Full Patching(entire residual stream), Attention Only, and MLP Only.

We monitor intervention effects by extracting logit scores for answer choice tokens at the decision token position. For each test case, we compute *clean logits* (from clean prompt) and *corrupted logits* (from corrupted prompt) for all answer choices. The *clean baseline logit difference* is $\text{clean\_logit\_diff} = \text{logit}(\text{correct\_choice}) - \text{logit}(\text{incorrect\_choice})$ from clean logits, and the *corrupted baseline logit difference* is $\text{corrupted\_logit\_diff} = \text{logit}(\text{correct\_choice}) - \text{logit}(\text{incorrect\_choice})$ from corrupted logits. The *baseline gap* is $\text{baseline\_gap} = \text{clean\_logit\_diff} - \text{corrupted\_logit\_diff}$, representing the maximum possible improvement from patching. After patching at layer l, we compute *patched logit difference* $\text{patched\_logit\_diff}_l$ and *logit difference improvement* $\Delta\text{logit\_diff}_l = \text{patched\_logit\_diff}_l - \text{corrupted\_logit\_diff}$. We normalize improvements by baseline gap: $\text{normalized\_improvement}_l = \Delta\text{logit\_diff}_l/\text{baseline\_gap}$, which represents the fraction of the gap recovered (0 = no improvement, 1.0 = full recovery). We identify the onset layer l^* as the earliest layer where patching causes the model's output choice to flip from the corrupted baseline to the correct answer. We assess patching effectiveness through success rates (percentage of cases where patching at any layer flips the answer) and normalized logit difference improvement, enabling component-level attribution of causal effects to attention mechanisms versus MLP circuits. Complete prompt templates are provided in Appendix A.2.

4.2 Onset Analysis

Onset Layer Progression: The onset layer progression is observed in Figure 2, where the most common onset layers are marked for different intervention types. For full patching, the onset layer progresses later as task complexity increases. The mean onset layers increase with task complexity: Coordinate Patching (33.2), Stability Judge (35.6), and Valence Verifier (36.1). This onset progression reflects the transformer's layer-wise abstraction hierarchy. Early layers establish geometric representations from raw coordinates and later layers apply abstract chemical constraints.

The Figure 3 shows the normalized improvement over patching layers $(\text{patched\_logit\_diff} - \text{corrupted\_logit\_diff})/\text{baseline\_gap}$, where $\text{baseline\_gap} = \text{clean\_logit\_diff} - \text{corrupted\_logit\_diff}$ represents the maximum possible recovery. We also observe consistent recovery pattern that all tasks recover approximately half of the baseline gap at onset layers and $\sim$100% at later layers. We

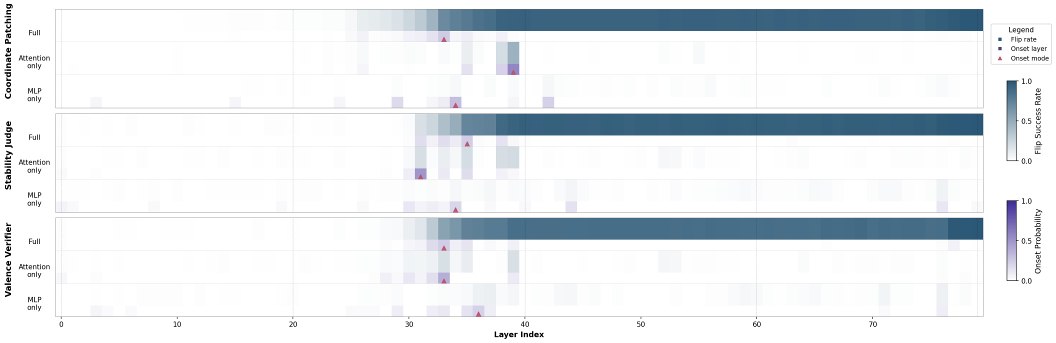


Figure 2: Progression of crystallographic reasoning onset, showing success rate (percentage) over layers.

also observe consistent mean normalized improvements across tasks (Coordinate: 53.0%, Stability: 54.0%, Valence: 53.5%).

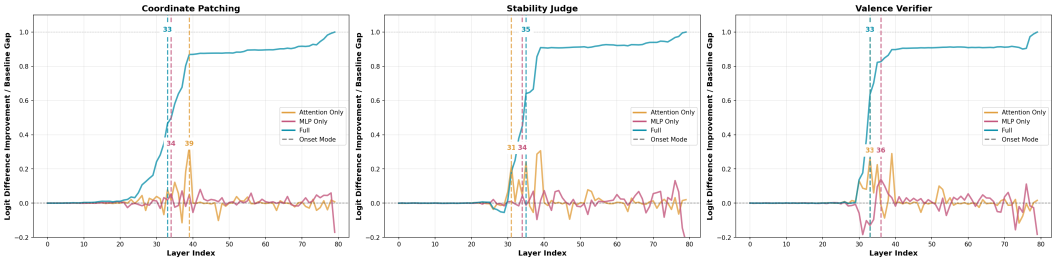


Figure 3: Layer-wise normalized logit difference improvement for three patching conditions: Full Patching (blue), Attention Only (yellow), MLP Only (red).

Circuit Specialization: As shown in Figure 2, for Coordinate Patching and Stability Judge tasks, Attention Only patching achieves high success rates (42.1% and 34.5% respectively) while MLP Only patching shows less causal effect, confirming that these tasks rely primarily on information routing and geometric comparison. In comparison, in the Valence Verifier experiment, MLP Only patching shows comparable success rate to Attention Only performance. When partial patching is applied, normalized improvements can be negative as shown in Figure 3, meaning patching makes performance worse than the corrupted baseline. It confirms that MLP and attention outputs are interdependent. The observed component specialization supports our hypothesis that information routing tasks require attention heads to move and compare geometric data across the prompt, while abstract reasoning tasks require MLPs to access parametric knowledge about chemical principles and apply mathematical rules like charge neutrality.

Our causal analysis suggests that crystallographic reasoning emerges through activation of specialized neural circuits. The division of labor observed between attention and MLP has profound implications for developing controllable crystal structure generation systems. For tasks requiring geometric optimization or structural comparison, interventions should target attention mechanisms that facilitate information integration across structural representations. For tasks requiring adherence to chemical principles or thermodynamic constraints, interventions should focus on MLP circuits that encode abstract domain knowledge. The onset progression further suggests that effective control strategies must respect the transformer's natural computational flow. Early interventions can influence geometric processing, middle-layer interventions can affect thermodynamic integration, and late-layer interventions can modify abstract chemical reasoning. These findings inform our subsequent intervention experiments, where we demonstrate how toleverage this mechanistic understanding to approach targeted control over specific crystallographic properties during structure generation.

5 Experiment 3: Crystal Structure Generation with Onset Layer Intervention

Our preceding experiments establish the computational foundations for controlled generation: hierarchical knowledge emergence (Experiment 1) and precise localization of crystallographic reasoning circuits (Experiment 2). We now conduct a preliminary test of whether mechanistic insights can enable limited manipulation of the generative process to influence crystal structure stability.

We designed a targeted intervention that extracts stability representations from the model and injects them during generation. For each unstable prototype (Parent A, $E_{\text{hull}} \geq 0.1$ eV/atom), we construct a stability vector from an ensemble of five stable reference structures (Parent B, $E_{\text{hull}} \leq 0.05$ eV/atom):

$$\vec{v}_{\text{stability}} = \frac{1}{N} \sum_{i=1}^{N} h_{l^*}(\text{Parent B}_i) - h_{l^*}(\text{Parent A})$$

During autoregressive generation, we modify the residual stream activation at layer 39, selected based on the onset analysis where the LLM recovers approximately 90% of the baseline gap:

$$h'_{l^*} = h_{l^*} + \alpha \cdot \vec{v}_{\text{stability}}$$

We conducted comprehensive experiments across 1,818 total interventions, testing intervention strengths $\alpha \in [0.0, 1.0, 2.0]$ and temperatures $T \in [0.0, 0.6]$ across three experimental runs targeting different compositional complexities (binary, ternary, and quaternary systems: ele2, ele3, ele4), where $\alpha = 0.0$ serves as the baseline control condition. This corresponds to 303 parent A structures (40, 145, and 118 for ele2, ele3, and ele4 respectively), each tested under 6 intervention conditions (3 alphas × 2 temperatures).

Table 1: Performance of Onset Layer Intervention on Stability. Stability improvement measured as reduction in E_{hull} from Parent A to generated child. % Novel & Stable indicates the percentage of generated structures that are both novel (not identical to parent) and thermodynamically stable ($E_{\text{hull}} \leq 0.05$ eV/atom). % Valid POSCARs indicates the percentage of generated text that successfully parses into valid crystal structures.

α	Temp.	Avg. Stability Improv. (eV/atom)	% Novel & Stable	% Valid POSCARs
0.0	0.0	-0.701 ± 1.620	4.5%	80.2%
0.0	0.6	-0.898 ± 1.731	3.3%	80.5%
1.0	0.0	-0.759 ± 1.711	4.5%	81.2%
1.0	0.6	-0.967 ± 1.720	3.6%	81.8%
2.0	0.0	-0.788 ± 1.783	**5.0%**	79.9%
2.0	0.6	-0.866 ± 1.414	3.8%	77.6%

Our analysis indicates measurable but limited intervention effectiveness, with 60 metastable results achieved in total. Interventions with $\alpha > 0$ yield improvements over baseline ($\alpha = 0$) conditions ($p < 0.05$), though practical effect sizes are modest. Results exhibit substantial case-to-case variability (±1.4–1.8 eV/atom), underscoring limited reliability and sensitivity to setup choices. We observe 80% valid POSCAR generation with successful parsing for valid structures; however, this does not imply general robustness of the overall intervention outcome.

Figure 4 illustrates intervention effectiveness in a ternary halide system where an unstable BaMnBr_3 structure ($E_{\text{hull}} = 0.193$ eV/atom) achieves substantial stability improvement. The stability vector was constructed from five stable Ba-Mn-Br reference compounds: two MnBr_2 variants ($E_{\text{hull}} = 0.015$ eV/atom), Mn_4Br_8 ($E_{\text{hull}} = 0.019$ eV/atom), Ba_4Br_8 ($E_{\text{hull}} = 0.019$ eV/atom), and Br_4 ($E_{\text{hull}} = -0.062$ eV/atom). This ensemble captures stability patterns across different Ba:Mn:Br coordination environments.

The intervention with $\alpha = 1.0, T = 0.0$ transforms the structure into $\text{Ba}_2\text{Mn}_2\text{Br}_8$ with $E_{\text{hull}} = 0.052$ eV/atom, achieving a stability improvement of 0.140 eV/atom. The transformation maintains the halide framework while adjusting stoichiometry from Ba:Mn:Br = 1:1:3 to 2:2:8, effectively doubling the formula unit and optimizing the Br coordination environment.

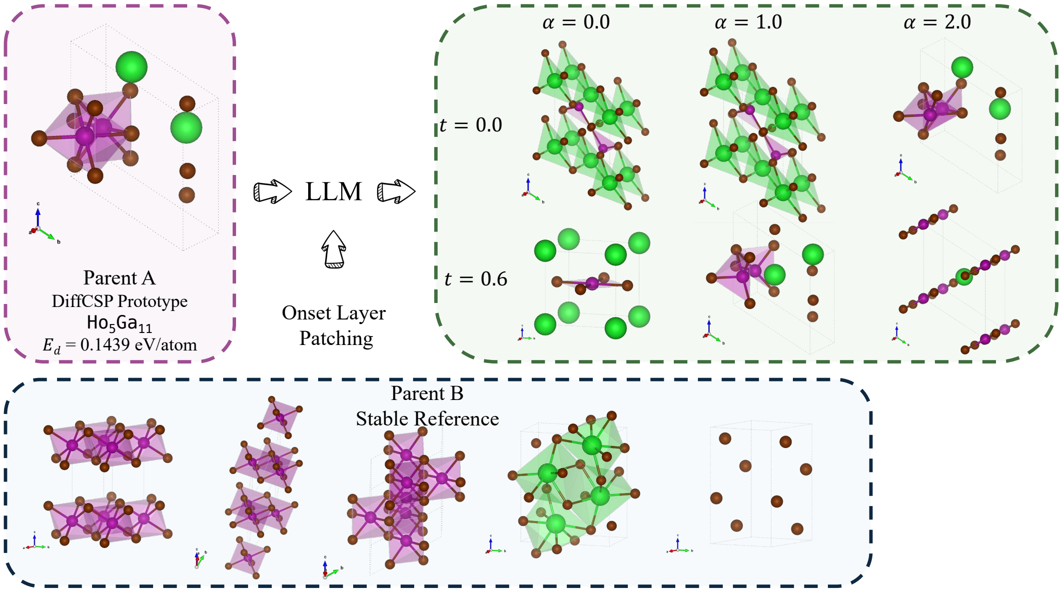


Figure 4: Intervention case study in a ternary halide system.

5.1 Parameter Optimization and Material System Analysis

Our analysis reveals distinct parameter sensitivities across different material systems. The best performing case achieved a stability improvement of 0.202 eV/atom ($CoF_4 \rightarrow Co_2F_8$ with $\alpha = 2.0, T = 0.6$), illustrating potential gains in isolated instances. Parameter effects show consistent trends in this dataset: $\alpha = 2.0, T = 0.0$ attains the highest observed stability rate (5.0%), while $\alpha = 1.0, T = 0.6$ yields the highest POSCAR generation success (81.8%).

Material system analysis reveals distinct success patterns across crystallographic families. Intermetallic systems show the highest stability rate (12.7%), followed by chalcogenides (5.9%), halides (3.4%), and oxides (2.9%). Nitrides show no stable results (0.0%), indicating system-dependent intervention effectiveness. Composition complexity analysis shows optimal performance for 3-element systems (4.9% stability rate), with decreasing effectiveness for simpler (2-element: 4.0%) or more complex (4-element: 3.4%) compositions.

5.2 Implications for Materials Discovery

This intervention results provide a preliminary proof-of-concept for influencing crystallographic properties during generation. It shows that the intervention can be effective in intermetallic and chalcogenide systems, suggesting potential applications in targeted discovery settings, although further validation is required to ensure synthesizability. The variability in result deformation energy highlights the limited reliability of current interventions and the need for careful parameter calibration. Success seems contingent on favorable alignment between target system chemistry and reference ensemble patterns.

6 Limitations and Conclusion

Our Latent Crystallography Microscope framework provides a first step toward mechanistic understanding of how large language models process crystallographic knowledge, revealing where and how LLMs tackle different crystallographic tasks. We show the capabilities of LLMs at different scales in conceptual understanding and crystallographic reasoning. Our activation patching experiments identify specialized circuits where attention heads route information while MLP blocks apply abstract crystallographic rules, with onset layers progressing later as task complexity increases. These mechanistic insights provide preliminary evidence of potential targeted layer-specific interventions on improving thermodynamic stability while preserving structural innovation. It provides a foundation for developing more controllable AI systems in materials science and broader scientific domains.

While our work demonstrates the mechanistic analysis of crystallographic reasoning in LLMs, several methodological limitations present valuable opportunities for future research. Our primary causal analysis focuses on a single model architecture (Llama 3.1-70B), limiting generalizability to other transformer architectures or model families. The intervention experiments involve simple compositions from DiffCSP prototypes, achieving limited success rates. In addition, our evaluation framework relies primarily on CHGNet metastability evaluation which may not capture all aspects of synthesizability. Our intervention methods operate at the residual stream level and cannot achieve feature-level precision, with success contingent on favorable alignment between target system chemistry and reference ensemble patterns. Despite these limitations, our work establishes mechanistic interpretability as a promising direction for controllable LLM-based scientific discovery, providing initial practical guidance while highlighting the challenges and opportunities for future research in developing more reliable and precise intervention methods.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR)*, 2017. 2, 3

[2] Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15:10570, 2024. 1, 3

[3] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x. 4

[4] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yuxin Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features. 2, 3

[5] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31 (9):3564–3572, May 2019. doi: 10.1021/acs.chemmater.9b01294. URL https://doi.org/10.1021/acs.chemmater.9b01294. 1, 3

[6] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, volume 36, pages 16318–16352. Curran Associates, Inc., 2023. 3

[7] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198/. 2, 3

[8] Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html. 2, 3

[9] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. 2022. URL https://arxiv.org/abs/2209.10652. 3

[10] Jingru Gan, Peichen Zhong, Yuanqi Du, Yanqiao Zhu, Chenru Duan, Haorui Wang, Daniel Schwalbe-Koda, Carla P. Gomes, Kristin A. Persson, and Wei Wang. Matllmsearch: Crystal structure discovery with evolution-guided large language models. 2025. URL https://arxiv.org/abs/2502.20933. 1, 3

[11] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495. Association for Computational Linguistics, nov 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/. 2

[12] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ward Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *International Conference on Learning Representations 2024*, 2024. 1, 3

[13] Akshat Gupta, Jay Yeung, Gopala Anumanchipalli, and Anna Ivanova. How do llms use their depth?, 2025. URL https://arxiv.org/abs/2510.18871. 2, 5

[14] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. 2024. URL https://arxiv.org/abs/2404.15255. 2, 3

[15] Frederick Klauschen, Jonas Dippel, Philipp Keyl, Philipp Jurmeister, Michael Bockmayr, et al. Toward explainable artificial intelligence for precision pathology. *Annu Rev Pathol*, 19:541–570, 2023. doi: 10.1146/annurev-pathmechdis-051222-113147. 4

[16] Antonio Majdandzic, Chandana Rajesh, and Peter K Koo. Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biology*, 24(1):109, 2023. doi: 10.1186/s13059-023-02956-3. 4

[17] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. URL https://openreview.net/forum?id=giMJzZIuzr. 3

[18] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262. 2, 3, 5, 6

[19] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. 3

[20] Trupti Mohanty, Maitrey Mehta, Hasan M Sayeed, Vivek Srikumar, and Taylor D Sparks. Crystext: A generative ai approach for text-conditioned crystal structure generation using llm. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2024-gjhpq. 4

[21] Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023. doi: 10.1038/s41576-022-00532-2. 4

[22] Christopher Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/mech-interp-essay. 2, 3

[23] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. In-context learning and induction heads. 2022. URL https://arxiv.org/abs/2209.11895. 2, 3

[24] Nathan J. Szymanski and Christopher J. Bartel. Establishing baselines for generative discovery of inorganic crystals. *Materials Horizons*, 2025. 2

[25] Adly Templeton, Trenton Bricken, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yuxin Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity`. 2, 3

[26] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=SJzSgnRcKX`. 2, 3

[27] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Language models linearly represent sentiment. 2024. URL `https://openreview.net/forum?id=Xsf6dOOMMc`. 3

[28] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020. 3, 5

[29] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=NpsVSN6o4ul`. 2, 3

[30] Magdalena Wysocka, Oskar Wysocki, Marie Zufferey, Dónal Landers, and André Freitas. A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. *BMC Bioinformatics*, 24(1):198, 2023. doi: 10.1186/s12859-023-05262-8. 4

[31] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018. doi: 10.1103/PhysRevLett.120.145301. URL `https://link.aps.org/doi/10.1103/PhysRevLett.120.145301`. 1, 3

[32] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 21961–21973, 2021. 2

[33] Keqiang Yan, Xiner Li, Hongyi Ling, Kenna Ashen, Carl Edwards, Raymundo Arróyave, Marinka Zitnik, Heng Ji, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Invariant tokenization of crystalline materials for language model enabled generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385. 4

[34] Sherry Yang, Simon Batzner, Ruiqi Gao, Muratahan Aykol, Alexander L Gaunt, Brendan McMorrow, Danilo Jimenez Rezende, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Generative hierarchical materials search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=PsPR4NOiRC`. 4

[35] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025. doi: 10.1038/s41586-025-08628-5. 2

[36] Jing Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. 3

A Appendix

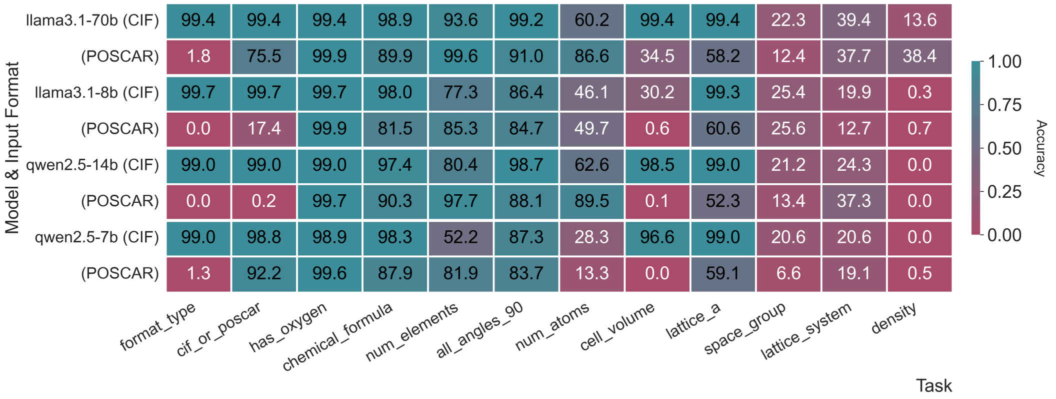
A.1 Format recognition and property extraction results

Format Recognition Tasks:
F1 (**format_type**): Open-ended question asking the model to identify the file format.
F2 (**cif_or_poscar**): Binary choice task asking whether the input is CIF or POSCAR format.

Property Extraction Tasks (Tier 1 - Simple Lookup):
P1 (**has_oxygen**): Determine if the structure contains oxygen atoms.
P2 (**chemical_formula**): Extract the chemical formula of the crystal.
P3 (**num_elements**): Count the number of unique element types.
P4 (**all_angles_90**): Determine if all lattice angles are 90°.

Property Extraction Tasks (Tier 2 - Lookup + Simple Calculation):
P5 (**num_atoms**): Count the total number of atoms in the unit cell.
P6 (**cell_volume**): Calculate the unit cell volume.

Property Extraction Tasks (Tier 3 - Requires Crystallographic Knowledge):
P7 (**lattice_a**): Extract the lattice parameter a (explicit in CIF, requires computing vector magnitude in POSCAR).
P8 (**space_group**): Identify the space group (explicit in CIF, requires symmetry analysis in POSCAR).
P9 (**lattice_system**): Classify the crystal system (cubic, tetragonal, etc.).
P10 (**density**): Calculate the material density (requires atomic mass lookup and volume).



Model & Input Format	format_type	cif_or_poscar	has_oxygen	chemical_formula	num_elements	all_angles_90	num_atoms	cell_volume	lattice_a	space_group	lattice_system	density
llama3.1-70b (CIF)	99.4	99.4	99.4	98.9	93.6	99.2	60.2	99.4	99.4	22.3	39.4	13.6
(POSCAR)	1.8	75.5	99.9	89.9	99.6	91.0	86.6	34.5	58.2	12.4	37.7	38.4
llama3.1-8b (CIF)	99.7	99.7	99.7	98.0	77.3	86.4	46.1	30.2	99.3	25.4	19.9	0.3
(POSCAR)	0.0	17.4	99.9	81.5	85.3	84.7	49.7	0.6	60.6	25.6	12.7	0.7
qwen2.5-14b (CIF)	99.0	99.0	99.0	97.4	80.4	98.7	62.6	98.5	99.0	21.2	24.3	0.0
(POSCAR)	0.0	0.2	99.7	90.3	97.7	88.1	89.5	0.1	52.3	13.4	37.3	0.0
qwen2.5-7b (CIF)	99.0	98.8	98.9	98.3	52.2	87.3	28.3	96.6	99.0	20.6	20.6	0.0
(POSCAR)	1.3	92.2	99.6	87.9	81.9	83.7	13.3	0.0	59.1	6.6	19.1	0.5

Figure 5: Accuracy of format recognition and property extraction tasks. Ten crystallographic properties tested with CIF or POSCAR input in four models: Llama 3.1-70B, Llama 3.1-8B, Qwen2.5-14B, and Qwen2.5-7B.

A.2 Experimental Prompts

A.2.1 Format Recognition and Property Extraction

All prompts follow a common template structure. The full prompt format is:

```
You are a crystallography expert. Analyze the following
crystal structure data and answer the question precisely.
Crystal Structure Data:
[CIF or POSCAR structure content]
Question: [question]
Answer:
```

Format Recognition Prompts:

F1 (format_type):

```
What file format is this crystal structure data in?
```

F2 (cif_or_poscar):

```
Is this structure data in CIF or POSCAR format?
```

Property Extraction Prompts:

P1 (has_oxygen):

```
Does this structure contain oxygen atoms?  Answer yes or no.
```

P2 (chemical_formula):

```
What is the chemical formula of this crystal structure?
```

P3 (num_elements):

```
How many different types of elements are in this structure?
```

P4 (all_angles_90):

```
Are all lattice angles equal to 90 degrees?  Answer yes or no.
```

P5 (num_atoms):

```
How many atoms are in the unit cell?
```

P6 (cell_volume):

```
What is the unit cell volume in cubic Angstroms?
```

P7 (lattice_a):

```
What is the lattice parameter 'a' in Angstroms?
```

P8 (space_group):

```
What is the space group of this crystal structure?
```

P9 (lattice_system):

```
What crystal system does this structure belong to?
```

P10 (density):

```
What is the density of this crystal in g/cm³?
```

A.2.2 Coordinate Patching

Clean Prompt (Coordinate Healing):

```
[INST]
You are an expert crystallographer.  The following POSCAR is
unstable because the [element] atom is misplaced.
POSCAR:
[Unstable POSCAR content]
The misplaced [element] atom is at ([x] [y] [z]).  Which
option is the correct, stable atomic position?
A) [distractor1_x] [distractor1_y] [distractor1_z]
B) [correct_x] [correct_y] [correct_z]
C) [distractor2_x] [distractor2_y] [distractor2_z]
D) [misplaced_x] [misplaced_y] [misplaced_z]
```

```
Output ONLY the SINGLE LETTER of the correct option.
[/INST]
Answer:
```

Corrupted Prompt (Coordinate Copying):

```
[INST]
You are an expert crystallographer.  The following POSCAR is
unstable because the [element] atom is misplaced.
POSCAR:
[Unstable POSCAR content]
The misplaced [element] atom is at ([x] [y] [z]).  Which of
the following is the misplaced position?
A) [distractor1_x] [distractor1_y] [distractor1_z]
B) [correct_x] [correct_y] [correct_z]
C) [distractor2_x] [distractor2_y] [distractor2_z]
D) [misplaced_x] [misplaced_y] [misplaced_z]
Output ONLY the SINGLE LETTER of the correct option.
[/INST]
Answer:
```

A.2.3 Stability Judge

Clean Prompt (Stable in Position A):

```
[INST]
You are an expert crystallographer evaluating crystal
stability.  Compare these two crystal structures and determine
which is more thermodynamically stable.
Structure A:
[Stable POSCAR content]
Structure B:
[Strained POSCAR content]
Based on crystallographic principles, which structure is more
thermodynamically stable?
Answer with SINGLE LETTER either ‘A’ or ‘B’.
[/INST]
Answer:
```

Corrupted Prompt (Stable in Position B):

```
[INST]
You are an expert crystallographer evaluating crystal
stability.  Compare these two crystal structures and determine
which is more thermodynamically stable.
Structure A:
[Strained POSCAR content]
Structure B:
[Stable POSCAR content]
Based on crystallographic principles, which structure is more
thermodynamically stable?
Answer with SINGLE LETTER either ‘A’ or ‘B’.
[/INST]
Answer:
```

A.2.4 Valence Verifier

Clean Prompt (Charge Neutral):

```
[INST]
You are an expert chemist.  Consider a chemical system
containing the following ions:
[charge-neutral ion list, e.g., “1x Sr2+, 1x Ti4+, 3x O2-”]
Based on the principle of charge neutrality, can these ions
form a stable, neutral compound?
A) Yes
B) No
Output ONLY the SINGLE LETTER of the correct option.
[/INST]
Answer:
```

Corrupted Prompt (Charge Imbalanced):

```
[INST]
You are an expert chemist.  Consider a chemical system
containing the following ions:
[charge-imbalanced ion list, e.g., “1x Sr2+, 1x Ti4+, 2x O2-”]
Based on the principle of charge neutrality, can these ions
form a stable, neutral compound?
A) Yes
B) No
Output ONLY the SINGLE LETTER of the correct option.
[/INST]
Answer:
```

A.2.5 Onset Layer Intervention

Generation Prompt:

```
Generate a new, more thermodynamically stable crystal
structure by improving the following unstable crystal
structure for [composition].
‘‘‘poscar
[Unstable Parent A POSCAR content]
‘‘‘
REQUIREMENTS:
- Primary goal:  Lower deformation energy
- Allow compositional changes for better stability
- Avoid copying parent coordinates
- Ensure valid coordination and no overlapping atoms
OUTPUT ONLY the POSCAR string:
```