Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI and Social Good

Mat-Instructions: A Large-Scale Inorganic Material Instruction Dataset for Large Language Models

Ke Liu[*1,2], **Shangde Gao**[1], **Yichao Fu**[1], **Xiaoliang Wu**[2], **Shuo Tong**[1], **Ajitha Rajan**[2], **Hao Xu**[*3]
[1]Zhejiang University
[2]The University of Edinburgh
[3]Harvard University
lk2017@zju.edu.cn, haxu@bwh.harvard.edu

Abstract

Recent advancements in large language models (LLMs) have revolutionized research discovery across various scientific disciplines, including materials science. The discovery of novel materials, particularly crystal materials, is essential for achieving sustainable development goals (SDGs), as they drive breakthroughs in climate change mitigation, clean and affordable energy, and the promotion of industrial innovation. However, unlocking the full potential of LLMs in materials research remains challenging due to the lack of high-quality, diverse, and instruction-based datasets. Such datasets are crucial for guiding these models in understanding and predicting the structure, property, and function of materials across various tasks. To address this limitation, we introduce Mat-Instruction, a large-scale inorganic material instruction dataset, specifically designed to unlock the potential of LLMs in materials science. Extensive experiments on fine-tuning LLaMA with our Mat-Instruction dataset demonstrate its effectiveness in advancing progress for materials science. The code and dataset are available at https://github.com/zjuKeLiu/Mat-Instructions.

1 Introduction

Recent advancements in Large Language Models (LLMs), including GPT-4 [Achiam *et al.*, 2023], LLaMA [Touvron *et al.*, 2023], and DeepSeek [Liu *et al.*, 2024] have significantly transformed the field of Natural Language Processing (NLP). These models, characterized by their vast parameter counts, are trained on extensive text corpora and excel in generating human-like text and understanding complex contexts. To adapt these general-purpose models for specific tasks, researchers have adopted instruction tuning techniques [Ouyang *et al.*, 2022; Sanh *et al.*, 2022], which involve training LLMs on specialized instruction datasets to enhance their performance in targeted domains.

Several instruction datasets have been developed to boost the efficiency of LLMs in general-purpose applications. The Stanford Alpaca dataset [Taori *et al.*, 2023] provides prompts and annotations for controllable text generation, while the GPT4All dataset [Anand *et al.*, 2023] encompasses diverse formats such as code, stories, and dialogue for training and evaluating general-purpose language models. Similarly, the COIG dataset [Zhang *et al.*, 2023] integrates multiple corpora, including translated texts, exam questions, and human value alignment instructions, specifically designed for Chinese language processing.

The success of LLMs in traditional text processing or general-purpose applications is largely attributed to the availability of extensive, high-quality instruction datasets. Despite the initial success of LLMs [Antunes *et al.*, 2024; Merchant *et al.*, 2023] in specific materials science problems such as crystal structure prediction (CSP), property analysis, and novel material discovery, their full potential in materials science is still limited by the lack of dedicated, domain-specific instruction datasets. This limitation arises from three primary challenges: (1) High Cost of Data Acquisition and Annotation: Material data is inherently complex and information-rich, making its collection and annotation resource-intensive. (2) Interdisciplinary Knowledge Requirements: Materials science integrates insights from crystallography, computational materials science, and materials engineering, necessitating a broad and specialized knowledge base. (3) Lack of Standardized Representations: Unlike NLP, materials science lacks a unified framework for representing materials and their computational properties, complicating the development of a universally applicable dataset.

To address these challenges, we introduce **Mat-Instruction**, a comprehensive instruction dataset to meet the unique requirements of LLM for crystal materials science. Mat-Instruction is structured around six core components:

- **Crystal Structure Prediction Instructions**: Focused on the properties and behaviors of crystal materials, this component addresses fundamental challenges in Crystal Structure Prediction (CSP) and material design.
- **Property Prediction Instructions**: Designed for predicting material properties, this component supports tasks related to structure and property prediction, facilitating data-driven material design.
- **Description-guided Crystal Design Instructions**: Tailored for NLP tasks in materials informatics, this com-

*Corresponding authors.

ponent includes information extraction and description-guided crystal structure design.

- **Crystal Reaction Instructions**: Centered on chemical reactions with crystal materials, this component provides guidance for predicting reaction products.
- **Crystal Retrosynthesis Instructions**: Dedicated to synthesis processes, this component offers instructions for designing synthetic routes, selecting reaction conditions, and anticipating outcomes for crystalline compounds.
- **Crystal Description Instructions**: Aimed at interpreting material data, this component provides instructions for describing the properties, structures, applications, and behaviors of crystal materials.

The development of Mat-Instruction involved curating material data from licensed sources and transforming them into task-specific instruction formats. By equipping LLMs with material-specific knowledge, Mat-Instruction enhances their capability to understand and predict material structures, properties, and functions, thereby revolutionizing the interpretation and discovery of material data.

To evaluate the effectiveness of Mat-Instruction, we conducted extensive experiments using a representative LLM as the base model. Instruction tuning was performed across the six components of instructions, and the results demonstrate that Mat-Instruction significantly improves the versatility and comprehension of LLMs in materials science. This work lays the foundation for novel scientific discoveries by enabling LLMs to understand and analyze crystal material science data with outstanding effectiveness.

2 Related Work

2.1 Existing Material Science Datasets

The emergence of datasets in materials science has provided critical support for data-driven research, driving innovations in material discovery, property prediction, and multi-scale modeling. Established repositories such as the NOMAD Repository integrate quantum mechanical calculations, experimental data, and machine learning models, providing multidimensional coverage of crystal structures and electronic properties [Miret and Krishnan, 2024]. The Materials Genome Initiative (MGI) Database accelerates material discovery by aggregating global research outputs and enabling high-throughput computational and experimental data storage [de Pablo *et al.*, 2019]. Traditional databases like MatWeb [Gao *et al.*, 2013], which focus on material performance parameters, are increasingly being supplanted by online platforms such as Materials Project [Jain *et al.*, 2013] and AFLOWlib [Curtarolo *et al.*, 2012], which automate data collection and analysis through integrated computational tools.

In contrast to traditional databases, newly emerged datasets such as LLM4Mat-Bench [Rubungo *et al.*, 2024] and OMat24 [Barroso-Luque *et al.*, 2024] further expand the boundaries for material science. LLM4Mat-Bench contains 1.97 million crystal structure entries spanning 45 material properties, supporting multimodal inputs (e.g., CIF files and textual descriptions) for task-specific predictive modeling. OMat24 aggregates 110 million Density Functional Theory (DFT) calculations with formation energy errors as low as 20 meV/atom, establishing a high-precision benchmark for stability prediction. However, existing datasets still face challenges such as data heterogeneity (e.g., inconsistent formats and units) and insufficient integration of multimodal representations (e.g., text, diagrams, and code).

2.2 Instruction Tuning in Materials Science

Instruction tuning has emerged as a pivotal approach for enhancing the reasoning capabilities of large language models (LLMs) in scientific domains. In chemistry, instruction datasets such as SMolInstruct and ChemLLMBench have demonstrated remarkable efficacy in molecular design and reaction prediction tasks, significantly improving model performance in chemical reasoning [Choi and Lee, 2024]. HoneyBee, trained on the MatSci-Instruct dataset, pioneers the specialization of billion-parameter LLMs for materials science. Its innovative iterative instruction generation and validation mechanism has achieved state-of-the-art accuracy in material property analysis [Song *et al.*, 2023]. However, they do not include the crystal structure in their work. General-purpose LLMs also exhibit promising performance in material instruction datasets via few-shot learning. For instance, GPT-4 achieves a classification accuracy of 96.1% in the battery material classification task with a small number of training samples. Despite the progress, existing instruction datasets (e.g., Stanford Alpaca [Taori *et al.*, 2023] and COIG [Zhang *et al.*, 2023]) primarily address general NLP tasks, lacking the fine-grained knowledge for material science domain-specific tasks, especially crystal structure-related problems. This limitation is particularly evident in handling complex scenarios involving symbolic diversity (e.g., multiple crystal structure representations) and cross-modal alignment (e.g., linking Crystallographic Information File (CIF) files to textual descriptions). This underscores the need for domain-specific instruction datasets like Mat-Instruction, proposed in this work.

2.3 Challenges in Material Science LLMs

The challenges of LLM for materials science applications are summarized below:

- **Data Credibility and Integration**: Material data originates from diverse sources (e.g., experiments, simulations, literature), necessitating standardized tools like Robocrystallographer to convert CIF files into parseable text and quality control mechanisms (e.g., expert validation).
- **Deep Domain Knowledge**: Material science requires interdisciplinary expertise (e.g., crystallography, computational chemistry), yet LLMs struggle with complex terms (e.g., Wyckoff positions) and unit conversions (e.g., Å vs. nm). [Miret and Krishnan, 2024] reveal error rates exceeding 60% in numerical tasks and 3D symmetry interpretation.
- **Multimodal and Open-Access Limitations**: Material data often combines text, diagrams, and videos, but high-quality datasets are frequently locked behind paywalls (e.g., closed-source journals). Open platforms like arXiv require extensive data cleaning and annotation.

Recent works attempted to address these issues through retrieval-augmented generation (RAG) and integration with

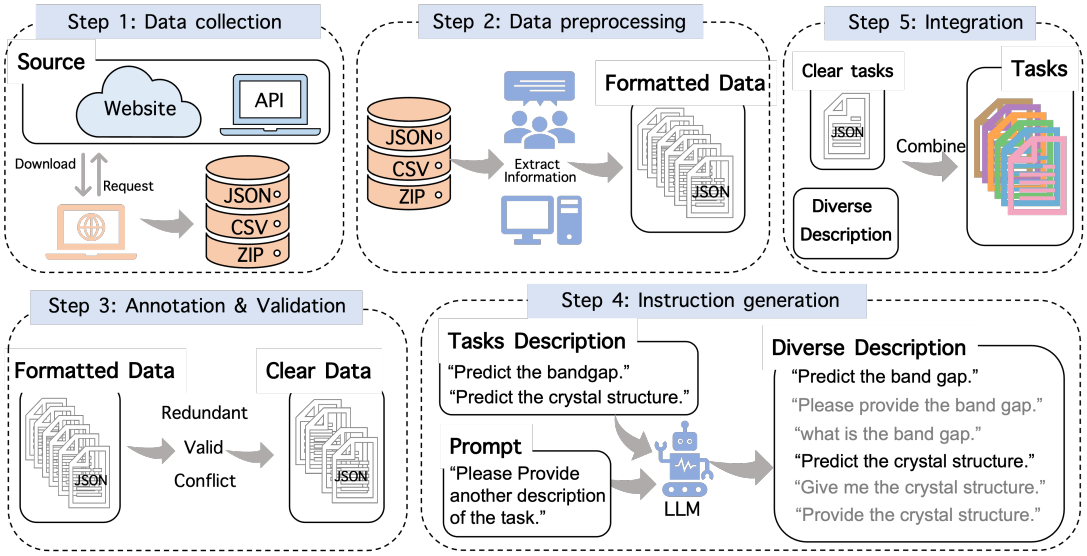

Figure 1: Overview of the Mat-Instruction dataset construction process, including the data collection, data preprocessing, annotation & validation, instruction generation, and integration.

Materials Project databases, enabling end-to-end material design [Miret and Krishnan, 2024]. However, challenges such as hallucinated references and incomplete external knowledge integration persist.

3 Mat-Instruction Construction

We construct Mat-Instruction by integrating data from multiple sources and transforming it into instruction-based formats. Different from [Song *et al.*, 2023], our Mat-Instruction consists of structure and description-related tasks. Structure-related tasks include crystal structure prediction, property prediction, and description-guided crystal design since properties of crystals are determined by their structures and these are important challenges in crystal material science [Tilley, 2020]. Description-related tasks consist of crystal reaction, crystal retrosynthesis, and crystal description prediction, which are tailored to synthesize desired materials. These tasks are essential for understanding and predicting the function, structure, and property of materials across various domains. Then we construct our Mat-Instruction dataset for these tasks by following steps: data collection, data preprocessing, instruction generation, annotation, and validation, as well as integration as shown in fig. 1. By following these steps, we create a robust and comprehensive dataset that empowers LLMs to excel in material science tasks, driving innovation and discovery in the field.

3.1 Data Collection

We collect data from various licensed sources, including Materials Project [Jain *et al.*, 2013], JDFT [Choudhary *et al.*, 2020], MatKG [Venugopal and Olivetti, 2024], Synthesis [Wang *et al.*, 2022], NERRE [Dagdelen *et al.*, 2024], and MGED-KG [Zhang *et al.*, 2024]. Each source provides unique insights into different aspects of material science, such as crystal structures, properties, and synthesis pathways. These data are carefully processed to ensure relevance and quality, allowing LLM to learn from diverse sources of information. All data sources are licensed under Creative Commons or GNU General Public License to ensure open access and sharing of data. Throughout the data collection process, we adhered to strict license agreements and ethical guidelines to ensure that the data was used in a responsible and ethical manner. To access the data, we mainly download the data from the official websites of these datasets or through the Accessible Programming Interfaces (APIs) provided by the dataset providers, as shown in fig. 1 Step 1.

3.2 Data Preprocessing

The collected data undergoes preprocessing to ensure consistency and quality as shown in fig. 1 Step 2. The row data from different sources may have different formats and units, which need to be standardized. This includes cleaning, normalization, and transformation into a unified format suitable for instruction tuning. First of all, we extract the data relevant to the tasks we defined in the previous step. Then we format the data to ensure that it is in a consistent structure and format. For example, the crystal structure in the Vienna Ab initio Simulation Package (VASP) [Hafner, 2008] and CIF files [Hall *et al.*, 1991] are different, so we convert the data into a standardized text format. The reaction data in the Synthesis dataset may contain multiple steps, we extract the key information and format it into a structured format. By preprocessing the data, we ensure that it is ready for instruction generation and tuning, enabling LLMs to learn from the data effectively.

3.3 Annotation and Validation

As shown in fig. 1 Step 3, to ensure the accuracy and relevance of the descriptions, we added annotations from domain experts to the data for each task. The annotated data is then validated through multiple iterations to refine the instructions

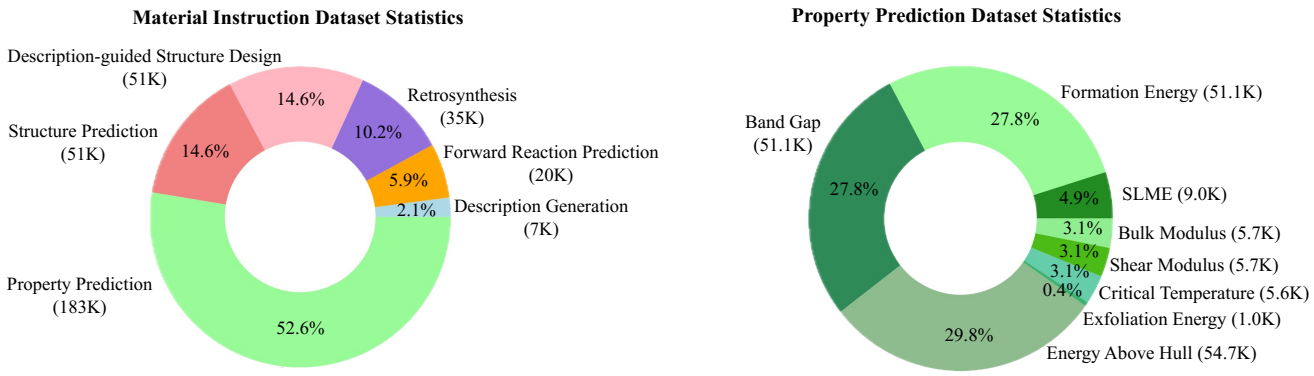

Figure 2: Mat-Instruction dataset statistics. Left: Overview of Mat-Instruction dataset statistics on six core components. Right: Statistics of property prediction part of the Mat-Instruction dataset.

and enhance their effectiveness. The validation process involves checking for redundancy, error, and inconsistencies, to ensure that LLMs can effectively learn from the data. For redundancy screening, we identify and remove duplicate instructions to prevent model overfitting and improve generalization. Error detection involves identifying and removing inaccurate descriptions and invalid crystal structures, ensuring that LLMs receive accurate and reliable information. Inconsistency resolution involves harmonizing conflicting instructions to provide coherent and unified learning for LLMs. So we remove the instructions that are not consistent for the same crystal in terms of structure, property, reaction, and synthesis. By annotating and validating the instructions, we ensure that LLMs can learn from the data effectively and make accurate predictions in various crystal material science domains.

3.4 Instruction Generation

An instruction consists of a task description *prompt*, a context *input*, and an expected output *target*. As shown in fig. 1 Step 4, we design specific and diverse prompts for each of the six core components via human-AI collaboration [Fang *et al.*, 2023]. Since the same task can be described in multiple ways in the real world, we use a combination of human-written and AI-generated prompts to ensure diversity and coverage. For example, a human expert writes a prompt like "Predict the crystal structure of a given material based on its chemical composition", and then we use an LLM to generate variations of this instruction, such as "Determine the crystal lattice structure from the provided chemical formula." This approach ensures that the dataset includes a wide range of instructions, covering different aspects and nuances of material science tasks. By generating diverse and specific instructions, we equip LLMs with the knowledge and guidance needed to excel in material science tasks.

3.5 Integration

With the diverse task descriptions (prompts), contexts (inputs), and expected outputs (targets) generated, we use a template, as shown in listing 1, to integrate the validated instructions into the Mat-Instruction dataset, as illustrated in fig. 1, Step 5 [Taori *et al.*, 2023]. Finally, we integrate the six tasks into our unified Mat-Instruction dataset. Each entry of the dataset contains a task description, context, and expected output, enabling LLMs to learn from the data effectively.

Listing 1 Python Code for Prompt Formatting

```
Instruction = f'Below is an instruction that
    describes a task, paired with an input
    that provides further context. Write a
    response that appropriately completes the
     request.\n\n ### Instruction:\n{Prompt}\
    n\n ### Input:\n{input}\n\n ### Response
    :\n'
```

4 Mat-Instruction Statistics

4.1 Data Overview

Mat-Instruction comprises a total of 349,090 instructions across the six core components, as shown in fig. 2 (left), including 183,654 property prediction instructions, 51,027 CSP instructions, 35,675 crystal retrosynthesis instructions, 20,502 crystal reaction instructions, 7,205 crystal description instructions, and 51,027 description-guided crystal design instructions.

4.2 Task Specific Statistics

Property Prediction Instructions. The property prediction instruction component is the largest (52.6%), aligning with the central importance of property prediction in materials science research. This task involves predicting material properties based on their crystal structures. It consists of diverse property prediction tasks as shown in fig. 2 (right), including band gap, formation energy, elastic constant, etc. Formation energy and energy above the hull are two of the most common property prediction tasks, which are essential for understanding the stability and reactivity of materials. Meanwhile, bandgap also accounts for 27.8%, which is an important property in materials science, affecting the electronic structure and conductivity of materials. The proportions of other properties are relatively small because they are less studied in materials science research. However, they remain important components of the field. By continuously

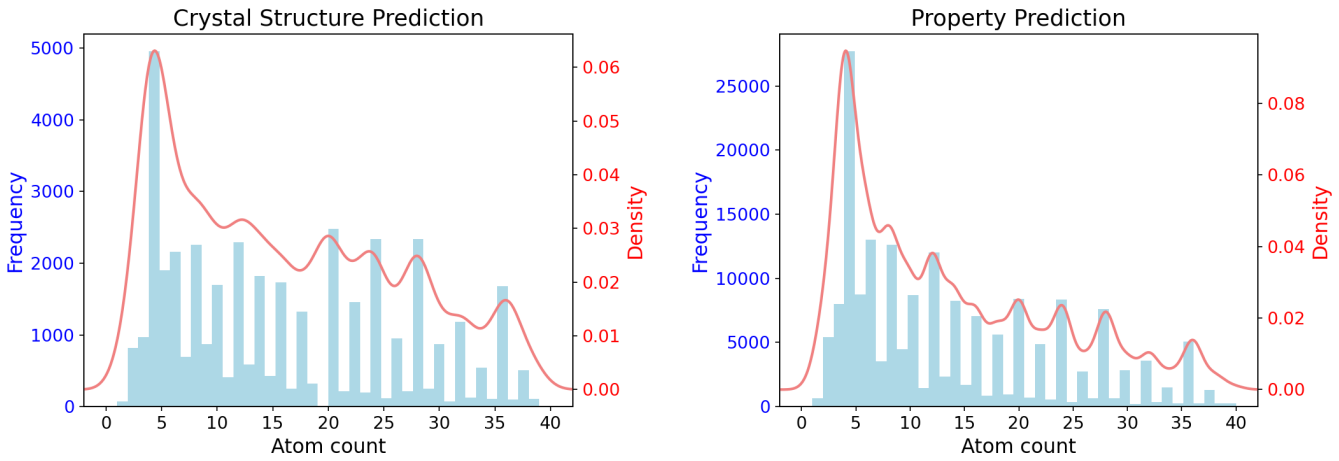

Figure 3: Mat-Instruction dataset atom count statistics. Left: Crystal Structure Prediction part. Right: Property prediction part.

collecting and organizing data in future work, we will further expand Mat-Instruction to cover a broader range of material property prediction tasks. For example, after inputting the crystal structure to the LLM and asking it to predict the band gap of the material, the LLM would then generate a response that predicts the band gap value of the input crystal structure.

Crystal Structure Prediction Instructions. The CSP instruction component is the second largest (14.6%), fulfilling the needs of CSP in material design and discovery. This task involves predicting the crystal structure of a material based on its chemical composition, which is essential for understanding its properties and behavior. For instance, we input the chemical composition of a material to the LLM and ask it to predict the crystal structure of the material. The LLM will then generate a response that describes a stable crystal structure of the material based on the input chemical composition.

Description-guided Crystal Design Instructions. 14.6% of the instructions pertain to description-guided crystal design, which involves predicting the crystal structure of a material based on its description. For example, this includes designing a crystal structure with a specific property or space group. This task is crucial for developing materials with tailored properties, enabling researchers to create novel materials with desired characteristics. Different from the CSP instructions, the description-guided crystal design instructions provide additional context and constraints for the LLM to generate a crystal structure that meets specific requirements. For example, we input a description of a material with specific properties to the LLM and ask it to design a crystal structure that exhibits these properties. The LLM will then generate a response that describes a crystal structure that meets the requirements specified in the input description.

Crystal Retrosynthesis Instructions. The crystal retrosynthesis instruction component is the fourth largest (10.2%), supporting the development of new crystal synthesis approaches in materials science research. This task involves predicting synthesis pathways, conditions, and outcomes of crystal materials, allowing researchers to design and fabricate materials with specific properties. For example, we input the chemical composition of a material into the LLM and ask it to predict the synthesis pathway of the material. The LLM will then generate a response that describes the steps and conditions required to synthesize the material based on the input chemical composition.

Crystal Reaction Instructions. The crystal reaction instruction component is the fifth largest (5.9%), offering insights into crystal reactions in materials science research. This task involves predicting reactions and their outcomes in crystal materials, providing insights into the chemical processes that govern material behavior. For example, we input the reactants of a chemical reaction to the LLM and ask it to predict the reaction product. The LLM will then generate a response that describes the products of the reaction based on the input reactants.

Crystal Description Instructions. The crystal description instruction component is the fifth largest (2.1%), supporting the study of crystal description in materials science research. This task involves interpreting material data and describing the properties, structures, and behaviors of crystal materials, enabling researchers to analyze and understand material properties. For example, we input the chemical composition of a material to the LLM and ask it to describe the material. The LLM will then generate a response that describes the properties, structures, and applications of the material based on the input chemical composition.

4.3 Atom Count Distribution

We select the crystal materials containing fewer than 40 atoms for our Mat-Instruction to ensure a balance between computational efficiency, and experimental feasibility for machine learning modeling, thereby enhancing the scientific value and practical usability of the dataset. The atom count distribution of the CSP instructions is shown in fig. 3. The distribution is skewed towards smaller atom counts, with the majority of instructions containing 1-10 atoms. This reflects the prevalence of crystals with a small number of atoms in CSP tasks, which are often used as building blocks for more

Model	Reaction				Retrosynthesis				Description		
	Right(%)	Rouge	BLEU	Bert	Right (%)	Rouge	BLEU	Bert	Rouge	BLEU	Bert
opt-6.7B	0.00	0.000	0.000	0.791	0.00	0.000	0.000	0.757	0.000	0.000	0.775
Mistral-7B	12.60	0.186	0.000	0.846	0.10	0.056	0.000	0.786	0.126	0.000	0.825
HoneyBee-7B	13.40	0.042	0.000	0.838	2.70	0.075	0.000	0.798	0.106	0.002	0.817
vicuna1.5-7B	12.50	0.160	0.000	0.841	2.33	0.115	0.000	0.802	0.082	0.002	0.811
LLaMA-7B	13.55	0.044	0.000	0.838	2.18	0.076	0.000	0.798	0.103	0.002	0.809
LLaMA2-13B	13.60	0.055	0.000	0.840	1.50	0.044	0.000	0.791	0.134	0.001	0.827
Galactica-6.7B	16.96	0.061	0.000	0.803	4.97	0.087	0.000	0.790	0.097	0.003	0.783
LLaMA-13B	16.09	0.022	0.000	0.827	2.35	0.049	0.000	0.782	0.097	0.002	0.802
Gemma-7B	17.40	0.032	0.000	0.841	1.40	0.065	0.000	0.805	0.123	0.001	0.824
Qwen2-7B	19.60	0.214	0.000	0.844	2.00	0.118	0.000	0.800	0.077	0.001	0.807
LLaMA3-8B	21.10	0.126	0.000	0.841	2.80	0.064	0.000	0.806	0.135	0.001	0.828
LLaMA-7B (FT)	49.73	0.545	0.012	0.931	42.06	0.206	0.008	0.838	0.418	0.209	0.885

Table 1: Results on the test set of the Mat-Instruction dataset.

Model	Property Prediction		Crystal Structure Prediction				Description guided			
	R^2	MAE	Match(%)	Valid(%)	Str	Comp	Match(%)	Valid(%)	Str	Comp
Mistral-7B	-0.8726	80.64								
LLaMA2-13B	-0.8122	80.30								
Gemma-7B	-0.9743	82.86		N/A				N/A		
HoneyBee-7B	-1.0416	77.86								
Qwen2-7B	-0.6775	74.14								
Vicuna1.5-7B	-0.8293	59.71								
LLaMA-7B (FT)	0.7896	3.472	1.29	19.83	0.395	0.975	0.02	44.74	0.429	0.817

Table 2: Results on the test set of the Mat-Instruction dataset, where Str and Comp are the average string and composition similarity scores. (N/A indicates the models lack the ability to generate valid crystal structures.)

complex materials. The distribution also includes a small number of instructions with larger atom counts, indicating the presence of more complex materials in the dataset. By covering a wide range of atom counts, Mat-Instruction enables LLMs to learn from diverse material structures and predict crystal structures across different domains.

5 Mat-Instruction Potential

To evaluate the efficacy of Mat-Instruction in enhancing LLM performance in material science tasks, we conduct extensive instruction tuning experiments on a representative LLM. We use the LLaMA-7B as the base model and fine-tune it on the Mat-Instruction dataset across the six core components. The results demonstrate the effectiveness of Mat-Instruction in improving the LLM's understanding and prediction of crystal materials, thus advancing progress within the field.

5.1 Setup

We conduct instruction tuning experiments on the representative opensource LLM, LLaMA-7B, using the Mat-Instruction dataset [Touvron *et al.*, 2023]. The LLaMA-7B model is a large language model with 7 billion parameters, pre-trained on a diverse range of text corpora.

Data Splitting. We split the Mat-Instruction dataset into training, validation, and test sets with a ratio of 80%, 10%, and 10%, respectively. The training set is used to fine-tune the model, while the validation set is used to tune the hyperparameters and monitor the model performance during training. The test set is used to evaluate the model's performance on unseen data and assess its generalization ability. We report the model performance on the test set across various tasks.

Baseline Model. We compare the performance of the fine-tuned LLaMA-7B with the base LLaMA-7B models, opensource LLMs (including LLaMA-13B, LLaMA2-13B, LLaMA3-8B [Touvron *et al.*, 2023], Qwen [Bai *et al.*, 2023], Mistral-7B [Jiang *et al.*, 2023], Gemma-7B [Team *et al.*, 2024], and Vicuna-7B [Chiang *et al.*, 2023]), and science LLMs (Galactiva-6.7B [Taylor *et al.*, 2022] and HoneyBee-7B [Song *et al.*, 2023]).

Evaluation Metrics. For crystal property prediction tasks, we mainly employ the **Mean Absolute Error (MAE)** and **Pearson relation coefficient (R^2)** to evaluate the model's performance. For structure-related tasks, including CSP and description-guided crystal design, we report the **Match rate**, **Valid rate**, **structure similarity**, and **composition similarity**. **Match rate** indicates the percentage of predicted structures that match the ground truth structures. **Valid rate** indicates the percentage of valid structures predicted by the model. **Structure similarity** measures the similarity between the predicted and ground truth crystal structures. **Composition similarity** measures the similarity between the predicted and ground truth chemical compositions of the crystal struc-

tures. For description-related tasks, including crystal description, retrosynthesis, and reaction, we report the **BLEU score**, **ROUGE score**, and **BERT score**, which are commonly used evaluation metrics in natural language processing (NLP) and text generation tasks. They measure the similarity between a generated text and a reference text, providing insights into the prediction quality. Besides, we report the **correctness rate (Right)** for reaction and retrosynthesis tasks, which indicates the percentage of correct predictions made by the model.

5.2 Experimental Results

We present the experimental results of the fine-tuned LLaMA-7B model on the Mat-Instruction dataset across the six core components. The results demonstrate the effectiveness of Mat-Instruction in enhancing the LLM's performance in material science tasks, enabling it to understand and predict crystal materials with high accuracy and precision.

Description-related Tasks. The fine-tuned LLaMA-7B model demonstrates strong performance in crystal reaction prediction tasks, retrosynthesis prediction tasks, and crystal description tasks, as shown in table 1. For crystal reaction prediction tasks, the correctness rate is 49.73%, with ROUGE, BLEU, and BERT scores of 0.545, 0.012, and 0.931, respectively, all of which significantly outperform the baseline model. For retrosynthesis prediction tasks, the correctness rate is 42.06%, with ROUGE, BLEU, and BERT scores of 0.206, 0.008, and 0.838, respectively, also demonstrating a substantial improvement over the baseline model. For crystal description tasks, the ROUGE, BLEU, and BERT scores are 0.418, 0.209, and 0.885, respectively, again exhibiting a significant performance gain over the baseline model. Comparing the fine-tuned LLaMA-7B model with the baseline model, we observe a substantial improvement in the model's performance across all tasks, highlighting the effectiveness of Mat-Instruction in enhancing the LLM's understanding and prediction of crystal materials.

Structure-related Tasks. The fine-tuned LLaMA-7B model achieves high accuracy in predicting material properties, generating valid crystal structures, and describing crystal materials as shown in table 2, where N/A indicates the model's lack of the ability to predict crystal structures. For the task of property prediction, the R^2 score achieves 0.789, and the MAE is 3.472, indicating the model's high accuracy in predicting material properties. For CSP tasks, the match rate is only 1.29%, and the valid rate is 19.83%, demonstrating the model's ability to generate valid crystal structures, while it's hard to match the ground truth structure. The structure similarity and composition similarity of the model in the CSP task achieves 0.395 and 0.975, respectively, which indicates the ability of the fine-tuned model to generate crystal structures following the given composition. For description-guided crystal design tasks, the match rate is only 0.02%, while the valid rate achieves 44.74%, which also indicates the model's ability to generate valid crystal structures. The structure and composition similarity achieve 0.429 and 0.817 respectively. Compared to CSP, structural similarity is higher due to the inclusion of more structural information in the instructions. However, composition similarity is lower, as in practice, only specific properties may be required, and composition details are not always provided in the instructions. The other models that are not fine-tuned on Mat-Instruction, can hardly generate valid structures or make correct property predictions.

6 Discussion

6.1 Crystal Structure Representations

In this work, we employ a concise yet comprehensive crystal structure format, VASP [Hafner, 2008], to represent the crystal structures in the Mat-Instruction dataset. This format includes lattice parameters, atomic species, and fractional coordinates, providing a standardized and detailed representation of crystal materials. While crystal structures can be represented in various formats, like CIF and VASP formats, which include additional details like symmetry operations, the representation used in our dataset remains complete and can be converted into any other format as needed. Future research on LLMs in materials science could take advantage of diverse crystal structure representations to further enhance the understanding and representation learning of crystal materials.

6.2 Model Performance & Task Complexity

The fine-tuned LLaMA-7B model demonstrates significant improvements in performance across all tasks, as shown in table 1 and table 2. However, the match rate and valid rate for CSP and description-guided crystal design tasks are relatively low, indicating the complexity and challenges associated with these tasks for LLMs. We attribute this to the intricate nature of crystal structures and the diverse requirements of the tasks, which may involve multiple constraints and conditions. Future research on LLMs in materials science could explore ways to address these challenges with our Mat-Instruction dataset.

7 Conclusion

In this work, we introduce Mat-Instruction, a comprehensive instruction dataset tailored to the unique challenges of crystal materials science. Mat-Instruction comprises six core components, each designed to address specific materials science tasks, including crystal structure prediction, property analysis, and crystal design. To construct Mat-Instruction, we integrate data from multiple sources, transform it into instruction-based formats, and validate the instructions with domain experts. Our extensive experiments on open-source LLMs demonstrate the effectiveness of Mat-Instruction in enhancing LLM performance in materials science tasks, enabling more accurate understanding and prediction of crystal materials. By equipping LLMs with domain-specific knowledge, Mat-Instruction would pave the way for new scientific discoveries and innovations in material science. Thus, we plan to open source Mat-Instruction, allowing the community to further expand and refine it. Our goal is to further improve LLM's understanding and prediction of crystal materials, with the expectation of driving significant advancements in materials science research and fostering societal progress.

Acknowledgments

We thank the PaddlePaddle for providing computing resources, and this work utilizes PaddlePaddle as the deep learning framework.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Anand *et al.*, 2023] Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. *GitHub*, 2023.

[Antunes *et al.*, 2024] Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):1–16, 2024.

[Bai *et al.*, 2023] 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.

[Barroso-Luque *et al.*, 2024] L Barroso-Luque, M Shuaibi, X Fu, BM Wood, M Dzamba, M Gao, A Rizvi, CL Zitnick, and ZW Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771*, 2024.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[Choi and Lee, 2024] Jaewoong Choi and Byungju Lee. Accelerating materials language processing with large language models. *Communications Materials*, 5(1):13, 2024.

[Choudhary *et al.*, 2020] Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.

[Curtarolo *et al.*, 2012] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H Taylor, Lance J Nelson, Gus LW Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, et al. Aflowlib. org: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, 2012.

[Dagdelen *et al.*, 2024] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024.

[de Pablo *et al.*, 2019] Juan J de Pablo, Nicholas E Jackson, Michael A Webb, Long-Qing Chen, Joel E Moore, Dane Morgan, Ryan Jacobs, Tresa Pollock, Darrell G Schlom, Eric S Toberer, et al. New frontiers for the materials genome initiative. *npj Computational Materials*, 5(1):41, 2019.

[Fang *et al.*, 2023] 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. *arXiv preprint arXiv:2306.08018*, 2023.

[Gao *et al.*, 2013] Zhi-Yu Gao, Guo-Quan Liu, et al. Recent progress of web-enable material database and a case study of nims and matweb. *Journal of Materials Engineering*, 3(11):89–96, 2013.

[Hafner, 2008] Jürgen Hafner. Ab-initio simulations of materials using vasp: Density-functional theory and beyond. *Journal of computational chemistry*, 29(13):2044–2078, 2008.

[Hall *et al.*, 1991] Sydney R Hall, Frank H Allen, and I David Brown. The crystallographic information file (cif): a new standard archive file for crystallography. *Foundations of Crystallography*, 47(6):655–685, 1991.

[Jain *et al.*, 2013] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013.

[Jiang *et al.*, 2023] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[Liu *et al.*, 2024] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

[Merchant *et al.*, 2023] 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.

[Miret and Krishnan, 2024] Santiago Miret and Nandan M Krishnan. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder,

Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[Rubungo *et al.*, 2024] Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. Llm4mat-bench: benchmarking large language models for materials property prediction. *arXiv preprint arXiv:2411.00177*, 2024.

[Sanh *et al.*, 2022] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[Song *et al.*, 2023] Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. Honeybee: Progressive instruction finetuning of large language models for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5724–5739, 2023.

[Taori *et al.*, 2023] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model. *GitHub repository*, 2023.

[Taylor *et al.*, 2022] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *CoRR*, abs/2211.09085, 2022.

[Team *et al.*, 2024] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[Tilley, 2020] Richard JD Tilley. *Crystals and crystal structures*. John Wiley & Sons, 2020.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

[Venugopal and Olivetti, 2024] Vineeth Venugopal and Elsa Olivetti. Matkg: An autonomously generated knowledge graph in material science. *Scientific Data*, 11(1):217, 2024.

[Wang *et al.*, 2022] Zheren Wang, Olga Kononova, Kevin Cruse, Tanjin He, Haoyan Huo, Yuxing Fei, Yan Zeng, Yingzhi Sun, Zijian Cai, Wenhao Sun, et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Scientific data*, 9(1):231, 2022.

[Zhang *et al.*, 2023] Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, Wenhao Huang, and Jie Fu. Chinese open instruction generalist: A preliminary release. *CoRR*, abs/2304.07987, 2023.

[Zhang *et al.*, 2024] Yuwei Zhang, Fangyi Chen, Zeyi Liu, Yunzhuo Ju, Dongliang Cui, Jinyi Zhu, Xue Jiang, Xi Guo, Jie He, Lei Zhang, et al. A materials terminology knowledge graph automatically constructed from text corpus. *Scientific Data*, 11(1):600, 2024.