
FOUNDATIONAL LARGE LANGUAGE MODELS FOR MATERIALS RESEARCH

Vaibhav Mishra^{1,*}, Somaditya Singh^{1,*}, Dhruv Ahlawat^{1,*}, Mohd Zaki^{2,*}, Vaibhav Bihani³, Hargun Singh Grover³, Biswajit Mishra⁴, Santiago Miret⁵, Mausam^{1,3,#}, N. M. Anoop Krishnan^{2,3,#}

¹Department of Computer Science and Engineering, ²Department of Civil Engineering

³Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi

⁴Cerebras Systems, Inc., ⁵Intel labs

#Corresponding authors: {mausam, krishnan}@iitd.ac.in

*Authors contributed equally.

Abstract

Materials discovery and development are critical for addressing global challenges in renewable energy, sustainability, and advanced technology. Yet, the exponential growth in materials science literature comprising vast amounts of textual data has created significant bottlenecks in knowledge extraction, synthesis, and scientific reasoning. Large Language Models (LLMs) offer unprecedented opportunities to accelerate materials research through automated analysis and prediction. Still, their effective deployment requires domain-specific adaptation for language understanding and solving domain-relevant tasks. Here, we present LLAMAT, a family of foundational models for materials science developed through continued pretraining of LLAMA models on an extensive corpus of materials literature and crystallographic data. Through systematic evaluation, we demonstrate that LLAMAT excels in materials-specific natural language processing and structured information extraction while maintaining general linguistic capabilities. The specialized LLAMAT-CIF variant demonstrates unprecedented capabilities in crystal structure generation, predicting stable crystals with high coverage across the periodic table. Intriguingly, despite LLAMA-3's superior performance in comparison to LLAMA-2, we observe that LLAMAT-2 demonstrates unexpectedly enhanced domain-specific performance across diverse materials science tasks, including structured information extraction from text and tables, more particularly in crystal structure generation—suggesting a potential “adaptation rigidity” in overtrained LLMs. Altogether, the present work demonstrates the effectiveness of domain adaptation towards the development of practically deployable LLM copilots for materials research. Beyond materials science, our findings reveal important considerations for domain adaptation of LLMs—model selection, training methodology, and domain-specific performance—that may influence the development of specialized scientific AI systems.

1 Introduction

Materials innovation can address ten of the seventeen United Nations Sustainable Development Goals through advances in sustainable energy systems, advanced electronics, and environmentally conscious manufacturing. This imperative for accelerated materials discovery coincides with an unprecedented expansion in the scientific literature—exceeding 6 million materials science publications—presenting both opportunities and challenges for materials informatics [1, 2, 3]. Obtaining actionable insights from this information explosion requires advanced computational tools that can effectively process vast scientific literature, the majority of which is unstructured text data and tables of varying structure.

Large Language Models (LLMs), also referred to as foundation models, have demonstrated remarkable capabilities in text processing, analysis, and generation [4]. In the field of materials, LLMs can enhance the research and discovery process through (i) rapid literature-based identification of materials [5, 6] and synthesis pathways [7], (ii) crystal structure generation [8, 9, 10], (iii) autonomous experimental planning [11, 12, 13], and (iv) results analysis [14, 15]. Recent advances [16, 17, 18, 19] have demonstrated efficacy of LLMs in materials concept comprehension, domain-specific query resolution [18, 20, 21], and simulation code generation [14]. However, critical analyses of the performance of these general-purpose LLMs reveal their inability to address domain-specific challenges [3, 18, 22, 14], including the correct interpretation of scientific phenomena such as physical laws or theories, specialized terminologies, and viable crystal structures [8, 23].

Effectively leveraging LLMs for materials research requires specialized domain adaptation to address their limitations in materials-specific information processing [3]. Initial efforts toward domain adaptation of LLMs by fine-tuning them for specific tasks in materials research have yielded promising breakthroughs in structured information extraction [24], materials-specific natural language processing [25, 16, 26], experimental data analysis [27, 22], and crystal structure generation [9, 8, 28, 29]. These achievements highlight the potential for a unified materials foundation model that integrates these capabilities to accelerate research and development.

Here, we introduce LLAMAT—a family of domain-adapted language models demonstrating generalist material science capabilities. Through a systematic approach combining pretraining, instruction fine-tuning, and task-specific fine-tuning, LLAMAT enables advanced scientific natural language processing, information extraction, and crystal generation. Our comprehensive evaluation demonstrates that these models outperform existing approaches across diverse materials science tasks and exhibit emergent capabilities that bridge the gap between human expertise and automated materials discovery.

2 Results

2.1 LLAMAT: A Family of Large Language Model for Materials

LLAMAT is developed by systematically embedding materials domain knowledge on LLAMA base models—specifically LLAMA-2-7B [30] and LLAMA-3-8B[31], hereafter referred to as LLAMA-2 and LLAMA-3, respectively. While larger LLAMA variants, such as the 70B models, could yield superior performance, our model selection optimizes the balance between computational demands for training and inference, available pretraining data volume, and practical deployment considerations for the larger materials community. To develop LLAMAT, we employed a rigorously designed three-stage pretraining-finetuning process (see Fig. 1, Methods). The initial stage comprised continued pretraining (CPT) on the base LLAMA models with an extensive and meticulously curated corpus, namely R2CID (see Methods and Tab. A.1 in App. A for details) with greater than 30 billion tokens of materials science (MatSci) knowledge, encompassing approximately 4 million peer-reviewed publications (94.43%), crystallographic information files (2.499%), and materials science community discourse (0.019%). Additionally, we incorporated a strategic 3% subset of REDPAJAMA data, the original training corpus of LLAMA models, to preserve fundamental linguistic capabilities while concurrently mitigating catastrophic forgetting.

Subsequently, we implemented two distinct finetuning pathways to develop specialized LLAMAT variants. The first variant, LLAMAT-Chat, underwent comprehensive instruction finetuning (IFT) across multiple domains, including general English comprehension, mathematical reasoning, and MatSci-specific datasets (see Methods and App. A.1). This model was further finetuned on a single corpus comprising several materials-relevant downstream tasks (see Tab. A.2), resulting in a materials research copilot with demonstrated proficiency in natural language tasks related to materials science, including named entity recognition, relation classification, and text classification to name a few, as well as structured information extraction from scientific text and tables (App. A.2). Concurrently, we developed LLAMAT-CIF models through IFT of LLAMAT models on crystallographic information files, a hand-curated dataset comprising five syntactic and four semantic tasks. Following this, parameter-efficient finetuning (PEFT) was employed on LLAMAT-CIF to enable crystal generation, a task of importance in materials discovery (see Methods(Section 4) for details).

To obtain the best-performing models, we conducted extensive experiments balancing the datasets, both during CPT and IFT (Appendix D), with the goal of developing a model that provides the best performance on MatSci tasks while not losing its original English capabilities. To this end, in IFT, we included datasets on general English comprehension (using OpenOrca[32]) and mathematical reasoning (using MathQA[33]) alongside materials science-specific tasks (see App. D), including datasets from MatSciNLP [34] and original datasets on question-answering on materials domains such as MatBookQA (3000 QA pairs), MaScQA (2000 QA pairs), and MatSciInstruct (170k QA pairs) [35] (see App. A). Systematic evaluation of model performance

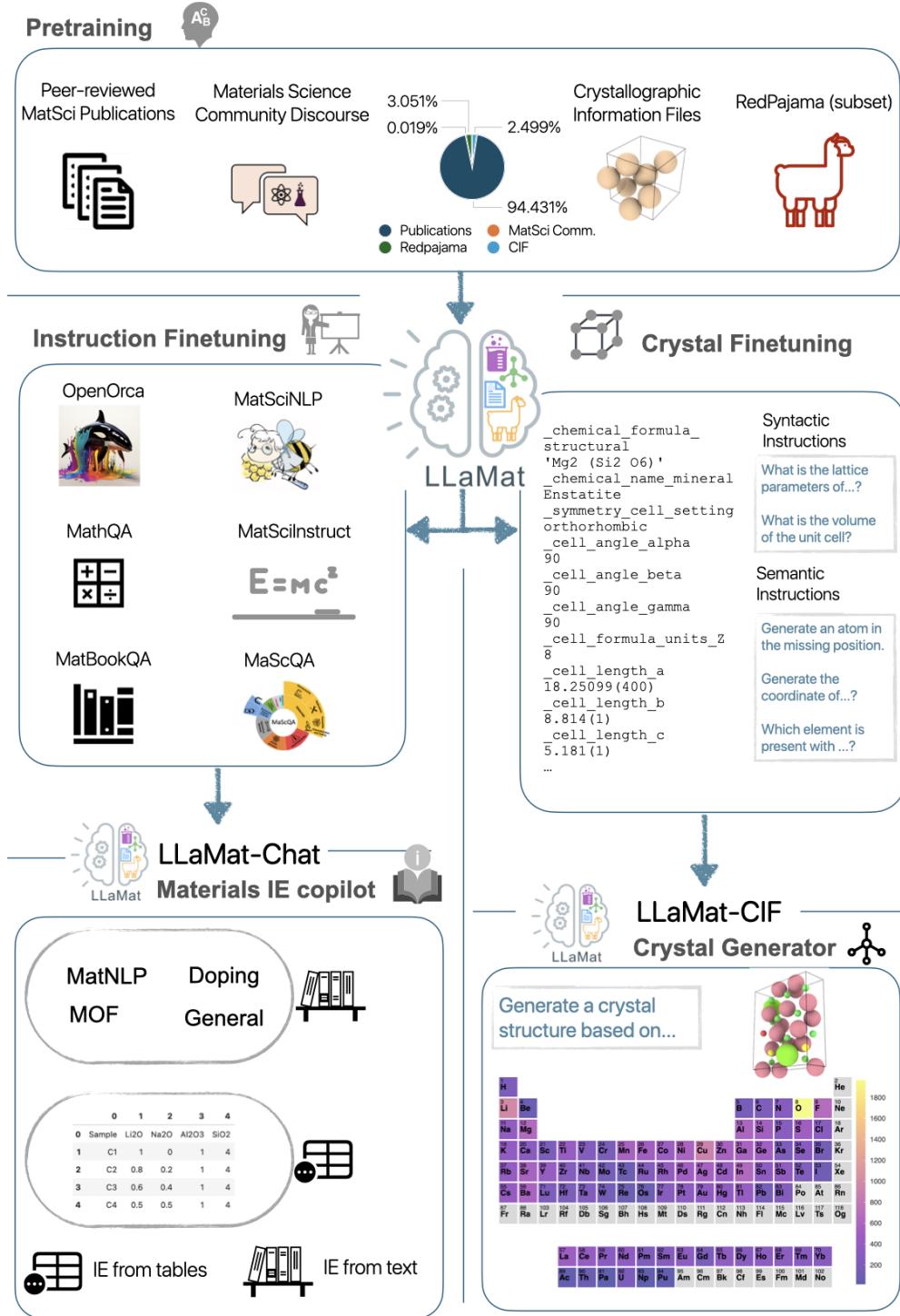


Figure 1: Development pipeline and capabilities of LLaMat for materials science applications. The schematic illustrates the two-stage development of LLAMAT, beginning with continuous pretraining on materials science corpora (top), followed by specialized instruction finetuning pathways (left and right). The pretraining dataset composition is shown in the pie chart, comprising peer-reviewed publications (94.43%), crystallographic information files (CIF, 2.50%), and a subset of RedPajama (3.051%). Two distinct finetuning pathways yield LLAMAT-Chat, a materials research copilot capable of structured information extraction and materials NLP tasks (left branch), and LLAMAT-CIF, specialized in crystal structure analysis and generation (right branch). Representative examples demonstrate the dataset details and model's capabilities in handling diverse materials science queries and tasks.

during CPT and IFT stages revealed the critical role of dataset distribution and learning rates (see App. C). More importantly, the hyperparameters and dataset distribution influencing model performance were found to be distinct for LLAMA-2 and LLAMA-3 (see Apps. C and D).

The CPT and IFT of LLAMAT models revealed several notable insights into the domain adaptation process of LLMs. Compared to intermediate checkpoints, CPT on domain-specific corpus consistently demonstrated superior performance metrics for both LLAMA-2 and LLAMA-3 architectures. During IFT on OpenOrca, model-specific behavioural patterns were observed: while LLAMA-2 showed substantial improvements across evaluation metrics, LLAMA-3 demonstrated minimal performance gains across materials science and general language tasks (see Tab. D.2). Models trained without MathQA[33] in their finetuning regime exhibited severe degradation in mathematical reasoning capabilities—failing to solve even elementary arithmetic problems despite maintaining reasonable linguistic performance relative to their respective base models (Tab. D.3). This finding underscores the presence of datasets pertaining to diverse capabilities during the domain adaptation process.

Interestingly, following the IFT OpenOrca and MathQA, additional IFT of LLAMAT models on materials-specific datasets, such as Honeybee [35] did not yield significant performance improvements of LLAMAT models on either English or MatSci tasks (see Tab. D.3 in Appendix). This unexpected observation suggests a fundamental distinction between domain knowledge acquisition and instruction-following capabilities: while domain adaptation through pretraining and finetuning effectively enhances field-specific performance, the development of robust instruction-following competency appears to be independently trainable through generic question-answer datasets. Through rigorous parametric optimization studies, we identified Pareto-optimal dataset configurations for each base model, effectively maximizing materials science task performance while maintaining robust general language capabilities (App. D).

2.2 Materials Research Copilot

To assess the model’s efficacy as a materials research copilot, we conducted systematic evaluations across two critical domains: Materials Natural Language Processing (MatNLP) and Materials Structured Information Extraction (MatSIE). These evaluations specifically targeted the model’s ability to comprehend complex materials science concepts and extract structured information from both textual and tabular data in scientific publications, representing fundamental capabilities required for materials research automation.

Materials Language Processing. MatNLP encompasses three fundamental natural language processing task families: entity recognition, extraction, and classification. The evaluation framework comprises ten materials-specific and four English datasets, totaling 14,579 test instances. These tasks systematically assess the model’s capability to extract granular information from materials literature—including synthesis protocols, characterization methods, and application-specific entities. They also include classification tasks (for instance, whether a particular document is related to a topic in materials) and entity relationship comprehension. The English dataset provides a complementary assessment of general language capabilities through question-answering and multiple-choice tasks.

We evaluate the performance of LLAMAT-2 and -3 models and their chat variants on this dataset and compare them to their respective base models and two closed source models: Claude-3 Haiku and Gemini-1.5 Flash-8B. For a fair comparison, we also finetuned (FT) both pretrained and chat variants of LLAMA-2 and LLAMA-3 on the training dataset of downstream task. Figure 2 presents a comprehensive performance analysis of LLAMAT in comparison to the finetuned LLAMA variants. The micro and macro F1 scores (Figures 2a,b) reveal LLAMAT-3 -Chat’s superior performance compared to non-chat variants, demonstrating the effectiveness of our domain-specific CPT-IFT strategy. Further, the performance of closed source models is inferior compared to LLAMAT models except for one extraction task. Triple experimental iterations yield minimal standard deviations, evidenced by compact error bars, confirming robust model performance (see App. E). The error bars are not associated with the closed source models because the inference was done only once using these models.

Our performance analysis reveals interesting architectural dependencies in domain adaptation capabilities. While LLAMAT-3 variants show greater relative improvement from their base model compared to LLAMAT-2 implementations, the finetuned LLAMAT-2 models consistently outperform their LLAMA-3 counterparts. This counterintuitive pattern persists even in CPT models without IFT, where LLAMAT-2 demonstrates superior performance. This observation suggests a potential domain adaptation limitation in LLAMA-3, possibly stemming from its extensive pretraining (~ 3 orders of magnitude more data) despite superior base model performance. This phenomenon, referred to hereafter as “adaptation rigidity,” a recurring observation

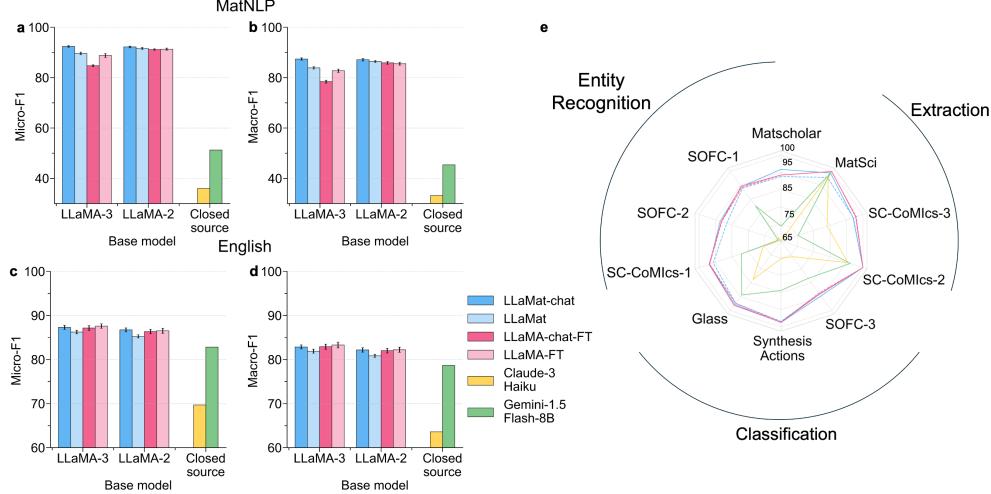


Figure 2: Comparative performance analysis of LLAMAT and LLAMA models across materials science and general language tasks with closed source models: Claude and Gemini. a, Micro-F1, and b, Macro-F1 scores demonstrate performance on materials science tasks, with error bars representing a standard deviation from three independent evaluations. c, Micro-F1, and d, Macro-F1 scores on general English language tasks, showing maintained capabilities across domain adaptation. e, Radar plot illustrating task-specific performance across diverse materials science applications, including entity recognition, extraction, and classification tasks. Dark and light blue lines represent LLAMAT-Chat and LLAMAT, respectively; dark and light pink lines represent LLAMA-Chat and LLAMA variants. The yellow and green lines represent the closed source models Claude-3 Haiku and Gemini-1.5 Flash-8B. Solid lines indicate chat models, while dashed lines represent non-chat variants. For materials science tasks, higher scores indicate better performance in extracting domain-specific information, identifying relationships between materials entities, and classifying scientific text. Results demonstrate that domain-specific pretraining enhances materials science task performance while preserving general language capabilities.

as discussed in later results, underscores the complex relationship between model architecture, pretraining scale, and domain adaptation efficacy[36, 37]. Nevertheless, both LLAMAT variants consistently surpass their respective base LLAMA models in performance metrics.

While LLAMAT models demonstrate superior performance on MatNLP tasks, it is important to analyze whether this improvement is at the cost of their performance on general language tasks. We observe that the English language task performance (Figures 2c,d) exhibits minimal cross-implementation variance, validating our strategic use of the subset of REDPAJAMA dataset during CPT and OpenOrca during IFT. The radar plot (Figure 2e) provides a granular analysis of micro-F1 scores across MatNLP dataset subsets, with solid and dotted lines differentiating chat and non-chat variants. Most notably, LLAMAT-3 -Chat model demonstrates consistent performance advantages across diverse materials science tasks, including entity recognition, classification, and extraction tasks, establishing their efficacy for broader materials science applications.

Structured Information Extraction from Text. The materials science literature contains vast amounts of critical information about material compositions, synthesis protocols, and properties embedded within unstructured text. Extracting this information in a structured format is essential for accelerating materials discovery but traditionally requires extensive manual annotation and specialized model development for each extraction task. This challenge is particularly acute in specialized domains such as doping studies and metal-organic frameworks (MOFs), where precise extraction of chemical compositions, structural relationships, and functional properties is crucial. While recent studies have demonstrated the potential of finetuned commercial LLMs for these tasks [16, 38], their proprietary nature and associated costs limit scalable deployment across the millions of articles in materials literature, necessitating the development of open-source alternatives optimized for materials science applications [6].

Having established the superior performance of LLAMAT-Chat models in MatNLP tasks, we now evaluated their structured information extraction capabilities. Figure 3a demonstrates the performance of LLAMAT-Chat

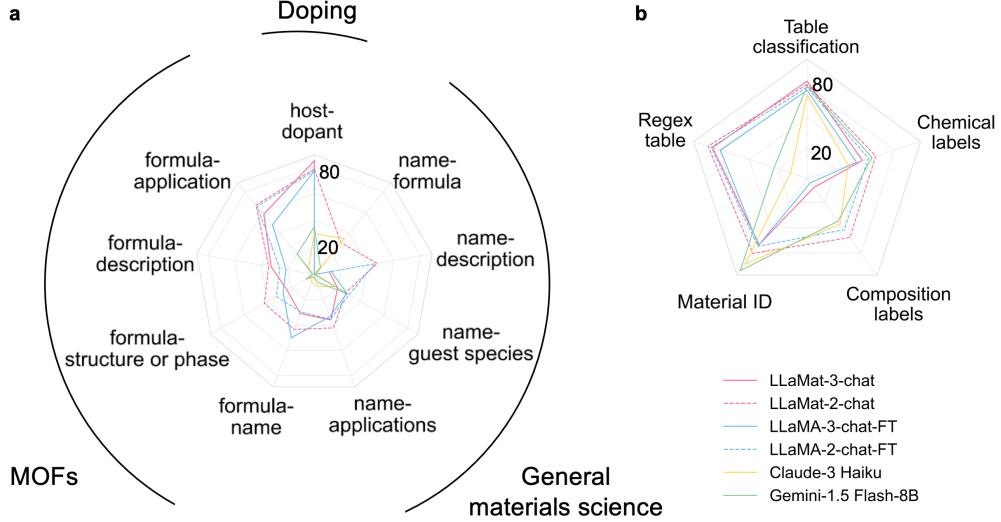


Figure 3: Performance evaluation of structured information extraction capabilities across materials science subdomains. a, Radar plot showing F1-scores for various information extraction tasks in metal-organic frameworks (MOFs), doping studies, and general materials science. b, Performance comparison on structured information extraction from materials science tables, including table classification, chemical label identification, and composition extraction. Solid and dashed lines represent chat and finetuned variants of LLAMAT (pink) and LLAMA (blue) models, respectively, with radial axes indicating F1-scores.

models and closed source models across nine distinct extraction tasks in the Doping, metal-organic framework (MOF), and general materials domains showcasing better capabilities of the former compared to closed source models. The results of all the variants of LLAMA and LLAMAT models are provided in App. E. The radar plots reveal that both LLAMAT-2 and LLAMAT-3 chat variants consistently outperform their finetuned LLAMA counterparts in extracting relationships between host materials and dopants, formula-structure mappings, and application-specific information.

Notably, LLAMAT-2-chat exhibits particularly strong performance in formula-application relationships and host-dopant associations, while LLAMAT-3-chat slightly outperforms in formula-name tasks. This performance pattern aligns with our earlier observations of the “adaptation rigidity” phenomenon, where the LLAMAT-2-Chat model exhibits significantly enhanced capabilities compared to its successor after domain adaptation through CPT and IFT. This consistent trend across evaluation metrics reinforces our hypothesis about the inverse relationship between the initial pretraining scale and domain adaptation efficacy.

Information Extraction from Tables. Tables in the materials domain serve as structured repositories of composition–property data yet present unique challenges due to their heterogeneous formats and complex organizational schemas across publications [39, 22]. This inherent variability in tabular data representation demands advanced language models capable of understanding the context of materials science and extracting structured information with high fidelity.

We now evaluate the capability of LLAMAT models to extract meaningful information from materials tables. To this end, we consider five critical capabilities: compositional table classification, chemical constituent localization, composition extraction, material identifier recognition, and regex-amenable information identification. We consider a set of 737 tables from peer-reviewed publications to evaluate the same. These tables were presented in a challenging manually annotated benchmark dataset for information extraction from tables [39]. Figure 3b confirms a recurring pattern: LLAMAT-2 and LLAMA-2 models consistently outperform their third-generation counterparts across all evaluation metrics, particularly in chemical label identification and composition extraction tasks. This observation aligns with our previous findings regarding the enhanced domain adaptability of second-generation architectures, suggesting that this advantage extends to structured data interpretation tasks. Detailed performance metrics and task-specific analyses are provided in App. F.

2.3 Crystal Generation

Crystal structure prediction represents a fundamental challenge in materials discovery, traditionally addressed through computationally intensive methods such as density functional theory (DFT) calculations, generative models [40, 41, 42, 43], and Graph Neural Networks [44, 45, 46, 47]. Language models offer an alternative paradigm despite lacking explicit crystallographic optimization. Recent works [9, 8, 10] demonstrate the potential of LLMs toward crystal generation.

Here, we evaluate the capability of LLAMAT to generate crystal structure. To this end, we developed LLAMAT-CIF through a comprehensive three-phase optimization strategy: CIF pretraining with natural language descriptions, crystallographic instruction finetuning, and PEFT for structure generation. Quantitative evaluation reveals superior performance of LLAMAT-2-CIF across multiple metrics (Tab. 1), achieving exceptional composition validity (0.995), high coverage (0.986 recall, 0.996 precision), and improved stability prediction (49.49% stable structures). The performance patterns reinforce our earlier observations of “adaptation rigidity” in LLAMAT-3: despite its ability to generate more complex structures, it exhibits lower structural validity (0.674) and generation efficiency, requiring roughly 33,000 attempts versus 13,000 for LLAMAT-2-CIF to produce 10,000 structures fit for further evaluation pipeline. Notably, LLAMAT-2-CIF demonstrates optimal performance across metrics, though inter-model variations suggest hyperparameter sensitivity [48] (see App. C.2 for the loss curve).

Analysis of the generated structures reveals distinct characteristics (see Fig. 4a) demonstrate contrasting behaviours: LLAMAT-3-CIF generates structures with higher elemental complexity (peaks around 24-32 elements) while LLAMAT-2-CIF favours more straightforward compositions (peaks around 6-12 elements). Energy profiles show both models generate thermodynamically reasonable structures, with distributions centered near 0 eV/atom, though LLAMAT-2-CIF exhibits a tighter distribution, suggesting more consistent stability.

Crystallographic system analysis (Fig. 4b) reveals a consistent preference hierarchy across both models: rhombohedral structures dominate (~4,000 instances), followed by monoclinic and orthorhombic systems (~2,000 each). Interestingly, this distribution is distinct from the CIF dataset used for CPT and IFT. The periodic table visualization (Fig. 4c) for LLAMAT-2-CIF exposes systematic compositional biases that align with chemical intuition: minimal actinide incorporation (<50 instances), balanced representation across transition elements (200-400 instances), and predominant oxygen presence (>1,600 instances)—patterns reflecting natural abundance and synthetic accessibility. Beyond structure generation, LLAMAT-CIF models demonstrate versatility in various CIF-related tasks (details in App. H).

Table 1: Comparison of crystal structure generation capabilities across different model architectures. Performance evaluation using multiple metrics: validity (structural integrity and composition correctness), coverage (recall and precision of generated structures), property distribution (Wasserstein distance for density (ρ) and number of elements (N_{el})), and thermodynamic stability (percentage of structures predicted stable by M3GNet). Arrows indicate metrics’ desired direction (\uparrow : higher is better, \downarrow : lower is better). The top section shows baseline results from state-of-the-art methods [9]. LLAMAT-2-CIF demonstrates superior performance across most metrics, particularly in composition validity (0.995) and stability prediction (49.49%), while maintaining high coverage (0.986 recall, 0.996 precision). Bold values indicate the best performance for each metric.

Method	Validity		Coverage		Property Dist.		Stability
	Struct. \uparrow	Comp. \uparrow	Recall \uparrow	Prec. \uparrow	$\rho\downarrow$	$N_{el}\downarrow$	M3GNet \uparrow
CDVAE [9]	1.000	0.867	0.991	0.995	0.688	1.43	28.8%
LLAMA-2 [9]							
7B ($\tau = 1.0$)	0.918	0.879	0.969	0.960	3.850	0.96	35.1%
7B ($\tau = 0.7$)	0.964	0.933	0.911	0.949	3.610	1.06	35.0%
13B ($\tau = 1.0$)	0.933	0.900	0.946	0.988	2.200	0.05	33.4%
13B ($\tau = 0.7$)	0.955	0.924	0.889	0.979	2.130	0.10	38.0%
Present work							
LLAMAT-2-CIF	0.878	0.995	0.986	0.996	0.623	0.023	49.49%
LLAMAT-3-CIF	0.674	0.693	0.925	0.994	12.355	0.261	42.95%

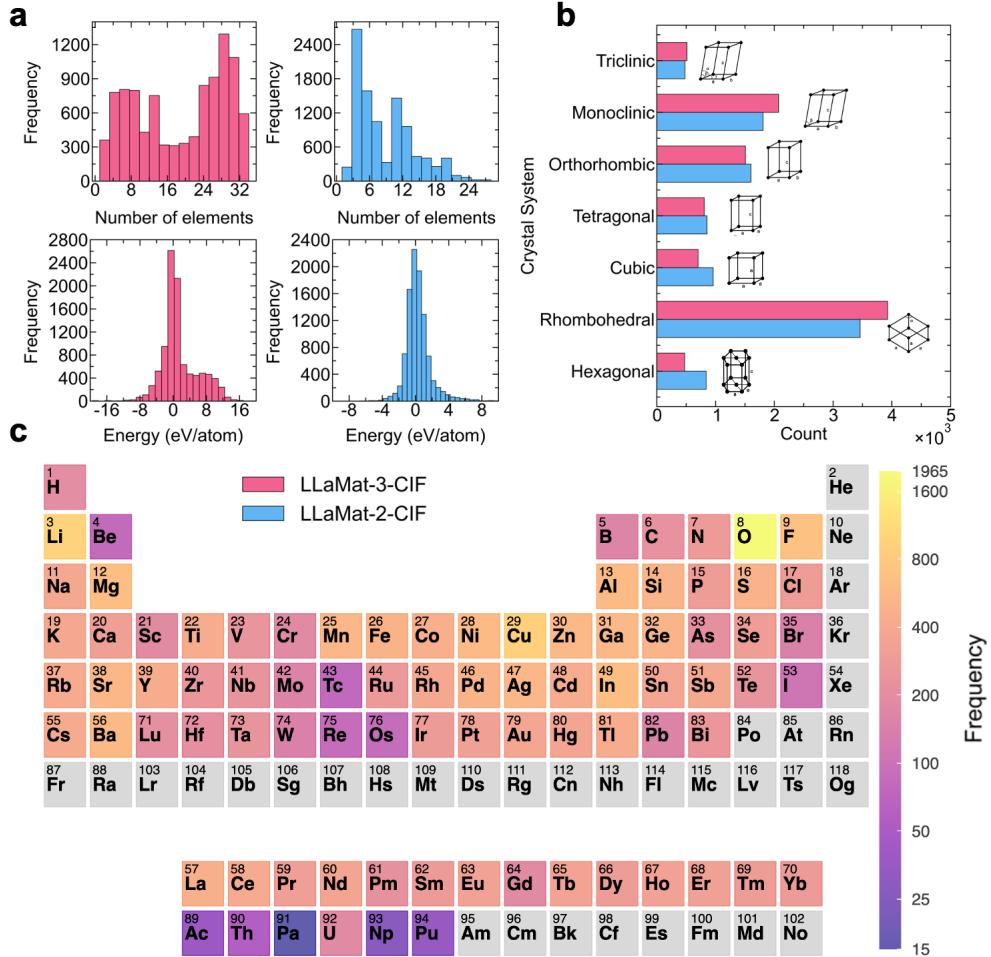


Figure 4: Compositional and structural analysis of crystal structures generated by LLaMat models. a, Distributions characterizing 10,000 generated structures from LLAMAT-3-CIF (pink) and LLAMAT-2-CIF (blue). Top: Frequency distributions of elemental complexity showing the number of unique elements per structure. Bottom: Energy distributions after M3GNet relaxation, indicating thermodynamic stability of generated structures. b, Distribution of Bravais lattice systems in generated structures with corresponding unit cell representations, revealing preferential generation of rhombohedral and monoclinic systems. c, Periodic table heat map visualizing elemental frequency across 10,000 structures generated by LLAMAT-2-CIF, where color intensity represents generation frequency. Grey cells indicate elements absent in generated structures.

3 Discussion

Altogether, employing a comprehensive pretraining-finetuning strategy, we demonstrate the development of domain-adapted foundational language models for materials science. A comprehensive evaluation of LLAMAT on several tasks, including entity recognition, entity extraction, and information extraction from text and tables, demonstrates that strategic domain adaptation through CPT and targeted IFT can transform general-purpose language models into specialized scientific tools without compromising their foundational capabilities. Moreover, the fact that the present work relies on the smaller models of LLAMA family suggests that adapting smaller models toward a specific domain might be a more economical and practical solution than relying on general-purpose LLMs.

The LLAMAT-CIF models represent a particularly significant advance in materials structure prediction. While LLAMAT-3 excels in generating complex structures with higher atomic counts and near-zero relaxation energies, its lower generation stability compared to LLAMAT-2 (requiring 33,000 versus 13,000 attempts for 10,000 valid structures). The models’ demonstrated ability to implicitly learn realistic chemical constraints—evidenced by systematic trends in elemental compositions and crystal system preferences—suggests potential for accelerating materials discovery while maintaining physical and chemical validity.

A significant finding emerges in the differential performance between model generations. Despite LLAMAT-3’s superior baseline capabilities, LLAMAT-2 variants demonstrate enhanced adaptability across multiple tasks, particularly in tabular information extraction and crystal structure generation. This raises an interesting question about the ability of highly over-trained models, such as LLAMA-3 to adapt to a new domain through CPT [36, 37]. This observation, referred to as “adaptation rigidity”, reported for the first time to the best of the authors’ knowledge, challenges the conventional scaling assumptions in LLMs. We hypothesize that the loss landscape[49] in the local vicinity of the minima in over-trained LLAMA-3 models may have a notably different character in comparison to those of LLAMA-2. A more detailed investigation of this aspect could provide further insights into the domain adaptation of LLMs.

These advances have broader implications beyond materials science. The successful development of domain-adapted language models while maintaining general capabilities provides a blueprint for creating specialized scientific AI systems across disciplines. The phenomenon of “adaptation rigidity” suggests the need to reevaluate scaling strategies in domain-specific AI applications, potentially influencing the development trajectory of specialized language models across scientific domains.

However, several challenges require attention to realize the full potential of these models. The observed hyperparameter sensitivity in PEFT optimization[50] indicates the need for more robust finetuning methodologies. The models’ preferential generation of certain crystal systems suggests the importance of developing comprehensive materials space exploration strategies. Understanding the fundamental principles underlying the adaptation rigidity phenomenon could provide crucial insights for optimizing domain adaptation strategies in large language models.

Looking ahead, this work establishes a foundation for integrating AI systems into materials research workflows. The demonstrated capabilities in automated literature analysis, extraction, and crystal structure prediction suggest the potential for accelerating materials discovery pipelines [3]. Future development should focus on enhancing model robustness, expanding capabilities to broader materials science applications, and developing theoretical frameworks for understanding domain adaptation in LLMs. The insights gained from this study—toward developing foundational LLMs for materials—may inform fundamental principles for developing specialized AI systems across scientific domains, potentially transforming how we approach domain adaptation of large language models for scientific applications.

4 Methods

4.1 Dataset Preparation

4.1.1 Pretraining Dataset: R2CID

The performance of foundation models is fundamentally determined by their pretraining dataset composition, necessitating meticulous curation of the constituent data sources. Our pretraining dataset, designated R2CID, integrates three distinct components: scientific literature from materials research publications, a curated subset of RedPajama (the original pretraining corpus for LLAMA models), and crystallographic information files (CIF). The scientific literature provides comprehensive materials characterization and synthesis protocols, while the RedPajama subset help prevent the catastrophic forgetting of the English language processing capabilities. The CIF datasets provide information on crystal structures, including atomic positions, lattice parameters, and symmetry operations. This tripartite combination enabled continued pretraining to generate the LLAMAT models. The specific composition and characteristics of each dataset component are detailed below.

a. Research Papers. Our corpus comprises over 4 million peer-reviewed articles sourced from approximately 500 Elsevier [51] and 300 Springer [52] journals. Selection criteria included full-text accessibility in XML format for Elsevier publications and HTML format for Springer publications. Journal selection was made manually based on the relevance to the materials domain. Article acquisition utilized the CrossRef API [53] to extract Digital Object Identifiers (DOIs), facilitating subsequent retrieval of full-text content in publisher-specific formats.

b. RedPajama. The RedPajama dataset [54], which served as the primary training corpus for the LLAMAT-2 [30], encompasses diverse textual sources, including arXiv preprints, GitHub repositories, StackExchange discussions, Wikipedia articles, and sanitized Common Crawl data. To preserve the model’s foundational linguistic capabilities while preventing catastrophic forgetting, we extracted a representative subset of approximately 700 million tokens. This strategic sampling maintains the model’s general-purpose functionality while facilitating domain-specific knowledge acquisition.

c. Crystallographic Information Files. Despite the existence of multiple text-based crystal representations [18], crystallographic information files (CIF) remain the definitive standard for structural data derived from diffraction studies. These standardized files encode essential parameters, including unit cell dimensions, interaxial angles, space group symmetry operations, and atomic position coordinates. Our dataset incorporates 470,000 CIF files, augmented with natural language descriptions generated via RoboCrystallographer [55]. These files were aggregated from three major sources: the Materials Project [56], GNoME-based ab-initio configurations [57], and the American mineralogist crystal structure database (AMCSD) [58].

d. R2CID Dataset Integration. The integration protocol implemented a structured mixing strategy to optimize training efficiency and maintain model robustness. Research paper content was systematically interspersed with RedPajama text, maintaining a ratio of 2.4 million RedPajama tokens per 100 million research paper tokens. Crystallographic data integration occurred within the terminal 10% of the dataset, where CIF files and their descriptions were interleaved with research paper content.

4.1.2 Instruction Finetuning

The IFT protocol incorporated multiple specialized datasets encompassing materials science and general question-answering tasks. We developed two novel domain-specific datasets: MatBookQA, consisting of materials science questions and answers generated via GPT4 using contextual prompting, and a comprehensive question bank derived from the Graduate Aptitude Test in Engineering (GATE). GATE is a standardized examination for postgraduate admissions at premier Indian institutions and select international universities. The constituent datasets are detailed below.

a. OpenOrca. The OpenOrca corpus encompasses 800,000 high-fidelity instruction-response pairs spanning diverse technical domains. Previous investigations [32] have demonstrated that models finetuned on this dataset exhibit superior performance across multiple evaluation frameworks, including Big-Bench Hard and AGIEval. This enhanced performance manifests in improved technical comprehension, complex query resolution, and domain-appropriate response generation. Dataset optimization procedures were implemented to determine the optimal training sample size for our specific application (see App. D).

b. Mathematics Corpus (MathQA). To enhance the model’s quantitative reasoning capabilities, we incorporated 7,500 selected problems from the Math dataset [33]. This curated subset consists of advanced

competition-level mathematical problems chosen to develop robust problem-solving abilities across various mathematical domains.

c. Materials Science Instruction Sets (MatSciInstruct). The materials science instruction corpus integrates multiple specialized datasets, including a novel collection generated through GPT-4 (gpt-4-0613) using open-source materials science textbooks as source material. This approach generated contextually rich questions spanning diverse materials science subdomains. The corpus incorporates MatSciInstruct[35], which employs a two-phase development framework: an initial Generation phase utilizing an instructor model to create domain-specific instruction data, followed by a Verification phase wherein a distinct verifier model assesses instruction quality across multiple dimensions including accuracy, relevance, completeness, and logical consistency. The instruction set is further augmented with the MatSciNLP training corpus and our custom-developed MatBookQA dataset.

d. MatBookQA. The MatBookQA dataset was systematically developed using a comprehensive materials science textbook[59]. The development protocol employed chapter-wise GPT-4 prompting using twenty distinct prompt templates (detailed in Appendix G), equally divided between generating short and extended responses. This methodology yielded 2,069 question-answer pairs, comprising 1,887 concise responses and 182 comprehensive explanations.

e. Materials Science Question Answering (MaScQA). The MaScQA dataset encompasses 1,585 questions from Indian undergraduate engineering examinations, specifically 1,036 from civil engineering and 549 from chemical engineering curricula. Answer validation was performed using the GPT-4o model (2024-02-01), with only verified correct responses retained in the final dataset. As detailed in Zaki et al.[14], the question taxonomy includes four distinct categories: traditional multiple-choice, correlation-based matching, numerical multiple-choice, and open-ended numerical problems.

f. Crystallographic Information File (CIF) Dataset. To train the language models to generate crystals, we created a new set of tasks that enable the language models to train on various aspects of CIF. Specifically, we developed instruction-output pairs from CIF files sourced from AMCS, Google GNoME, and the Materials Project to enhance LLAMAT's crystallographic comprehension and natural language query resolution capabilities. To this extent, we developed an instruction set implementing a dual-task framework comprising syntactic and semantic components. Syntactic tasks address the structural interpretation of CIF files. In contrast, semantic tasks, inspired by Gruver et al. (2024)[9], focus on crystal stability principles, including elemental co-occurrence patterns, atomic spatial distributions, and stability-determining properties. This methodology generated approximately 7 million instruction-output pairs (6,941,865 training instances and 27,183 validation instances). The complete task framework, with corresponding system prompts detailed in Appendix H, encompasses:

Syntactic Analysis Tasks:

- Atomic frequency quantification within crystal structures.
- Spatial coordinate-based atomic identification.
- Crystal parameter determination: dimensional analysis, volumetric calculation, and space group classification.
- Site occupancy equivalence evaluation.
- Structure-based chemical formula derivation.

Semantic Analysis Tasks:

- Property-conditioned crystal structure generation.
- Positional atomic prediction using MASK token methodology.
- Structural dimension prediction for stability optimization.
- Element-constrained crystal structure synthesis.

4.1.3 Materials Natural Language Processing (MatNLP)

The model evaluation employed a comprehensive dual-stage assessment protocol encompassing both materials science and general language capabilities. The primary evaluation phase compared multiple model iterations to optimize architectural decisions, while the secondary phase benchmarked performance against contemporary state-of-the-art materials science models. The primary evaluation corpus comprised 14 specialized materials

science tasks, supplemented with four general-purpose reasoning and comprehension assessments to preserve broad linguistic capabilities.

Table A.2 and App. B delineate the task taxonomy, dataset specifications, and sample distribution across training and validation sets. The evaluation framework encompasses multiple task categories, namely, sentence classification (SC), relation extraction (RE), named entity extraction (NER), synthesis action retrieval (SAR), paragraph classification (PC), entity extraction (EE), slot filling (SF), question answering (Q&A), and multiple choice question answering (MCQ). Detailed task specifications are documented in App. B and Ref. [34]. Model evaluation incorporated single-epoch fine-tuning on the training corpus prior to validation assessment to ensure instruction comprehension.

The secondary evaluation phase utilized the MatSciNLP dataset [60], which reformulates these tasks as multi-class classification problems. This meta-dataset enables direct performance comparison with existing materials science language models. To maintain evaluation integrity, distinct model instances were trained for each evaluation phase due to potential dataset overlap. Performance assessment followed the methodology established in Ref. [35], implementing single-epoch training on a condensed training set followed by evaluation on a comprehensive 170,000 sample validation corpus. Task-specific examples are provided in Appendix I.

4.1.4 Structured Information Extraction Dataset (**MatSIE**)

The extraction of structured information facilitates automated data processing and machine-readable format conversion. Given the domain expertise and structured data comprehension acquired through instruction fine-tuning, LLAMAT models were hypothesized to demonstrate robust performance in structured extraction tasks. To further analyze this capability, we performed the evaluation of the models using instruction-output pairs derived from four specialized datasets: (i) Doping, (ii) General materials, (iii) metal-organic frameworks (MOF) [24], and (iv) DiSCoMAT [39].

The initial three datasets focus on entity recognition and relationship extraction within materials science texts. The DiSCoMAT dataset provides annotated tables extracted from materials science publications. For the entity-relationship datasets, we developed six system prompts serving as prefixes to query-response pairs, where responses conform to standardized JSON schemas as established in Ref. [24] (see App. F). The DiSCoMAT dataset, originally developed for alternative applications, was transformed to generate JSON-structured annotations suitable for the language models (format specifications in App. F).

4.2 Model Development Methodology

4.2.1 Continued Pretraining

The pretraining corpus underwent hierarchical prioritization based on materials science relevance ($P1 > P2 > P3$). This corpus integrated materials science community discourse data and incorporated RedPajama subset to mitigate catastrophic forgetting, supplemented with 470,000 crystallographic information files for structural comprehension. The integration methodology employed a dual-phase mixing strategy:

- Primary phase: 90% of P1 content integrated with P2 and P3 datasets through stochastic shuffling.
- Secondary phase: Remaining 10% of P1 content combined with the CIF dataset through stochastic shuffling.

The resultant dataset underwent final integration with RedPajama using a token-ratio methodology: approximately 0.15M RedPajama tokens per 5M materials science tokens. The details of the pretraining dataset, along with the number of tokens, are mentioned in D.1.

4.2.2 Instruction Finetuning

The LLAMAT-Chat models, initialized with corresponding LLAMAT model weights, underwent tri-phase instruction finetuning:

- **Phase I:** Single-epoch finetuning on OpenOrca dataset to establish general instruction-following capabilities
- **Phase II:** Three-epoch finetuning on mathematical questions, optimizing quantitative reasoning capabilities. The limited dataset size enabled the observation of continuous validation loss reduction.

- **Phase III:** Single-epoch finetuning on an integrated corpus comprising MatSciInstruct, MatSciNLP, MatBookQA, and MaScQA, focusing on materials science-specific instruction comprehension

Implementation utilized the Megatron-LLM framework with learning rate initialization at 2×10^{-6} , scaling to 2×10^{-5} over initial 10% iterations, followed by cosine decay. This protocol was replicated for LLAMAT-2 and LLAMAT-3 chat model development.

4.2.3 Task Finetuning

a. LLaMat-Chat. The final development phase incorporated combined training on the training set of MatNLP and MatSIE datasets. This phase employed a 10^{-5} learning rate with cosine decay over two epochs. The intention of this stage was to familiarize the LLAMAT-Chat models with a wide range of tasks relevant to materials research, including scientific natural language processing, structured information extraction, and tabular information extraction. All the training data from these datasets were mixed to form a single task dataset on which the LLAMAT-Chat models were finetuned.

b. LLaMat-CIF. Crystal structure generation capabilities were implemented through parameter-efficient finetuning [9]. Optimal LLAMAT checkpoints underwent instruction finetuning using the dataset detailed in Section 4.1.2, with model selection based on minimal validation loss (Fig. C.2). Comprehensive finetuning specifications and hardware configurations are documented in Sections C.2 and C.3.

4.3 Baselines

In order to compare the performance of LLAMAT with existing general-purpose models, we considered LLAMA, Gemini-1.5 Flash-8B, and Claude-3 Haiku. Note that these models were chosen as they were the closest comparable ones in the respective families with LLAMAT models in terms of the number of parameters. To assess the effect of finetuning, LLAMA models were evaluated both with and without finetuning (FT).

4.4 Evaluation Metrics

a. Loss function. The loss function used to train the models for CPT, IFT, and task finetuning is the cross-entropy loss.

b. MatNLP and MatSIE. To evaluate the performance of models on the downstream tasks in MatNLP and MatSIE, precision, recall, and F1 scores are used with the annotated data as the ground truth.

c. Crystal generation. To evaluate the performance of LLMs for crystal generation, we rely on the following metrics.

1. Validity check: Structural validity and compositional validity are calculated as described in [40]. The former indicates that the distance between the centres of two atoms is greater than the sum of their atomic radii. The compositional validity is obtained using SMACT[61], which identifies if the given material is charge neutral based on all possible charge combinations.
2. Coverage: We use two coverage metrics, COV-R (recall) and COV-P (precision), described in [40] to measure the similarity between ensembles of generated materials and ground truth materials in the test set. COV-R Measures the percentage of ground truth materials being correctly predicted, and COV-P measures the percentage of predicted materials having high quality as described in [40]
3. Property statistics: We compute the Wasserstein distance between the property distributions of the generated materials and the test materials. We use density (in g/ cm³) and number of unique elements (#elem) as the properties.
4. Stability Check: We used M3GNet ([62]) to approximate force, energy, and stress in crystal unit cells. We use the predicted energy of the final structure as our stability metric since those having low predicted absolute energy (< 0.1 eV/atom \hat{E}_{hull}) are likely to be stable. While other potentials could be used, we relied on M3GNet to ensure direct comparison with the baselines.

5 Code availability

Codes used in this work are shared in the LLAMAT GitHub repository: <https://github.com/M3RG-IIID/llamat>.

Acknowledgments

N. M. A. K. acknowledges the funding support received from BRNS YSRA (53/20/01/2021-BRNS), ISRO RESPOND as part of the STC at IIT Delhi, Google Research Scholar Award, Intel Labs, and Alexander von Humboldt Foundation. M. acknowledges grants by Google, IBM, Microsoft, Wipro, and a Jai Gupta Chair Fellowship. M. Z. acknowledges the funding received from the PMRF award by the Ministry of Education, Government of India. The authors thank Microsoft Accelerate Foundation Models Research (AFMR) for access to OpenAI models. The authors thank the High-Performance Computing (HPC) facility at IIT Delhi for computational and storage resources. This work was partially supported by the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Programme at the University of Edinburgh. The EIDF provided access to Cerebras CS2 clusters for training the language models.

References

- [1] NM Anoop Krishnan, Hariprasad Kodamana, and Ravinder Bhattoo. *Machine Learning for Materials Discovery: Numerical Recipes and Practical Applications*. Springer Nature, 2024.
- [2] Vineeth Venugopal and Elsa Olivetti. MatKG: An autonomously generated knowledge graph in Material Science. *Scientific Data*, 11(1):217, February 2024.
- [3] Santiago Miret and NM Krishnan. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.
- [4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [5] Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.
- [6] Mara Schilling-Wilhelmi, Martíño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T Koch, José A Márquez, and Kevin Maik Jablonka. From text to insight: large language models for materials science data extraction. *arXiv preprint arXiv:2407.16867*, 2024.
- [7] Sheshera Mysore, Zach Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. *arXiv preprint arXiv:1905.06939*, 2019.
- [8] Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, December 2024.
- [9] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.
- [10] Qianggang Ding, Santiago Miret, and Bang Liu. Matexpert: Decomposing materials discovery by mimicking human experts. In *AI for Accelerated Materials Design-NeurIPS 2024*.
- [11] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.
- [12] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024.
- [13] Malcolm Sim, Mohammad Ghazi Vakili, Felix Strieth-Kalthoff, Han Hao, Riley J Hickman, Santiago Miret, Sergio Pablo-García, and Alán Aspuru-Guzik. Chemos 2.0: An orchestration architecture for chemical self-driving laboratories. *Matter*, 7(9):2959–2977, 2024.
- [14] Mohd Zaki, NM Anoop Krishnan, et al. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024.
- [15] Andrew D White, Glen M Hocky, Heta A Gandhi, Mehrad Ansari, Sam Cox, Geemi P Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, et al. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2(2):368–376, 2023.
- [16] 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, February 2024.
- [17] Hasan M Sayeed, Wade Smallwood, Sterling G Baird, and Taylor D Sparks. Nlp meets materials science: Quantifying the presentation of materials data in literature. *Matter*, 7(3):723–727, 2024.
- [18] Nawaf Alampara, Santiago Miret, and Kevin Maik Jablonka. Mattext: Do language models need more than text & scale for materials modeling? In *AI for Accelerated Materials Design-Vienna 2024*, 2024.
- [19] Yoel Zimmermann, Adib Bazgir, Zartashia Afzal, Fariha Agbere, Qianxiang Ai, Nawaf Alampara, Alexander Al-Feghali, Mehrad Ansari, Dmytro Antypov, Amro Aswad, et al. Reflections from the 2024 large language model (llm) hackathon for applications in materials science and chemistry. *arXiv preprint arXiv:2411.15221*, 2024.
- [20] Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Benedict Emoekabu, Aswanth Krishnan, Mara Wilhelmi, Macjonathan Okereke, Juliane Eberhardt, Amir Mohammad Elahi, Maximilian Greiner, et al. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*, 2024.

- [21] Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135*, 2024.
- [22] Kausik Hira, Mohd Zaki, Dhruvil Sheth, NM Anoop Krishnan, et al. Reconstructing the materials tetrahedron: challenges in materials information extraction. *Digital Discovery*, 3(5):1021–1037, 2024.
- [23] Nawaf Alampara, Mara Schilling-Wilhelmi, Martíño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *arXiv preprint arXiv:2411.16955*, 2024.
- [24] 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.
- [25] Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. Honeybee: Progressive instruction finetuning of large language models for materials science. *arXiv preprint arXiv:2310.08511*, 2023.
- [26] Hasan M. Sayeed, Trupti Mohanty, and Taylor D. Sparks. Annotating Materials Science Text: A Semi-automated Approach for Crafting Outputs with Gemini Pro. *Integrating Materials and Manufacturing Innovation*, 13(2):445–452, June 2024.
- [27] Defne Circi, Ghazal Khalighinejad, Anlan Chen, Bhuwan Dhingra, and L Catherine Brinson. How well do large language models understand tables in materials science? *Integrating Materials and Manufacturing Innovation*, 13(3):669–687, 2024.
- [28] Sterling G Baird, Hasan M Sayeed, Joseph Montoya, and Taylor D Sparks. matbench-genmetrics: A python library for benchmarking crystal structure generative models using time-based splits of materials project structures. *Journal of Open Source Software*, 9(97):5618, 2024.
- [29] Kamal Choudhary. Atomgpt: Atomistic generative pretrained transformer for forward and inverse materials design. *The Journal of Physical Chemistry Letters*, 15(27):6909–6917, 2024.
- [30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [31] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [32] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.
- [33] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [34] Yu Song, Santiago Miret, and Bang Liu. MatSci-NLP: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3621–3639, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [35] Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. HoneyBee: Progressive instruction finetuning of large language models for materials science. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5724–5739, Singapore, December 2023. Association for Computational Linguistics.
- [36] Shamane Siriwardhana, Mark McQuade, Thomas Gauthier, Lucas Atkins, Fernando Fernandes Neto, Luke Meyers, Anneketh Vij, Tyler Odenthal, Charles Goddard, Mary MacCarthy, et al. Domain adaptation of llama3-70b-instruct through continual pre-training and model merging: A comprehensive evaluation. *arXiv preprint arXiv:2406.14971*, 2024.
- [37] Firat Öncel, Matthias Bethge, Beyza Ermis, Mirco Ravanelli, Cem Subakan, and Çağatay Yıldız. Adaptation odyssey in LLMs: Why does additional pretraining sometimes fail to improve? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19834–19843, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

- [38] Maciej P Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569, 2024.
- [39] Tanishq Gupta, Mohd Zaki, Devanshi Khatsuriya, Kausik Hira, N M Anoop Krishnan, and Mausam . DiSCoMaT: Distantly supervised composition extraction from tables in materials science articles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13465–13483, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [40] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022.
- [41] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.
- [42] Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Mikhail Galkin, Santiago Miret, and Siyamak Ravankhahsh. Symmcmd: Symmetry-preserving crystal generation with diffusion models. In *AI for Accelerated Materials Design-NeurIPS 2024*.
- [43] Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M Wood. FlowMM: Generating materials with riemannian flow matching. In *Forty-first International Conference on Machine Learning*, 2024.
- [44] Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, Matthew Spellings, Mikhail Galkin, and Santiago Miret. Matsciml: A broad, multi-task benchmark for solid-state materials modeling. *arXiv preprint arXiv:2309.05934*, 2023.
- [45] Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael Bronstein. A hitchhiker’s guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.
- [46] Santiago Miret, Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, and Matthew Spellings. The open matsci ML toolkit: A flexible framework for machine learning in materials science. *Transactions on Machine Learning Research*, 2023.
- [47] Vaibhav Bihani, Sajid Mannan, Utkarsh Pratiush, Tao Du, Zhimin Chen, Santiago Miret, Matthieu Micoulaut, Morten M Smedskjaer, Sayan Ranu, and NM Anoop Krishnan. Egraffbench: evaluation of equivariant graph neural network force fields for atomistic simulations. *Digital Discovery*, 3(4):759–768, 2024.
- [48] Ayan Sengupta, Vaibhav Seth, Arinjay Pathak, Natraj Raman, Sriram Gopalakrishnan, and Tanmoy Chakraborty. Robust and efficient fine-tuning of llms with bayesian reparameterization of low-rank adaptation. *arXiv preprint arXiv:2411.04358*, 2024.
- [49] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6391–6401, 2018.
- [50] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.
- [51] ScienceDirect.com | Science, health and medical journals, full text articles and books.
- [52] Springer Nature Developer Portal | APIs for Research Papers.
- [53] Isaac Farley. Documentation.
- [54] togethercomputer/RedPajama-Data-1T · Datasets at Hugging Face, July 2024.
- [55] Alex M. Ganose and Anubhav Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, 2019.
- [56] Anubhav Jain, Joseph Montoya, Shyam Dwaraknath, Nils ER Zimmermann, John Dagdelen, Matthew Horton, Patrick Huck, Donny Winston, Shreyas Cholia, Shyue Ping Ong, et al. The materials project: Accelerating materials design through theory-driven data and tools. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pages 1751–1784, 2020.
- [57] 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.

- [58] American Mineralogist Crystal Structure Database.
- [59] Sabar D. Hutagalung. *Materials Science and Technology*. IntechOpen, Rijeka, Mar 2012.
- [60] Yu Song, Santiago Miret, and Bang Liu. Matsci-nlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. *arXiv preprint arXiv:2305.08264*, 2023.
- [61] Daniel W Davies, Keith T Butler, Adam J Jackson, Jonathan M Skelton, Kazuki Morita, and Aron Walsh. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019.
- [62] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.
- [63] Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Maruscyk, and Lukas Lange. The sofc-exp corpus and neural approaches to information extraction in the materials science domain. *arXiv preprint arXiv:2006.03039*, 2020.
- [64] Kyosuke Yamaguchi, Ryoji Asahi, and Yutaka Sasaki. Sc-comics: A superconductivity corpus for materials informatics. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6753–6760, 2020.
- [65] Vineeth Venugopal, Sourav Sahoo, Mohd Zaki, Manish Agarwal, Nitya Nand Gosvami, and NM Anoop Krishnan. Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, 2(7), 2021.
- [66] Zheren Wang, Kevin Cruse, Yuxing Fei, Ann Chia, Yan Zeng, Haoyan Huo, Tanjin He, Bowen Deng, Olga Kononova, and Gerbrand Ceder. Ulsa: unified language of synthesis actions for the representation of inorganic synthesis protocols. *Digital Discovery*, 1(3):313–324, 2022.
- [67] Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702, 2019.
- [68] Ankan Mullick, Akash Ghosh, G Sai Chaitanya, Samir Ghui, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Computational Materials Science*, 233:112659, 2024.
- [69] Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, Joel Hestness, et al. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023.

Appendices

A Dataset details

A.1 Pretraining and IFT dataset

Table A.1 contains details about the datasets we used for pretraining, followed by instruction finetuning to infuse the materials domain knowledge to the model while also giving our model the capability to follow instructions and answer queries through chat.

Table A.1: details about Instruction finetuning and pretraining datasets. for more detailed info, see Sec. 4.2.2

Pretraining Dataset	Token Length		
Elsevier/Springer	30B	-	Tokens sourced from material science research papers on Elsevier and Springer.
RedPajama	300M	-	A part of the Original Llama-2 corpus. We interleave this at regular intervals in the pretraining corpus: 10M research paper tokens followed by 0.1M RedPajama tokens.
Mat Sci Community Discourse	30M	-	Tokens sourced from MSCD, which is a forum for questions and answers for material science.
<hr/>			
IFT Dataset	Train Size	Val Size	Description
OpenOrca	576,000	-	The standard instruction finetuning dataset. A subset of the FLAN dataset is augmented with answers from GPT-4. It contains generic instructions following tasks.
MathQA	7500	5000	Contains numerical math questions. We train on this dataset to improve the mathematical ability of our model.
MatSciInstruct	52658	-	A collection of NLP tasks in the material science domain generated using ChatGPT, Claude, and GPT-4[35]
MatSciNLP	19942	170594	A collection of NLP tasks in the material science domain
MatBookQA	150 + 1800	32 + 87	Long and short questions and answers generated by GPT-4 on chapters of an open-source material science book.
MaScQA × 4	1022 × 4	1022 × 4	comprises 1036 and 549 questions from the civil and chemical engineering undergraduate-level exams in India, respectively. Only those questions that were answered correctly by GPT4 are taken; the total count is, hence, 1022.
Crystal finetuning Dataset	6,941,865	27,183	Semantic and syntactic instruction-output pairs based on CIF files. Details are provided in Appendix H

A.2 MatNLP, MatSIE, and Crystal generation datasets

Table A.2 contains details about the individual datasets and tasks used for training and evaluating the models. Figure C.2 shows the distribution of Bravais lattice on the CIF dataset used to train LLAMAT.

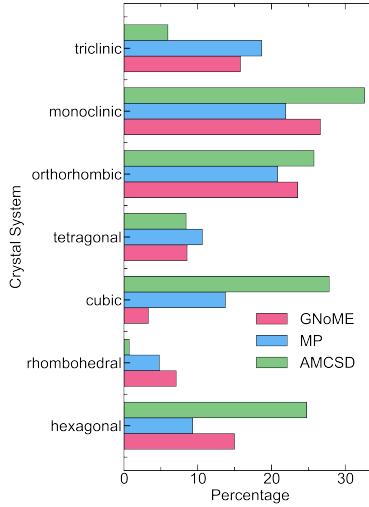


Figure A.1: Distribution of the Bravais lattice of CIF training dataset.

Table A.2: Task Descriptions and evaluation dataset sizes. For a detailed description of each task type, see B

Task	Dataset	Train Size	Eval Size	Task Description
MatNLP				
Entity recognition				
Matscholar	Matscholar	1062	1061	Named entity recognition tasks over data taken from matscholar.
SOFC-1	sofc-token	175	177	Named entity recognition over sentences from a corpus with data pertaining to "solid oxide fuel cells" [63]
SOFC-2	sofc-token	175	179	Identify slot fillers from sentences using a predefined set of semantically meaningful entities. Each sentence describes an experiment frame.
SC-CoMics-1	sc-comics	937	936	Named entity recognition over sentences from a corpus on "superconductivity" [64].
Classification				
Glass	glass-non-glass	300	299	Paragraph classification: Determine whether a given paragraph pertains to glass science. This task is adapted from [65]
Synthesis Actions	SAR	565	569	Classify word tokens into one of eight predefined synthesis action categories. SAR data adapted from [66]
SOFC-3	sofc-sent	1893	1889	Sentence classification: Identify sentences that describe relevant experimental facts. The task data is adapted from [63]
Extraction				
SC-CoMics-2	sc-comics	287	288	Extract event arguments and their roles based on specified event triggers.
SC-CoMics-3	sc-comics	376	373	Predict the most relevant relation type for a given span pair.
MatSci	structured-re	1788	1786	Predict the most relevant relation type for a given span pair.
English				
Q&A	squad	1042	1042	English questions and answers based on reading comprehension.
MCQ	hellaswag	981	980	English tasks on multiple choice question answering based on common sense.
MCQ	boolqa	500	499	Dataset with naturally occurring yes/no questions.
MCQ	story-cloze	500	501	MCQ for common-sense evaluation for story understanding and generation. Choose the correct ending for a 4-sentence story.
SIE Doping				
NER	basemats	322	59	Entity recognition of the base material used in a sentence referencing the use of doping.
NER	dopants	385	66	Entity recognition of the dopant used in a sentence referencing the use of doping.
RE	triplets	327	62	Relation extraction between base materials and dopants.
SIE General				
NER	acronym	45	13	Entity recognition of the acronym for a material used in the input.
NER	applications	443	53	Entity recognition of the applications for material in the input.
NER	name	216	34	Entity recognition of the name of a material in the input.
NER	formula	417	63	Entity recognition of the formula of a material in the input.
NER	structure or phase	325	47	Entity recognition of the structure or phase of a material in the input.
NER	description	358	49	Entity recognition of the description of a material in the input.
RE	formula-name	103	8	Relation extraction to get which formula corresponds to which material name in the input.
RE	formula-structure/phase	427	52	Relation extraction to get which material formula corresponds to which structure/phase description in the input.
RE	formula-application	811	56	Relation extraction to get which material formula in the input corresponds to which applications.
RE	formula-description	399	41	Relation extraction to get which material formula in the input corresponds to which description.
SIE MOFs				
NER	name of MOF	511	65	Entity recognition of the name for a MOF material in the input.
NER	MOF formula	100	16	Entity recognition of a MOF formula for a material in the input.
NER	MOF description	267	22	Entity recognition of description for a MOF material in the input.
NER	guest species	201	26	Entity recognition of guest species for MOF material mentioned in the input.
NER	applications	1024	128	Entity recognition of applications for a MOF material mentioned in the input.
RE	name-guest species	255	34	Relation extraction of name and guest species mentioned in the input.
RE	name-application	1004	137	Relation extraction of name and applications mentioned in the input.
RE	name-description	168	16	Relation extraction of name and description mentioned in the input.
DiSCoMaT				
Table	comptable			Detect whether the input table has material compositions.
Table	regex			Detect whether compositions are extractable using a regular expression parser.
Table	gid	5146	737	Detect which column/row is a material identifier present in.
Table	composition			Identify all columns/rows containing complete material composition information.
Table	chemical			Identify all columns/rows reporting values of constituent chemicals of the material.

B Task category description

Table B.1: Descriptions of NLP tasks in the MatNLP dataset, with task data adapted from various sources [34]

Task Type	Description
Named Entity Recognition (NER)	The NER task requires models to extract summary-level information from materials science text and recognize entities, including materials, descriptors, material properties, and applications, among others. Identify the best entity type label for a given text span, including handling non-entity spans with a “null” label. NER task data in downstream tasks is adapted from [67, 63, 7, 64]
Relation Extraction (RE)	Predict the most relevant relation type for a given span pair (e.g., s_i , s_j). MatSci-NLP contains relation classification task data adapted from [7, 64, 68].
Event Argument Extraction (EE)	Extract event arguments and their roles based on specified event triggers, accounting for potential multiple events in a given text. MatSci-NLP task data is adapted from [7, 64]
Paragraph Classification (PC)	Determine whether a given paragraph pertains to glass science. This task is adapted from [65]
Synthesis Action Retrieval (SAR)	Classify word tokens into one of eight predefined synthesis action categories. SAR data in MatSci-NLP is adapted from [66]
Sentence Classification (SC)	Identifying sentences that describe relevant experimental facts. The task data is adapted from [63]
Slot Filling (SF)	Extract slot fillers from sentences using a predefined set of semantically meaningful entities. Each sentence describes an experiment frame, and the model predicts slots for that frame. Task data is adapted from [63]

C Hyperparameter optimization

C.1 Pretraining

The pretraining to obtain LLAMAT-2 and LLAMAT-3 models was performed for 14369 and 13812 steps, respectively. The details of learning rates, warmup ratio, epochs, and the learning rate scheduler are mentioned in Table C.1. Considering the stability of the LLAMAT-2 model from the loss curve shown in Fig. C.1, we took the last checkpoint for further evaluation. In the case of LLAMAT-3, we evaluated intermediate checkpoints to arrive at the final model for downstream evaluation and development of the chat model. The results in table C.2 calculated for LLAMA-3 were computed just after CPT and before any instruction-finetuning for chat capabilities was done. This experiment informed that the last checkpoint of LLAMA-3, i.e., after 13812 steps, is the best one, and hence, we chose it as our base LLAMA-3 model.

Table C.1: Hyperparameter details for pretraining of LLAMaT-2 and LLAMaT-3

Hyperparameters	LLaMat-2	LLaMA-3
max_lr	3e-04	7e-05
warmup_ratio	0.1	0.1
min_lr	3e-05	7e-06
epoch	1	1
scheduler	cosine	cosine

Table C.2: Results on downstream dataset after direct finetuning of the pretrained models

Model	MatNLP-Micro-F1	MatNLP-Macro-F1	English-Micro-F1	English-Macro-F1
4k	89.035	82.57	84.54	79.93
8k	88.731	82.91	83.015	78.38
13k	89.595	84.349	84.707	80.282
13812	90.02	84.752	84.06	79.547

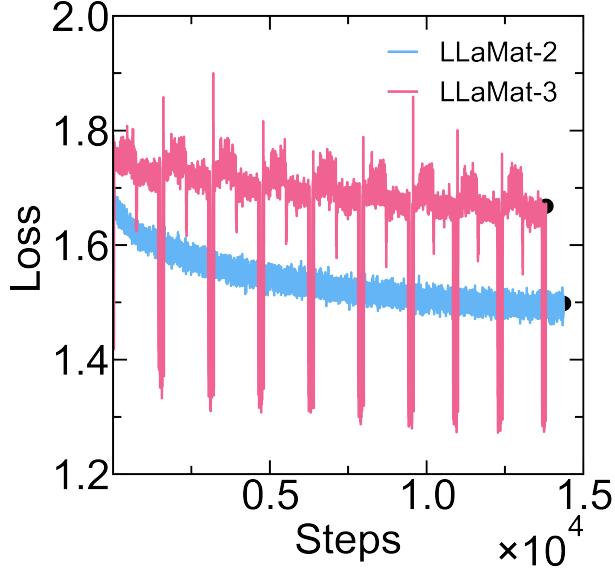


Figure C.1: Loss Curve for pretraining

C.2 Finetuning

This section shows the loss curves obtained after CIF-IFT of LLAMAT on the CIF-IFT dataset. It can be seen in Figure C.2 a and b that the minimum validation loss occurred at 17000 and 15000 steps, respectively. These models were further used to perform the parameter efficient finetuning to evaluate the performance of the crystal generator on the unconditional crystal structure generation task [9].

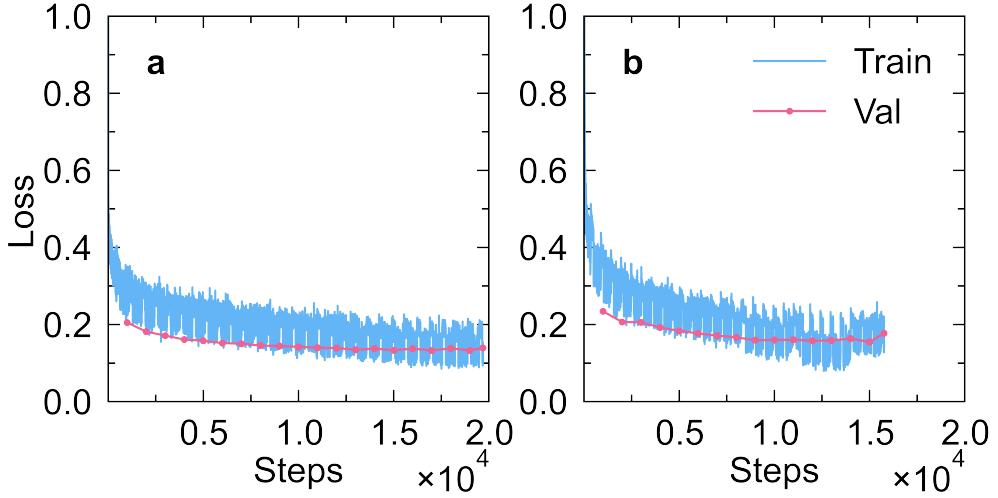


Figure C.2: Visualizing the loss curves of (a) LLAMAT-2-CIF and (b) LLAMAT-3-CIF models

C.3 Hardware setup and training time

The training times and hardware setup for each task are as follows :

- Pretraining LLAMAT-2: 8 NVIDIA A100 80GB GPUs for ~17 days
- Pretraining LLAMAT-3: 2 CS2 Cerebras Wafer Scale Cluster for ~3 days
- LLAMAT-IE-Copilot (see 4.2.2)
 1. Instruction fine tuning (stage 1): ~8 hours on 8 NVIDIA-A100 80GB GPUs.
 2. Instruction fine tuning (stage 2): ~1 hour 30 minutes on 8 NVIDIA-A100 80GB GPUs.
 3. Task finetuning (stage 3): 1 hour 10 minutes on NVIDIA-A100 80GB GPUs.
- LLAMAT-CIF: 2 CS2 Cerebras Wafer Scale Cluster for ~3 days

For continuous pretraining of LLAMA-2 models, we have used 8 NVIDIA A100 80GB GPUs as mentioned above. Since the dataset size and number of parameters are quite large, we use a distributed training methodology to efficiently utilize the storage and computational resources. Table C.3 lists our experiments to obtain optimal levels of data (DP), tensor (TP), and pipeline parallelisms (PP). We achieved the best token consumption rate of 27.1k tokens/second by considering PP=4, TP=1, and DP=2. Based on these experiments, we can also state that TP was less effective in our case than DP and PP.

In the case of LLAMA-3 pretraining, we used 2 CS2 Cerebras Wafer Scale Cluster. Here, we did not require parallelism as used in GPUs because of the linear scaling of the compute performance with change in the number of accelerators[69]. During the pretraining, we used batch_size of 960 and micro_batch_size of 80 as suggested by the training script provided at Cerebras Model Zoo on GitHub.

Nodes x GPUs	DP	TP	PP	tokens/s (k)	tokens/s/gpu (k)
1x2	1	1	2	7.5	3.75
1x2	1	2	1	4.8	2.4
1x2	2	1	1	OOM	OOM
1x4	1	1	4	12	3
1x4	1	1	1	5.6	1.4
1x4	4	1	1	OOM	OOM
1x4	2	2	1	9.4	2.35
1x4	2	1	2	13.8	3.45
1x4	1	2	2	12.5	3.125
1x8	1	1	8	22.9	2.8625
1x8	1	1	1	5	0.625
1x8	8	1	1	OOM	OOM
1x8	2	2	2	17.5	2.1875
1x8	1	2	4	14.2	1.775
1x8	1	1	4	9.9	1.2375
1x8	2	4	2	23.4	2.925
1x8	1	2	4	13	1.625
1x8	2	4	1	14.9	1.8625
1x8	4	2	1	21.4	2.675
1x8	2	4	1	27.1	3.3875

Table C.3: GPU Performance Metrics. OOM stands for Out-Of Memory error.

D Dataset distribution optimization

D.1 Pretraining

Table D.1: Details of pretraining datasets for obtaining LLAMAT-2 and LLAMAT-3

Dataset	# samples		# tokens (LLAMA-2)		# tokens (LLAMA-3)	
	train	val	train	val	train	val
P1	2,686,786		18,872,303,847			
P2	1,055,330	106,395	9,050,611,308	413,927,438	7,831,900,364	442,507,226
P3	225,634		1,864,471,418		1,589,414,318	-
MSCD	36,875		5,975,502	0	5,212,659	0
RedPajama	651,356	279,158	962,319,047	414,815,173	805,636,840	347,375,685
CIF	470,222	9,598	788,427,184	16,124,004	633,237,003	12,947,445
Total	5,126,203	395,151	31,544,108,306	844,866,615	10,865,401,184	802,830,356

D.2 Finetuning

The first step in instruction finetuning our models is training on OpenOrca, a general instruction finetuning dataset. We trained the model for different steps between 0-800k, then finetuned it on the downstream dataset again before evaluation.

Table D.2 shows the results on LLAMAT-3 and LLAMAT-2. We observed that LLAMAT-2’s English capability increases with more steps in general, while for LLAMAT-3, there is no such observation. Also, LLAMAT-3’s score in MatNLP is lower than its score at 0 steps. This could be because OpenOrca is a general-purpose IFT dataset unrelated to our downstream tasks. Since LLAMAT-3 already had a high score on both English and MatNLP initially, we don’t notice a significant further increase. From the results of LLAMAT-3 D.2, we decided to fix 576k training samples for Open-Orca instruction finetuning for LLAMAT-3, and 448k training samples for LLAMAT-2. Further IFT processes are described in the methodology section.

We also conducted experiments with different training samples for the MathQA and honeybee datasets for LLAMAT-2 .

Table D.3: Results for training with MathQA and different sample size of honeybee dataset on downstream evaluation

Model	Pretrain	OpenOrca	MathQA	Honeybee	MicroF1-MatNLP	MacroF1-MatNLP	MicroF1-English	MacroF1-English
LLAMA-2					84.24	77.75	80.63	77.01
LLAMAT	10B	0	0	0	85.43	79.68	78.8	75.33
LLAMAT	30B	0	0	0	87.85	82.26	82.23	78.73
LLAMAT	30B	448k	0	0	89.51	84.66	83.69	80.04
LLAMAT	30B	448k	0	32k	88.52	83.24	83.25	79.48
LLAMAT	30B	448k	0	48k	88.52	83.02	84.5	80.8
LLAMAT	30B	448k	0	96k	88.6	83.04	83.97	80.17
LLAMAT	30B	448k	0	144k	88.44	83.12	84.38	80.5
LLAMAT	30B	448k	7500*3	0	89.66	84.77	82.59	78.67
LLAMAT	30B	448k	7500*3	32k	87.89	82.27	83.56	79.66
LLAMAT	30B	448k	7500*3	48k	88.28	82.9	84.37	80.68
LLAMAT	30B	448k	7500*3	96k	88.04	82.39	84.17	80.25
LLAMAT	30B	448k	7500*3	144k	88.24	82.8	83.8	79.84

Table D.2: Performance of LLAMAT-2 and LLAMAT-3 on MatNLP and English validation sets after instruction-finetuning on Open-Orca dataset to varying degrees. The optimal dataset size is chosen based on the Pareto optimal performance on both MatNLP and Eng datasets.

Steps	MicroF1-MatNLP	MacroF1-MatNLP	MicroF1-Eng	MacroF1-Eng
LLaMat-2				
0k	87.85	82.26	82.23	78.73
64k	88.44	83.07	82.94	79.32
128k	88.72	83.35	83.31	79.47
192k	89.08	83.71	83.20	79.55
256k	89.34	84.09	83.60	79.79
320k	88.14	82.68	84.22	80.32
384k	88.48	83.54	84.05	80.43
448k	89.51	84.66	83.69	80.04
512k	89.07	83.96	84.04	80.30
576k	89.09	83.76	84.47	80.82
640k	88.60	83.30	84.95	81.30
768k	89.23	84.05	84.34	80.55
800k	88.48	83.12	85.02	81.22
LLaMat-3				
0k	89.70	83.71	84.56	80.24
64k	88.40	82.85	85.31	80.57
128k	86.39	80.29	83.63	79.24
192k	88.48	82.67	84.20	79.38
256k	85.97	80.32	84.68	79.81
320k	88.03	82.10	85.10	80.49
384k	87.42	81.95	85.40	80.67
448k	86.85	81.64	85.06	80.25
512k	87.89	82.37	84.74	80.24
576k	88.79	83.09	84.74	80.01
640k	88.40	82.85	85.31	80.57
768k	86.96	81.27	85.50	80.63
800k	87.70	82.48	85.07	80.19

E Model Performance

Table E.1: F1-score results on all our datasets. SIE = Structured information extraction. FT = Finetuned

Downstream Micro-F1		LLaMat-3 chat		LLaMA-3 chat FT		LLaMA-3 FT		LLaMat-2 chat		LLaMat-2 FT	
Task	sub-dataset	LLaMat-3 chat	LLaMat-3	LLaMA-3 chat FT	LLaMA-3 FT	LLaMat-2 chat	LLaMat-2	LLaMat-2 chat FT	LLaMat-2 FT		
MatNLP											
Entity Recognition	Matscholar	86.97	81.72	83.4	81.13	81.74	81.46	80.63	80.75		
Entity Recognition	SOFC-1	89.73	87.16	89.45	88.06	89.3	87.59	86.27	88.42		
Entity Recognition	SOFC-2	83.28	77.95	76.79	77.37	81.98	81.26	80.56	80.14		
Entity Recognition	SC-CoMics-1	90.73	86.74	70.87	89.03	91.31	90.9	90.67	90.93		
Classification	Glass	92.56	90.89	89.67	85.11	93.67	91.33	92.89	91.56		
Classification	Synthesis Actions	96.68	95.33	94.9	95.75	96.44	96.1	96.16	96.04		
Classification	SOFC-3	93.24	92.52	92.68	92.09	93.85	93.26	93.19	93.2		
Entity Extraction	SC-CoMics-2	91.95	91.22	66.38	86.26	93.78	94.4	91.33	92.34		
Entity Extraction	SC-CoMics-3	98.8	92.58	83.01	93.3	100.0	99.84	99.92	99.92		
Entity Extraction	MatSci	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0		
SIE Doping											
NER	basemats	0.889	0.895	0.854	0.912	0.884	0.880	0.904	0.910		
NER	dopants	0.952	0.869	0.825	0.884	0.870	0.840	0.865	0.855		
RE	triplets	0.846	0.794	0.769	0.816	0.785	0.785	0.772	0.753		
all	exact-match	0.742	0.656	0.613	0.694	0.629	0.597	0.629	0.629		
SIE General											
NER	acronym	0.353	0.500	0.522	0.222	0.444	0.600	0.500	0.600		
NER	applications	0.577	0.583	0.596	0.490	0.754	0.702	0.724	0.697		
NER	name	0.475	0.595	0.600	0.535	0.624	0.720	0.674	0.607		
NER	formula	0.623	0.584	0.702	0.640	0.755	0.667	0.634	0.751		
NER	structure or phase	0.372	0.507	0.275	0.408	0.661	0.724	0.655	0.606		
NER	description	0.404	0.304	0.330	0.404	0.383	0.469	0.375	0.433		
RE	formula-name	0.308	0.316	0.5	0.16	0.435	0.267	0.296	0.286		
RE	formula-structure/phase	0.233	0.312	0.268	0.247	0.429	0.412	0.326	0.413		
RE	formula-application	0.584	0.615	0.483	0.491	0.673	0.606	0.66	0.684		
RE	formula-description	0.325	0.241	0.214	0.341	0.344	0.323	0.26	0.238		
SIE MOFs											
NER	name of mof	0.678	0.822	0.661	0.746	0.735	0.752	0.717	0.794		
NER	mof formula	0.500	0.374	0.421	0.622	0.686	0.701	0.720	0.735		
NER	mof description	0.364	0.288	0.327	0.347	0.581	0.503	0.471	0.359		
NER	guest species	0.294	0.242	0.421	0.410	0.432	0.486	0.514	0.500		
NER	applications	0.675	0.653	0.638	0.640	0.671	0.668	0.646	0.640		
NER	exact match	0.098	0.137	0.157	0.137	0.098	0.098	0.078	0.098		
RE	name-guest species	0.2	0.154	0.28	0.311	0.269	0.356	0.298	0.269		
RE	name-application	0.36	0.36	0.334	0.368	0.422	0.403	0.355	0.454		
RE	name-description	0.113	0.074	0.129	0.421	0.475	0.449	0.457	0.26		
DiSCoMaT											
table	comptable	0.834	0.825	0.775	0.752	0.812	0.809	0.794	0.794		
table	regex	0.848	0.834	0.795	0.753	0.879	0.876	0.858	0.866		
table	gid	0.734	0.770	0.741	0.752	0.804	0.784	0.747	0.819		
table	composition	0.280	0.347	0.242	0.450	0.677	0.648	0.615	0.642		
table	chemical	0.582	0.626	0.543	0.636	0.673	0.656	0.623	0.642		
all	exact_match	396/598	393/585	349/520	376/568	557/701	551/700	539/710	548/705		

F DiSCoMat instruction and JSON Schema

We give the following instructions to the model before providing the question and table from which to answer. It includes the JSON schema of the output format in the form of a dictionary containing non-empty lists. The definition for each entry of the dictionary is also passed to the model.

Prompt:

```
You are an expert in materials science and extracting data from tables. You have to fill the following dictionary for the given table. Each key is defined as follows:
'comp_table'- If the input table has material compositions then return [1], else [0];
'regex_table'- If the input table has material compositions and they can be extracted using a regular expression parser, then return [1], else [0].
'composition_row_index'-The list containing the index of rows which have complete information about material composition.
'chemical_col_index'-The list containing the index of columns which report values of constituent chemicals of the material.
'composition_col_index'-The list containing the index of columns which have complete information about material composition.
'chemical_row_index'-The list containing the index of rows which report values of constituent chemicals of the material.
```

```
'gid_row_index'-The index of row having material identifier.  
'gid_col_index'-The index of column having material identifier.
```

```
dictionary =  
{'comp_table': [],  
'regex_table': [],  
'composition_row_index': [],  
'composition_col_index': [],  
'chemical_row_index': [],  
'chemical_col_index': [],  
'gid_row_index': [],  
'gid_col_index': []}
```

NOTE:The output will be the dictionary with keys having non-empty lists ONLY.

G Prompts for MatBookQA

Short Prompts

- You are a materials scientist. Use your expertise to generate concise answers to the following questions.
- As a materials scientist, provide short, precise answers to these questions.
- With your knowledge in materials science, answer the following questions succinctly.
- Given your background in materials science, provide brief, expert answers to these queries.
- Using your expertise in materials science, generate short answers for the following questions.
- Drawing from your experience in materials science, answer these questions with concise and accurate information.
- As an expert in materials science, provide quick, accurate answers to these questions.
- From your perspective as a materials scientist, generate short and precise answers to the following questions.
- Using your knowledge as a materials scientist, answer these questions briefly and accurately.
- Leverage your expertise in materials science to provide concise answers to these queries.

Long Prompts

- You are a materials scientist. Use your expertise in the field to generate detailed and comprehensive answers for the following questions.
- As a materials scientist, provide thorough and well-explained answers to these questions.
- With your knowledge in materials science, answer the following questions with detailed and extensive information.
- Given your background in materials science, provide long and comprehensive answers to these queries.
- Using your expertise in materials science, generate detailed and in-depth answers for the following questions.
- Drawing from your experience in materials science, answer these questions with elaborate and accurate information.
- As an expert in materials science, provide thorough and well-detailed answers to these questions.
- From your perspective as a materials scientist, generate long and comprehensive answers to the following questions.
- Using your knowledge as a materials scientist, answer these questions in detail and with full explanations.
- Leverage your expertise in materials science to provide extensive and well-explained answers to these queries.

H CIF IFT prompts

H.1 Syntactic tasks

- You are a Material Science expert who works with crystallographic files (CIF files). Use your understanding of the CIF file format to extract information about the unit cell structure.
- Utilize your expertise in Material Science to extract data regarding the unit cell structure from CIF files, drawing upon your comprehension of the file format.
- As a specialist in Material Science, employ your knowledge of CIF files to extract pertinent details concerning the unit cell structure.
- As a Material Science expert, utilize CIF file parsing to extract essential data regarding the unit cell configuration.
- Draw upon your Material Science expertise to extract unit cell structure information from CIF files, utilizing your understanding of the file format.
- Employ your understanding of Material Science and CIF file format to extract crucial information concerning the unit cell arrangement.
- As a specialist in Material Science, employ CIF file analysis to gather insights into the unit cell structure.
- Utilize your proficiency in Material Science to parse CIF files and extract relevant details regarding the unit cell configuration.
- Draw upon your expertise in Material Science to extract insights into the unit cell structure by analyzing CIF files.

H.2 Semantic tasks

H.2.1 Generative tasks

- You are a Material Science expert who works with crystallographic files (CIF files). Use your expertise to answer the following question related to the generation of stable materials when some information about it is described.
- Employ your expertise in Material Science, particularly in working with CIF files, to address the question concerning the creation of stable materials with partial descriptive information.
- Utilize your proficiency in Material Science and handling CIF files to provide insights into generating stable materials with limited descriptive data.
- Apply your knowledge as a Material Science specialist, specifically in manipulating CIF files, to respond to queries regarding the production of stable materials given incomplete information.
- Utilize your skills as a Material Science expert, with a focus on CIF files, to tackle the question concerning the development of stable materials based on partial descriptions.
- Employ your expertise in Material Science, particularly in the realm of CIF files, to address inquiries related to the creation of stable materials despite incomplete data.
- Utilize your proficiency in working with CIF files, as well as your background in Material Science, to answer questions regarding the generation of stable materials with limited descriptive details.
- Apply your knowledge and experience in Material Science, including your familiarity with CIF files, to provide solutions for generating stable materials when only partial information is available.
- Employ your specialized knowledge in Material Science, specifically your experience with CIF files, to tackle questions related to creating stable materials with partial information.
- Apply your skills as a Material Science expert, particularly in managing CIF files, to provide insights into generating stable materials despite incomplete descriptive data.

H.2.2 Infill tasks

- You are a Material Science expert who works with crystallographic files (CIF files). Use your expertise to answer the following question related to predicting the masked element in a CIF file.

- Utilize your expertise as a Material Science specialist, well-versed in CIF files, to address queries concerning the anticipation of the hidden element within a CIF file.
- Employ your proficiency in Material Science and crystallographic file analysis to tackle questions related to predicting the concealed element in a CIF file.
- Apply your knowledge in Material Science, particularly your experience with CIF files, to provide insights into predicting the masked element within a CIF file.
- Utilize your skills as a Material Science expert, specializing in CIF files, to offer solutions for predicting the undisclosed element in a CIF file.
- Employ your expertise in Material Science and crystallographic file manipulation to address questions concerning the forecast of the hidden element in a CIF file.
- Apply your specialized knowledge in Material Science, particularly your expertise with CIF files, to provide solutions for predicting the concealed element within a CIF file.
- Utilize your proficiency in crystallographic file analysis, coupled with your background in Material Science, to respond to questions regarding the prediction of the masked element in a CIF file.
- Apply your expertise in Material Science, particularly your familiarity with crystallographic files, to address inquiries concerning the prediction of the masked element in a CIF file.

Dimension task

- You are a Material Science expert who works with crystallographic files (CIF files). Use your expertise to answer the following question related to predicting the dimensions of a stable crystal conditioned on some information about the crystal.
- Utilize your expertise in Material Science and familiarity with CIF files to address the task of predicting the dimensions of a stable crystal based on the provided information.
- As a Material Science specialist working with CIF files, apply your knowledge to forecast the dimensions of a stable crystal given certain parameters.
- Employ your proficiency in crystallography and CIF file analysis to tackle the question of predicting the dimensions of a stable crystal conditioned on specific data.
- Utilize your expertise in Material Science and experience with CIF files to provide insights into predicting the dimensions of a stable crystal with given information.
- Apply your knowledge as a Material Science expert, particularly in working with CIF files, to answer questions related to predicting the dimensions of a stable crystal.
- Leverage your understanding of crystallographic principles and CIF files to address inquiries about predicting the dimensions of a stable crystal based on provided criteria.
- Utilize your expertise in Material Science, coupled with your familiarity with CIF files, to provide solutions for predicting the dimensions of a stable crystal conditioned on known parameters.
- Apply your knowledge as a Material Science specialist to analyze CIF files and predict the dimensions of a stable crystal given specific information.

Volume calculation task

- You are a Material Science expert who works with crystallographic files (CIF files). Use your expertise to compute the volume of a unit cell of the crystal described below.
- As a Material Science expert dealing with CIF files, please compute the unit cell volume for the given crystal.
- With your knowledge in Material Science and experience with crystallographic files, determine the volume of the crystal's unit cell.
- Given your background in Material Science and familiarity with CIF files, please find the volume of the described crystal's unit cell.
- As a Material Science specialist working with CIF files, calculate the volume of the unit cell of the provided crystal.

- With your proficiency in Material Science and crystallographic files, determine the unit cell volume for this crystal.
- Given your expertise in Material Science and knowledge of CIF files, compute the volume of the described crystal's unit cell.
- As an expert in Material Science and CIF files, calculate the unit cell volume for the given crystal.
- Using your Material Science and CIF file expertise, determine the volume of the unit cell of the crystal described.

I Examples

I.1 IFT Datasets

Example:OpenOrca

system: You are an AI assistant. You will be given a task. You must generate a detailed and long answer.,

question: Just when we thought Pharrell Williams couldn't get any cooler, he becomes the first guy to star in Chanel 's upcoming handbag campaign. Williams has already appeared in a short video by Karl Lagerfeld and even walked the Chanel runway, and now he's modelling the Gabrielle bag from the fashion house's spring collection. Set to join Kristen Stewart, Caroline de Maigret and Cara Delevingne in the campaign - with Lagerfeld taking on the role of photographer - Williams star turn is set to debut on Monday 3 April.

Based on that paragraph can we conclude that this sentence is true?
Lagerfeld will be taking pictures of the event where Pharrell along with three others in the fourth month of the year.

Choose from:

- Yes;
- It's impossible to say;
- No;

answer: Yes, based on the given paragraph, we can conclude that Lagerfeld will be taking pictures of the event where Pharrell along with three others in the fourth month of the year. The paragraph mentions that Lagerfeld will be the photographer for the upcoming Chanel handbag campaign, which will feature Pharrell Williams along with Kristen Stewart, Caroline de Maigret, and Cara Delevingne. The campaign is set to debut on Monday 3 April, which falls in the fourth month of the year. Therefore, the sentence is true.

Example:MathQA

```
"system": "Describe the logical steps needed to prove the following math theorem.",
```

```
"question": "What is  $\frac{1}{2} + \frac{1}{3}$ ? ",
```

```
"answer": "We want to have the least common denominator,  $2 \cdot 3 = 6$ , in the denominator when we add the two fractions. We write one half as  $\frac{1}{2} \cdot \frac{3}{3} = \frac{3}{6}$ . Also, we write one third as  $\frac{1}{3} \cdot \frac{2}{2} = \frac{2}{6}$ . Adding these, we obtain  $\frac{3}{6} + \frac{2}{6} = \frac{5}{6}$ , which is in the most simplified form. So,
```

$$\frac{1}{2} + \frac{1}{3} = \boxed{\frac{5}{6}}.$$

Example:MatSci-Instruct

system: You are a material scientist. Answer the following question based on the paragraph provided.

question: What are the two types of superconductors studied in the text?

We study the effects of quasidisorder and Anderson disorder on a two-dimensional topological superconductor with an applied external magnetic field. The cases of a *p*-wave superconductor and a noncentrosymmetric superconductor with mixed *p* and *s*-wave pairings and Rashba spin-orbit coupling are studied. We show that, for a perpendicular magnetic field, the introduction of quasidisorder leads to the appearance of topological phases in new regions, characterised by an integer value of the Chern number. For a parallel magnetic field, we identify regimes with the appearance of new Majorana flat bands and also new unidirectional Majorana edge states, as quasidisorder is introduced. We show that the Majorana flat bands have a quantized Berry phase of π and identify it as a topological invariant. Two topological transitions are identified and the values of the critical exponents z and ν are obtained. The fractal nature of the eigenstates is discussed both for Anderson disorder and Aubry-André disorder.

answer: The two types of superconductors studied in the text are a *p*-wave superconductor and a noncentrosymmetric superconductor with mixed *p* and *s*-wave pairings and Rashba spin-orbit coupling.

Example:MatSciNLP

system: Task: named entity recognition for doped. Request: You need to predict the entity type, and choices : b-cmt, i-cmt, b-mat, i-mat, b-dsc, b-pro, i-pro, i-dsc, b-smt, i-smt, b-apl, i-apl, b-spl, i-spl, b-material, b-device, i-device, b-experiment, b-value, i-value, i-material, i-experiment, meta, material, property-misc, nonrecipe-material, operation, number, amount-unit, brand, reference, characterization-apparatus, synthesis-apparatus, amount-misc, material-descriptor, property-unit, condition-unit, condition-misc, property-type, condition-type, apparatus-unit, apparatus-descriptor, apparatus-property-type, element, main, process, sc, characterization, property, value, doping

question: No appreciable distinction between undoped and O₃Y₂ - doped Ce - TZP powders was observed in either DTA or TGA thermograms .

answer: B-DSC

Example:MatBookQA

system: Using your expertise in materials science, generate short answers for the following questions.

question: Why is an X-ray microbeam technique used in the measurement system?

answer: An X-ray microbeam technique is used in the measurement system to achieve a smaller beam size, which allows for more precise targeting of the sample and higher spatial resolution in the measurements. This technique provides detailed structural information on a sub-micron scale, which is essential for studying the intricate changes in phase-change materials.

Example:MatBookQA

system: Solve the following question with highly detailed step by step explanation. Write the correct answer inside a dictionary at the end in the following format. The key 'answer' has a list which can be filled by all correct options or by a number as required while answering the question. For example for question with correct answer as option (a), return 'answer':[a] at the end of solution. For question with multiple options'a,c' as answers, return 'answer':[a,c]. And for question with numerical values as answer (say 1.33), return 'answer':[1.33]

input: Match the industrial processes in Group-1, with the catalyst materials in Group-2.

Group-1- P. Ethylene polymerisation Q. Petroleum feedstock cracking R. Oxidation of SO_2 to SO_3 S. Hydrogenation of oil

Group-2- I. Nickel II. Vanadium pentoxide III. Zeolite IV. Aluminium triethyl with titanium chloride promoter.

- (A) P-IV, Q-III, R-II, S-I (B) P-I, Q-IV, R-III, S-II
(C) P-I, Q-II, R-III, S-IV (D) P-II, Q-III, R-IV, S-I

output: To solve this matching problem, we need to correctly pair each industrial process in Group-1 with its corresponding catalyst material in Group-2. Let's analyze each process and its typical catalyst:

1. Ethylene polymerisation (P):

- This process typically uses a Ziegler-Natta catalyst, which is a combination of aluminium triethyl and titanium chloride.
- Therefore, the correct match is IV.

2. Petroleum feedstock cracking (Q):

- This process often uses zeolites as catalysts due to their ability to facilitate the breaking down of large hydrocarbon molecules.
- Therefore, the correct match is III.

3. Oxidation of SO_2 to SO_3 (R):

- This process is part of the contact process for producing sulfuric acid and typically uses vanadium pentoxide (V_2O_5) as the catalyst.
- Therefore, the correct match is II.

4. Hydrogenation of oil (S):

- This process typically uses nickel as a catalyst to hydrogenate unsaturated fats and oils.
- Therefore, the correct match is I.

Now, let's compile the matches:

- P (Ethylene polymerisation) - IV (Aluminium triethyl with titanium chloride promoter)
- Q (Petroleum feedstock cracking) - III (Zeolite)
- R (Oxidation of SO_2 to SO_3) - II (Vanadium pentoxide)
- S (Hydrogenation of oil) - I (Nickel)

The correct option that matches these pairs is:

- (A) P-IV, Q-III, R-II, S-I

Thus, the correct answer is:

{'answer': ['A']}

I.2 Downstream Datasets

Example: MatNLP

```
system: You are a linguist and a material scientist. You need to identify the named entity for each of the keywords given after WORDS in the input. Answer to the question should be from one of the provided options. Do not output anything else other than the answer. You should output the word entity pair separated by ":" in each line. Your options are: b-mat, i-mat, b-spl, i-spl, b-dsc, i-dsc, b-pro, i-pro, b-apl, i-apl, b-smt, i-smt, b-cmt, i-cmt. Answer for each word must be in a new line.
```

```
question: WORDS: x-ray, diffraction, XRD, transmittance, electron, microscopy, TEM, Zn, Ti, O, annealed, OZn, O2Ti, nanocomposites \n SENTENCE: moreover , x-ray diffraction ( XRD ) patterns and transmittance electron microscopy ( TEM ) analysis show that multiple Zn - Ti - O ternary compounds exist in the high - temperature annealed OZn - O2Ti nanocomposites.
```

```
answer: x-ray : b-cmt \n diffraction : i-cmt \n XRD : b-cmt \n transmittance : b-cmt \n electron : i-cmt \n microscopy : i-cmt \n TEM : b-cmt \n Zn : b-mat \n Ti : i-mat \n O : i-mat \n annealed : b-smt \n OZn : b-mat \n O2Ti : b-mat \n nanocomposites : b-dsc
```

Example: MatSIE (text)

```
system: You are a material scientist. Tackle the ensuing query with a well-organized and structured reply
```

```
question: The third type of Pr3+ site, observed only in the Ce-doped compounds, is tentatively ascribed to local inhomogeneities provoked by trapped electrons.
```

```
Uncover any references to doping in this sentence. The answer should be in the following schema: "basemats": {"h0": <host 0>, "h1": <host 1>}, "dopants": {"d0": <dopant 0>}, "dopants2basemats": {<dopant key>: [<basemat key>], }
```

```
answer: {"basemats": {"b0": "compounds"}, "dopants": {"d0": "Ce"}, "dopants2basemats": {"d0": ["b0"]}}
```

Example: MatSIE (tables)

```
system: As a materials science expert skilled in extracting information from tables, your objective is to complete the following dictionary based on the table provided. Define each key as follows:
```

```
'comp_table' Assign [1] if the table includes data on material compositions, otherwise [0].  
'regex_table' Assign [1] if material compositions are present and extractable via regex, otherwise [0].  
'composition_row_index' Indices of rows with full material composition details.  
'chemical_col_index' Indices of columns showing the constituent chemicals' values.  
'composition_col_index' Indices of columns with full material composition details.  
'chemical_row_index' Indices of rows showing the constituent chemicals' values.  
'gid_row_index' Index of the row with the material identifier.
```

```

'gid_col_index' Index of the column with the material identifier.

dictionary = {
    'comp_table': [],
    'regex_table': [],
    'composition_row_index': [],
    'composition_col_index': [],
    'chemical_row_index': [],
    'chemical_col_index': [],
    'gid_row_index': [],
    'gid_col_index': []
}

NOTE: Only keys with non-empty lists will be included in the output.
question:Caption: Composition of glasses (mol%)

Table: [[{'Sample': 'SiO2', 'CaO': 'ZrO2', 'V2O5': 'B-glass', '57.17': '36.75', '6.08': '0.00'}, {'V-0.1': '57.11', '36.71': '6.08', '0.10': '0.00'}, {'V-0.3': '57.00', '36.64': '6.06', '0.30': '0.06'}, {'V-0.5': '56.88', '36.57': '6.05', '0.50': '0.06'}, {'V-0.7': '56.77', '36.49': '6.04', '0.70': '0.04'}, {'V-1.0': '56.60', '36.39': '6.02', '0.99': '0.03'}, {'V-2.0': '56.05', '36.03': '5.96', '1.96': '0.02'}, {"V-5.0": "54.45", "35.00": "5.79", "4.76": "0.01"}]
Footer: {}
answer:
{ 'comp_table': [1],
  'composition_row_index': [1, 2, 3, 4, 5, 6, 7, 8],
  'chemical_col_index': [1, 2, 3, 4],
  'gid_col_index': [0],
  'regex_table': [0]}
}

```