Position Paper: Are LLMs Ready for Real-World Materials Discovery?

Anonymous Authors[1]

Abstract

Large Language Models (LLMs) create exciting possibilities for powerful language processing tools to accelerate research in materials science. While LLMs have great potential to accelerate materials understanding and discovery, they currently fall short in being practical materials science tools. In this position paper, we show relevant failure cases of LLMs in materials science that reveal current limitations of LLMs related to comprehending and reasoning over complex, interconnected materials science knowledge. Given those shortcomings, we outline a framework for developing Materials Science LLMs (MatSci-LLMs) that are grounded in materials science knowledge and hypothesis generation followed by hypothesis testing. The path to attaining performant MatSci-LLMs rests in large part on building high-quality, multi-modal datasets sourced from scientific literature where various information extraction challenges persist. As such, we describe key materials science information extraction challenges which need to be overcome in order to build large-scale, multi-modal datasets that capture valuable materials science knowledge. Finally, we outline a roadmap for applying MatSci-LLMs for real-world materials discovery through six interacting steps: 1. Materials Query; 2. Data Retrieval; 3. Materials Design; 4. Insilico Evaluation; 5. Experiment Planning; 6. Experiment Execution.

1. Introduction

The advent of large language models (LLMs) has started to reshape many technology development efforts and research roadmaps. Apart from significantly impacting the space of natural language processing, LLMs have had significant impact on many related fields, such as computer vision with text-to-image generators (Zhang et al., 2023), and have sparked the creation of efforts to integrate their capabilities into diverse industries. Examples of these efforts include task automation in healthcare (He et al., 2023), legal (Dahl et al., 2024), finance (Wu et al., 2023a), software engineering (Fan et al., 2023) and multiple scientific fields. The successful application of LLMs to materials science (MatSci) has the potential to transform the field by accelerating the discovery, synthesis and analysis of new materials that can address some of today's complex societal-scale challenges, such as climate change, energy security, sustainable agriculture and manufacturing, personalized medical devices, and access to more powerful computing systems. While recent research work has seen increased adoption of LLMs in adjacent scientific fields, such as chemistry (Jablonka et al., 2023) and diverse areas of biology (Lin et al., 2023; Hsu et al., 2022; Xu et al., 2023; Cui et al., 2023; Dalla-Torre et al., 2023), the application of LLMs in MatSci has been significantly slower. In this paper, we analyze current failures of LLMs in MatSci, and propose requirements for MATerials SCIence LLMs (MatSci-LLMs) along with a roadmap to enable the impactful application of MatSci-LLMs for furthering the field of materials.

1.1. MatSci-LLM Requirements

Material science is a highly interdisciplinary field – the first scientific branch that emerged out of fusion of different branches (physics, chemistry, mechanical engineering, and metallurgy) rather than a division – that studies how the interaction of atomic matter affects the properties and behavior of materials systems across diverse sets of conditions. The broad scope of materials systems encompasses multiple physical scales ranging from nanostructures with a small number of atoms which are used in modern computing devices, to continuum materials, such as metals and cement, that make up modern infrastructure like roads, bridges and buildings. Moreover, the wide range of materials application can range from electronic materials (e.g., transistors, batteries), to chemical materials (e.g., polymers) to biological materials (e.g., medical implants), each of which have very different requirements and application environments. As such MatSci often borrows concepts from related scientific fields, including physics, chemistry and biology, leading to vast and dynamic body of work covered by the field. This diversity also informs the wide range of tasks that materials scientists perform across different subfields, which further informs the requirements of MatSci-LLMs.

Given the diversity of MatSci knowledge and its associated technical tasks, we propose that MatSci-LLM should meet the following requirements, shown in Figure 1: 1. *Domain Knowledge & Grounded Reasoning:* MatSci-LLMs should

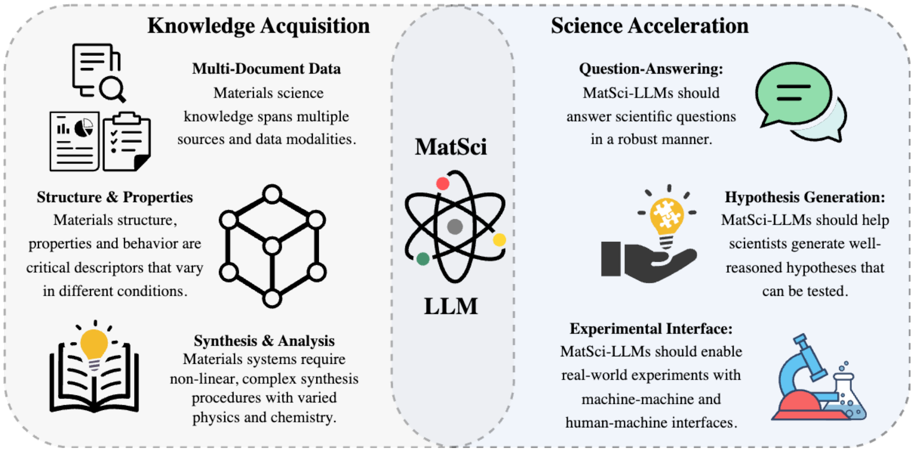

Figure 1. Overview of MatSci-LLM requirements related to knowledge acquisition and science acceleration. MatSci-LLMs require knowledge contained across multiple documents along multiple data modalities. Pertinent materials science knowledge includes understanding materials structure, properties and behavior covering diverse conditions, as well as materials synthesis and analysis procedures based on experimental descriptions. To effectively accelerate science, MatSci-LLMs should interact with human scientists as robust question-answering system and act as grounded hypothesis generators that augment a scientist's knowledge. Additionally, MatSci-LLMs should provide executable procedures for real-world experiments through machine-machine and human-machine interfaces.

have an understanding of MatSci as a subject area to provide useful information to material scientists and be able to reason over core MatSci principles, 2. *Augmenting Materials Scientists:* MatSci-LLMs should have the ability to perform useful tasks to accelerate materials science research that augments the work of materials scientists in reliable and interpretable manner.

Domain Knowledge & Grounded Reasoning: Earlier efforts on MatSci information extraction relied on rule-based (e.g. regex) approaches and domain-specific training. These approaches, however, could rarely scale beyond the specific materials family studied as described in Appendix A. Recent research efforts have yielded progress towards infusing MatSci domain knowledge into LLMs to solve text-based tasks. Early work, such as Walker et al. (2021); Gupta et al. (2022); Huang & Cole (2022) focused on training BERT (Devlin et al., 2019) architecture on domain-specific text corpora relevant to MatSci. A detailed analysis of these and other scientific language models by Song et al. (2023a) revealed that domain-specific pretraining did infuse the language models with relevant materials knowledge showing the importance of building high-quality MatSci language datasets. Recent works by Song et al. (2023b) and Xie et al. (2023) showed that instruction fine-tuning leveraging commercial scale LLMs can further improve performance on MatSci language tasks. While this suggests that practical MatSci knowledge can be extracted from today's commercial LLMs (e.g., GPT-3.5, Claude and GPT-4) their understanding of MatSci remains incomplete as shown by Zaki et al. (2024). As we discuss in Section 2, grounded reasoning over materials science principles remains a challenge.

Augmenting Materials Scientists: Research work in this direction has been primarily limited to adjacent domains, in particular chemistry. Jablonka et al. (2023) showed different applications of LLMs across diverse chemistry tasks, and Bran et al. (2023) and Boiko et al. (2023) provided examples of how a tool-agumented GPT-4 can perform real-world experimental planning and execution based on user queries. While such works show great relevance to MatSci, they are incomplete in addressing the vast diversity of knowledge and scientific tasks in MatSci, thereby falling short on the criteria needed to be considered MatSci-LLMs. One concrete difference between chemistry LLMs and MatSci-LLMs, for example, is the fact that IUPAC names in chemistry make the identification and tracking of molecular structures significantly easier. As discussed in Section 3, the materials domain contains much greater diversity including periodic crystal structures, surfaces, nanostructures, metals and alloys, and disordered materials like glasses, limiting the degree of systematization that can be achieved for indexing these materials with natural language.

1.2. What is needed for MatSci LLMs?

To fully unlock the power of LLMs for materials discovery, further progress is needed along the following directions.

1. Hypothesis Generation Grounded on MatSci Knowledge: MatSci-LLMs require understanding of diverse MatSci knowledge and the ability to reason over core domain principles. This will enable MatSci-LLMs to answer queries of domain-specific questions and generate useful hypotheses that can be verified with real-world experiments. Modern LLMs struggle to provide well-reasoned answers that reference relevant supporting evidence for in depth questions. While ongoing progress highlights the capability of modern LLMs to provide relevant MatSci knowledge (Song et al., 2023b), clear gaps remain (Zaki et al., 2024).

2. Multi-Modal MatSci Datasets: To enable MatSci-LLMs to capture the true extent of human knowledge of

MatSci, we require large-scale, high-quality, multi-modal datasets based primarily on peer-reviewed MatSci publications. Much of the knowledge in the MatSci literature is contained in tables, figures and other modalities that express important materials properties and behavior in addition to the physical and chemical relationships that underlie them. Additionally, modern LLMs still struggle to understand domain-specific MatSci language and effectively process MatSci notation, including chemical formulas and mathematical formulas expressed in line with other text.

3. Real-World Materials Design: Connecting MatSci-LLMs to real-world simulation and experimental tools in creative ways can comprehensively accelerate materials design, synthesis and analysis by enabling end-to-end automation. As described in Section 5, MatSci-LLMs can provide powerful human-machine and machine-machine interfaces for increased automation in MatSci research, as well as commercial materials production. Recent work in chemistry LLMs (Bran et al., 2023; Boiko et al., 2023) has shown promise in deploying tool-augmented LLMs in for end-to-end materials design tasks driven by text-based user input.

2. Failure Cases of LLMs in MatSci

While recent work has shown great promise in applying LLMs to materials science and adjacent fields, it is important to understand the limitations of LLMs on a broad set of challenges in the MatSci domain. To this extent, several efforts have focused on tasks related to question-answering, code writing, named-entity recognitions, abstract classification, and composition extraction from tables in materials literature (Zaki et al., 2024; Gupta et al., 2023; Song et al., 2023b;a). Here, we outline some of the failure cases demonstrating the need of developing robust MatSci-LLMs before deploying them for practical applications.

Proficient LLMs such as GPT-4 (Achiam et al., 2023) and LLaMA-2 (Touvron et al., 2023) are trained on vast corpora of publicly available text data (Computer, 2023). Accordingly, commercial LLMs should have had some reasonable exposure to the MatSci domain based on the wikipedia and other public domain information on MatSci. To evaluate the MatSci domain knowledge of LLMs, Zaki et al. (2024) curated a dataset of 650 questions requiring undergraduate-level knowledge to answer them. The analysis of the responses from LLMs revealed that GPT-4 with chain-of-thought (CoT) (Wei et al., 2022) reasoning performed the best with a score of 62% accuracy. Interestingly, GPT4-CoT performed the worst in numerical type questions with a score of only 39%. This suggests that present LLMs perform poorly in identifying the correct equations, grounding the context by substituting appropriate numerical values, and subsequently solving the original question. Moreover, a comparison of the performance of LLMs with human performance in the same exam revealed that GPT4-CoT, while outperforming other baselines, consistently obtained only 50% of the top performing human and ultimately never cleared the cutoff to pass the exams.

A task where LLMs are known to perform well is code generation (Chen et al., 2021; Lai et al., 2023; Zan et al., 2023). While code generation for chemistry tasks has yielded mostly promising results as detailed in White et al. (2023), Zaki et al. (2024) evaluated the performance of LLMs on code generation tasks as well and found that GPT-4 exhibited an accuracy of only 71% even in code writing tasks related to MatSci. A careful analysis of the questions and code-writing tasks where GPT4-CoT performed poorly reveals two salient conclusions: (i) LLMs have difficulty in solving complex numerical problems, not just because of the complex arithmetic (which could potentially be solved by translating the numerical question into code (Wu et al., 2023b)), but because of their inability to ground the available data in the given equation. This includes wrong substitution, incorrect conversion of units, and missing constants while converting units (e.g., taking into account the speed of light, Avagadro number, etc.). (ii) LLMs have difficulty in understanding the information related to 3D structure. Specifically, the crystal structures and the symmetries associated with materials are misinterpreted leading to incorrect conclusions. Zaki et al. (2024) observed these errors consistently in both code generation tasks and question-answering tasks for multiple LLMs.

To elucidate current LLM failure cases in MatSci, consider the question below. *"The materials belonging to which one of the following crystal classes would be both piezoelectric and ferroelectric: (A) 222 (B) 4mm (C) -1 (D) 2/m"*. Here, while (B) 4mm is the correct answer, GPT4-CoT suggests 2/m as the correct answer with the reasoning that *"(D) 2/m: This crystal class does not have a center of symmetry, which means it can be piezoelectric. Additionally, it is a polar crystal class, making it able to be ferroelectric as well."*. However, this reasoning is incorrect—2/m is centrosymmetric, and hence is not a polar crystal. Answering this question requires the knowledge of crystal structures, which needs to be grounded on the concept of polarity of atoms based on the crystal structure. Although the LLM understands this requirement, it performs poor reasoning and inference, potentially due to the lack of such data during training. Another example worth noting is the following question. *"The Miller indices of the first three Bragg peaks in the X-ray diffraction pattern obtained from a polycrystalline iron sample at room temperature are (A) (111), (200), (220) (B) (100), (110), (111) (C) (100), (110), (200) (D) (110), (200), (220)"*. While the correct answer is (D), GPT4-CoT suggests (C) as the answer, based on the reasoning that *"systematically listing all possible combinations of h, k, and l, and identifying those that satisfy the BCC selection rule (h + k + l = even), we get the following allowable Miller indices as: (100), (110), and (200)"*. However, this is contradictory as for the first Miller index (100) in the list, h + k + l = 1 + 0 + 0 = 1, which is odd and not even. Accordingly, while the understanding that the BCC selection rule requires (h + k + l) should be even is correct, the LLM was unable to ground this concept correctly and suggest the correct Miller indices. Ten such examples are listed in the Appendix D, where we provide the question, the answer and the reasoning provided by GPT-4, along with the correct solution. We observe consistent mistakes made by GPT-4 in numerical errors and reasoning inconsistencies both in conceptual MatSci questions and in MatSci code generation tasks. These observations strongly suggest that LLMs need to be exposed to more domain-specific information and develop greater reasoning capabilities for them to be usable for any real-world application in this domain.

LLM reasoning difficulties extend beyond the MatSci domain, including in general language tasks (BehnamGhader et al., 2023). Given the great importance of reasoning abilities to formulate and test scientific hypothesis, this remains a major gap in the development of MatSci-LLMs which also has implications for the broader LLM community. While LLMs have shown promise in solving concrete planning tasks, such as retrosynthesis planning (Boiko et al., 2023; Bran et al., 2023), and executing code-based functions as part of a greater workflow (Buehler, 2023; Yoshikawa et al., 2023), LLMs still rely on human intervention and correction in many cases. In addition to greater robustness, future LLMs can also benefit from greater interpretability to accelerate hypothesis testing and enable scientists to better understand the various aspects of the system they are interested in. This is especially important in MatSci where many experimental procedures can expose valuable information about the underlying material properties and behavior under diverse settings, which in turn can inform broader understanding of fundamental physical and chemical relationships of the underlying material system.

3. Grounding LLMs in Domain-Specific Language in MatSci

The MatSci domain requires great technical depth and breadth due to field's broad technical scope that interfaces with physics, chemistry, biology and various engineering disciplines. Hence, MatSci presents unique challenges for knowledge acquisition for domain-specific language models. While some of the challenges, such as domain-specific notations, are unique to MatSci, others have a broad intersection with multiple machine learning fields, such as multi-modal information extraction involving text, images, and video.

Domain-Specific Notations: Unlike IUPAC (Hellwich et al., 2020) nomenclatures in chemistry, there exist no standard notations for MatSci; for instance, $NaAlSi_2O_8$, $Na_2O.Al_2O_3.2SiO_2$, SiO_2–$0.5Na_2O$–$0.5Al_2O_3$ all represent the same material in different contexts. Moreover, several common domain-specific names are used to represent materials, including soda and lime for Na_2CO_3 and $CaCO_3$, respectively. Furthermore, some sub-domains, such as cement, employ their own chemical notations, that are contradictory with the general chemistry notation; C-S-H represents calcium silicate hydrate in cement chemistry, whereas in standard chemistry notation, it represents carbon, sulfur, and hydrogen, respectively. Similarly, there are several inconsistent abbreviations to refer to properties in MatSci that are used in the literature. These abbreviations are sometimes specific to a given document or can also be present in multiple documents. Thus, understanding these notations require grounding of the LLMs to different domains with the right context.

Incomplete Descriptions: Research publications regularly refer to previous works in the text to omit the specific information. This is exemplified by the sentences such as *"the fracture simulations were carried out using an established methodology as outlined in Griffith et al."*; a common practice in the MatSci literature. Such approaches are also extremely common when reporting experimental or simulation protocols, material compositions, syntheses and testing conditions, and optimizing process parameters. The context of the work reported in a manuscript, in this case, is thus grounded in a different manuscript. Moreover, the descriptions can occasionally be highly non-linear while referring to different sources in an extremely complicated manner. Finally, there could be reference to the manuals of different instruments and material sources, which are described elsewhere. Thus, an LLM needs to procure information across multiple sources and process it together in the proper context to create a sequential description of the process and the related analysis that was carried out.

Text-to-Structure: MatSci employs text to represent several 3D or 2D structures in different ways. For instance, the crystal structures are represented using the Wyckoff positions (Aroyo et al., 2006) –4mm in crystallography represents a crystal structure, whereas in general literature it might be confused with a distance metric (millimeter). Moreover, crystals are represented using a widely used format known crystal information file (CIF), which explains the details of the crystal along with the positions of the atoms as shown in Appendix C with a CIF of Silicon. Current LLMs are unable to read, interpret, or generate CIF in their entirety, which are a strong limiting factor for novel materials discovery. Similarly, there are several other approaches to represent the 3D structure using text, such as xyz files or other software specific files for diverse purposes. As such, the ability of LLMs to understand such files plays a crucial role for materials domain. Recent work by Gruver et al. (2023) shows that LLMs can generate correct CIF-formatted materials with targeted fine-tuning, but fail to provide context describing important details of the CIF file.

Multi-Modal Information Extraction: Text and Tables: Pertinent information for materials in peer-reviewed papers are generally represented using multiple modalities including text, tables, figures, and videos—audio is very rarely used, although not absent. Significant progress has been made in effectively extracting tabular information (Gupta et al., 2023; Zhao et al., 2023; Zhao & Cole, 2022) with work still remaining on how to best make use of extracted data to infuse the underlying knowledge of the tables into language models. Extraction can be particularly challenging for property based data, which often includes numbers that map to various types of scientific units. A simple exercise on information extraction from tables revealed that GPT-4 was able to extract only $\sim 55\%$ of materials properties from a table (Zaki et al., 2024) in contrast to $\sim 73\%$ by a model trained for the task (Gupta et al., 2023). This problem is amplified by the fact that more than 80% of material compositions and properties are reported in tables (Hira et al., 2023). Finally, in several cases table information is incomplete, even when read together with the caption. For instance, when reporting the composition of Magnesium alloys, many tables in research papers report the values of all the components other than Magnesium in a given alloy. Thus, the percentage of Magnesium in these alloys need to be computed as 100-(the sum of reported elements in the tables for a given alloy). Such a task requires the grounding of the table information along with the text, and then performing the necessary arithmetic to obtain the percentage.

Multi-Modal Information Extraction: Text and Figures: Materials are characterized by many experimental methods such Raman analysis, X-ray diffraction (XRD), X-ray fluoroscence, scanning electron micrographs or atomic force microscopy, the results of which are represented in figures. Such information could include the visual representation as captured by the measurement technique, such as an optical micrograph, or a plot that summarizes this information as in the case of XRD patterns. Moreover, this information would be further elaborated briefly in the figure caption followed by detailed explanation in the text. A sample text from a materials literature could be as follows: *"The XRD patterns shown in Figure XY(a) suggest that the sample is amorphous and does not have any crystalline content. Further, the XRF in Fig. XY(b) reveals the presence of clustering of Calcium instead of a uniform distribution"*. In this particular case, the text and figure need to be interpreted together. Moreover, the specific aspects of the image that allows one to make an interpretation such as the clustering of calcium or the specific feature of the XRD patterns that allows a researcher to identify that the material is amorphous is not labeled. In order for the LLM to learn these aspects, a large number of images and aligned text are need for joint training.

Multi-Modal Information Extraction: Text, Figures, Tables, and Videos: Additional modalities of presenting relevant information involve a combination of text, figures, and tables. In such cases, the information from the table should be read along with the figures while grounding the reasoning in the context of the manuscript, all of which further compound the cumulative importance of multi-modal information extraction. Occasionally, additional modalities such as audio-visual files are used to demonstrate material response. This includes material failure modes illustrating cracks, crystal growth, thermal response, response of actuators, and simulations amongst others. In this case, there are additional challenges that require the training LLMs to properly link the information in video, figures, and text and tables. These are open problems for which the community may need to develop novel methods and architectures.

While these challenges are far from exhaustive, we note that the tasks mentioned here reveal novel domain-specific problems whose solutions can have potentially strong impact on other scientific and machine learning domains as well.

3.1. Context Across Multiple Documents & Sources

Most of the information generated through the process of scientific inquiry is documented and communicated through natural language in peer reviewed publications. Understanding this text – and potentially extracting the information – presents several unique challenges distinct from other domains (Hira et al., 2023). One of the major challenges in this regard includes understanding the importance of context across multiple documents and sources of information. As highlighted above, this is additionally complicated by the fact that a given material can have multiple correct names and designations, meaning that important information can be contained across all of the names for a given material. Having unique IUPAC-style naming conventions can accelerate the deployment of LLMs as observed in chemistry (Jablonka et al., 2023), yet this is often not sufficient to capture all the relevant details needed to fully understand materials properties and synthesis leading to the development of additional data initiatives to fill that gap (Kearnes et al., 2021; Mercado et al., 2023). Hence while proposing an IUPAC-style naming convention might seem like a natural solution to aid the deployment of LLMs in MatSci, many proposals of such conventions have previously fallen short given the vast diversity of the field.

3.2. Diverse Experimental and Simulation Procedures

While understanding context is the first step towards automated discovery, the next natural step is to identify the appropriate experimental or simulation protocols relevant to the context. There are a multitude of experimental and simulation procedures to evaluate the same objective. As such, understanding them and choosing the relevant one or a set of relevant experiments is a challenging task, which requires the domain-specific information and reasoning. Early works have shown some promise in providing useful tools for creating synthesis procedures of specific materials (Olivetti et al., 2020; Jensen et al., 2021), but much work remains in effectively deploying LLMs for these tasks. Additionally, materials synthesis procedures provide additional challenges given the diversity of synthesis techniques and processing conditions that can be observed across materials design use cases even within similar materials classes. Similarly, effective *insilico* design and filtering of materials requires accurate simulations, and the inability to generate the correct codes grounded in the context can make this task challenging. This is exemplified by the mistakes in the code generation tasks where the LLMs are unable to either ground the concept correctly, or the generated code exhibits numerical errors. Thus, it is imperative that large amounts of curated and complete data associated with simulation and experimental procedures are used to train effective MatSci-LLMs.

4. Multi-Modal MatSci Corpus Building

The performance of a language model heavily depends on the quality of the dataset on which they are trained on. Thus, dataset creation remains an integral part in enabling progress across various deep learning subfields, including computer vision, graph learning, as well as natural language in both general and scientific domains. MatSci has several domain-specific variations in terms of text, and recent developments in multi-modal language models motivate the development of datasets that align the additional modalities such as figures, tables, and images. Given the vast amount of scientific information expressed in diverse modalities, multi-modal language models would unlock significantly more powerful capabilities for scientific language modeling.

The gold standard data for training these LLMs for MatSci is mostly contained in peer-reviewed publications at established editorials, such as the family of journals at Elsevier, Royal Society, American Society and Springer Nature amongst others. Unfortunately, much of the content of peer-reviewed publications at established journals remains inaccessible beyond paywalls limiting public access, which makes it difficult to access these valuable text data. Accordingly, it is unlikely that any general-purpose language models, such as GPT-4 or LLaMa, have had access to this data, which likely contributes to their poor performance on MatSci tasks highlighted in Section 2. While there have been promising development in making scientific text-data available via open-access through various preprint servers and portals like Semantic Scholar, text-data obtained from such sources often requires extensive cleaning and preparation processess before becoming useful for training scientific LLMs (Lo et al., 2020). Research work in making text-data more accessible has already produced meaningful results in making language model training and evaluation more successful (Song et al., 2023a;b). However, curating such data on which a LLM can be trained requires addressing additional data-related challenges described below.

1. Data Availability: While families of journals such as Elsevier and Springer provide a text and data-mining APIs based on paid subscription, machine readable formats such as xml files are available only for the manuscripts published in the 21st century. Our analysis of MatSci related articles in Springer and Elsevier journals revealed that $\sim$ 6 million peer-reviewed articles with a total of $\sim$ 20 billion words are available in machine readable format through institutional subscriptions (see Appendix B). While this corpus is significantly smaller compared to general language LLMs, the quality and domain relevance is significantly higher. Nevertheless, almost all the publications from 20th century or earlier are only available as PDFs or scanned files of the original hardcopy publication which are not easily machine readable. Thus, a vast amount of data representing the cumulative scientific endeavor before 21st century is not available for LLM training. Additionally, many other family of peer-reviewed journal do not either allow text and data mining, nor have a framework to support it. Data sourced from preprint servers, such as arXiv, often require extensive cleaning to make them amendable to LLM training.

2. Large-Scale Description & Annotation: While pretraining does not require annotations, several downstream tasks require high-quality annotations for finetuning the model or to employ it using in-context or few-shot learning. This is particularly important when data is obtained from multiple sources and in multiple modalities, such as tables, text, images, videos, and codes. For instance, a CIF document of Silicon should have a detailed description about what information is contained in the file, so that the LLM can learn to interpret the information in the CIF in addition to understanding its format. However, such large-scale annotations are currently unavailable and would require input from domain experts for them to be reliable. Given these challenges, to ensure scalable dataset generation, we propose: **1. Distant Supervision:** Here, sparse supervision signals are used to alleviate the amount of high-quality annotations required. This has shown promise in data extraction (Gupta et al., 2023) and can be applied along with existing materials databases, such as the Cambridge Structure Database (Groom et al., 2016). **2. In-Context Learning:** General purpose LLMs, such as GPT4, exhibit excellent in-context learning capabilities. Thus, prompt engineering along with in-context learning can be used to generate high-quality training data. Song et al. (2023b) demonstrated this approach for instructions-based finetuning of an LLM for materials science tasks. We provide several examples of composition extractions detailed in Appendix E, which show some of the capabilities and gaps of in context learning for GPT-4. **3. Domain-Specific Regex** can also be very useful to develop high-quality training data. Robocrystallographer (Ganose & Jain, 2019), for example, is a regex based approach that can produce a textual description for a CIF

file. Similarly, ChemDataExtractor (Swain & Cole, 2016) employs a regex based approach for chemical data extraction. Regex tools, along with LLMs and distant supervision, comprise promising methods for scalable data generation.
3. Linking Multiple Entities: While obtaining data on multiple entities is possible, linking them appropriately so they can be read together with relevant context is a non-trivial task. For instance, the description of a figure or a table in a manuscript could be spread across multiple paragraphs and even the supplementary material. This is in contrast to standard machine learning approaches where a figure and description are given together while training. Thus, developing a dataset that links multiple entities appropriately and employing an appropriate training scheme that respects the dataset and the context remains an open challenge for the research community.
4. Handling External References: Another major challenge while curating a dataset based on peer-reviewed publications relates to the use of external references. In a manuscript, references to multiple documents will be made to either support the claims made in the current work and thereby base the manuscript to be read in the context of the referred work. In other words, information in the present manuscript is grounded on the claims or observations in the reference manuscript, which in turn is grounded in its own references. Dataset and training schemes developed to train a MatSci-LLM should respect this fact to reduce hallucinations and provide well-reasoned, executable hypotheses. It is also worth noting that several LLMs often hallucinate when asked for references. Furthermore, when asked to generate a text in the form of a scientific manuscript, these LLMs generate fictitious references which are arbitrarily written to fit sentences generated in the process. This motivates the need to appropriately account for external references in the training data.

Altogether, these challenges suggest that the development of a dataset for MatSci-LLMs requires a close collaboration between multiple entities, including publishers, government, industry and academia. While development of high-quality datasets are indeed challenging, employing some of the aforementioned approaches along with the development of new methods for managing sparse datasets can spark research for novel solutions that effectively manage data sparsity in the MatSci domain. Moreover, MatSci-LLMs warrant machine learning solutions for several new challenges such as: (i) handling old and non-machine readable data through computer vision techniques that can convert the scanned documents to text respecting the original formats, and (ii) novel machine learning methods that can handle external references and multi-modal data that are linked together across different contexts. Such solutions could have implications beyond the MatSci domain, such as analyzing old documents relevant to the history, law, and finance.

5. Roadmap for Application of MatSci-LLMs

Figure 2 shows our proposed outline of an end-to-end materials discovery framework with MatSci-LLM at the core. Each part of the roadmap contains diverse research opportunities building towards performant MatSci-LLMs.

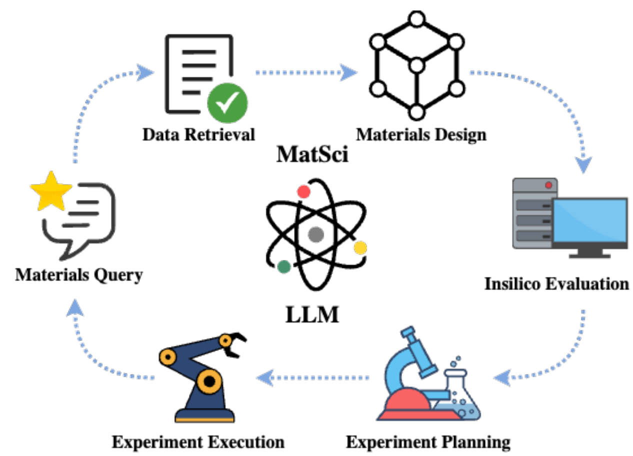

Figure 2. Roadmap of a Mat-Sci LLM based materials discovery cycle. The cycle starts with materials query from a researcher that specifies desired properties or an application. The MatSci-LLM then draws from external and internal knowledge bases to generate a materials design hypothesis which is evaluated in-silico. Next, the MatSci-LLM ingests the in-silico results and prepares an experimental plan to synthesize and characterize the material, after which the MatSci-LLM interfaces with the relevant machines to execute the experimental workflow. The final result is then shown to the user for evaluation and feedback. Each stage can interact with another for refinement and improvement by the MatSci-LLM.

1. Materials Query: Materials queries may span diverse user queries, such as, ”Suggest potential candidate materials for Li-solid state battery electrolytes”. Alternatively, more complex queries may require the MatSci-LLM to generate hypotheses based on presented experimental data, a text document or a combination of multiple such sources. As previously described, hypothesis generation remains an open research problem for MatSci-LLMs and may require interaction with other steps in the roadmap. Thus, materials query refers to the initial stage where the problem identification is performed either manually or through an LLM, presenting a research opportunity for novel, specialized querying methods. Emerging work may also build on top of novel human-machine interfaces for designing complex materials systems (Kanarik et al., 2023; Vasudevan et al., 2022) that enable greater efficiency and faster scientific understanding.
2. Data Retrieval: The second step toward materials discovery is to exploit the existing knowledge from MatSci-LLM internal and external knowledge bases. External knowledge includes databases of diverse materials properties (e.g., composition–structure–properties), unstructured data from text, or images, or a combination thereof. The vast amount of publications and knowledge sources makes it non-trivial to

extract such information. As such, externalizing the knowledge of MatSci-LLMs could be useful. Recent works by Cox et al. and Buehler (2024) have shown that intentional use of LLMs has the potential to broaden the availability of useful scientific knowledge by externalizing the knowledge of domain-specific LLMs as structured, human-interpretable knowledge. Cox et al. generated annotations for a database of $> 15,000$ protein-coding by using robust methods that rely on citations and cross-checking of the generated information against the scientific literature. Buehler (2024) and Venugopal & Olivetti (2024) externalized LLM knowledge in a structured knowledge graph, which can be probed by a scientist to further their understanding and potentially make corrections and adjustments as needed. Future work can build on top of these early approaches leading to more reliable retrieval methods along with new methods for LLM-based knowledge externalization.

3. Materials Design: Based on the query and retrieved information, MatSci-LLMs may directly generate structures based on CIF files (Gruver et al., 2023; Flam-Shepherd & Aspuru-Guzik, 2023; Antunes et al., 2023) or in combination with current graph or point-cloud based generative methods (Xie et al., 2021; Jiao et al., 2023; Govindarajan et al., 2023; Merchant et al., 2023; Zeni et al., 2023). As shown in Zeni et al. (2023); Govindarajan et al. (2023), conditional materials generation is a promising approach for designing materials with specific properties, where the properties can also be predicted from LLMs based on the textual description (Rubungo et al., 2023). The rapid progress on generative modeling techniques in adjacent fields such as computer vision, as well as the vast space of unsolved materials problems, such as designing complex real-world systems with multiple structures, provides great opportunities for future research work.

4. Insilico Evaluation: Code generation abilities can enable tool-agumented MatSci-LLMs to execute complex simulation workflows to evaluate materials designs. Recent work by Buehler (2023) shows a possible example of how LLMs can be integrated with relevant tools to perform end-to-end insilico design of polymer materials where an LLM can generate a new molecular compound in SMILES notation and then query different LLM-based agents to perform relevant calculations for the same materials. Further examples in Buehler (2023) show how the LLM system can also set up the relevant computation environment by installing relevant packages, showing further benefits of end-to-end tool integration. While many of these systems have not yet been tested at large scales, this represents a great step forward in constructing impactful MatSci-LLMs. Whereas recent work provides many reasons to be optimistic about LLMs capabilities to solve concrete in-silico MatSci problems, much published work fails to include negative results outlining current LLM failures. As such, it is important to continue to have broader analysis of LLMs' capabilities for code generation in MatSci similar to Zaki et al. (2024), which can point to relevant research directions for more robust integration of MatSci-LLMs into materials simulation workflows.

5. Experiment Planning: Following insilico evaluation, the next step involves the development of experiment plan to synthesize and characterize the materials. LLMs have shown promise in automating simple laboratory experiments (Boiko et al., 2023; Bran et al., 2023), while adjacent work aims to build necessary infrastructure for autonomous laboratories that perform experiments without human intervention (Sim et al., 2023; Szymanski et al., 2023). The selection of synthesis and characterization remains highly non-trivial and depends significantly on the material systems, length and time scales, and the properties of interest. As such, further research is needed to have MatSci-LLMs effectively identify realistic, detailed, and executable experimental procedures informed by the constraints of state-of-the-art technology and user needs. Moreover, the development of effective digital twins for different synthesis procedures will be immensely valuable for insilico testing and refinement of experimental plans. As described in Appendix A, data mining approaches sometimes provide useful directions for targeted use cases, but still require large amount of expert intervention, making them unsuitable for large-scale use.

6. Experiment Execution: The actual execution of the experiments require identification of the relevant commands and the sequence in which these should be given. This also involves intermediate analysis and appropriate modification of the protocol according to the real experimental conditions at a given time. For instance, melting of a silicate system would require continuous strirring until the system is bubble free; a standard protocol while preparing glasses. Thus, MatSci-LLMs should be able to make and change the plans dynamically as per the actual conditions of the experiment. Recent work by Yoshikawa et al. (2023) shows that LLMs can provide practical interfaces that more effectively connect scientists to robotic systems to perform chemical tasks. Current capabilities, however, are still in their early stages with significant future work needed to create robust control methods for a single machine, as well as machine-to-machine communication capabilities.

6. Conclusion

The virtuous cycle show in Figure 2 has the potential to enable impactful scientific discoveries through end-to-end automation while concurrently augmenting human knowledge through the discovery of new physical and chemical relationships for an expanding set of materials. The unique challenges of deploying LLMs in MatSci oulined in this paper, however, require further research to make MatSci-LLMs effective scientific assistants. Meaningful research advances will require advances along the interface many fields, including machine learning, materials simulation, materials synthesis, materials characterization and robotics.

7. Broader Impact

Among the 17 sustainable development goals (SDGs) proposed by United Nations, 10 can be achieved or accelerated through materials innovation. The development of an end-to-end framework for accelerating materials discovery can have broader impacts in several domains such as healthcare, agriculture, energy, sustainability, water filtration, and carbon capture, to name a few. Traditional materials discovery remains an extremely slow process often taking a period of 20-30 years from starting a design to final deployment. AI-driven materials discovery has a potential to reduce this cycle to a few years or even months, thereby providing a means to address major societal-scale challenges. Moreover, the availability of cloud computing and web services can make this process extremely efficient and economical leading to democratization of materials discovery. This could reduce the gap between large corporations and small and medium scale industries allowing materials discovery without the availability of large lab facilities and vast R&D budgets. While Mat-Sci LLMs have great promise in reducing the cost and democratizing materials discovery, great care must be taken to ensure that MatSci-LLMs and all related technologies be used in an ethical and harm-limiting manner. As such, we propose the following framework to reduce the risk of potentially harmful deployment of MatSci-LLMs:

1. **Transparency:** Encouraging sharing of models, datasets, and research methodologies to promote transparency, replication, and validation of findings by the broader scientific community is key to ensure risk mitigation. This approach facilitates peer review and enables independent verification of results and claims. Further, maintaining comprehensive logs of data sources, model training processes, and decision-making pathways should be encouraged to ensure accountability and traceability.

2. **Risk Assessment:** Developing methodologies for assessing the risks associated with the application of LLMs in MatSci, focusing on both direct impacts (e.g., safety of newly discovered materials) and indirect impacts (e.g., environmental effects of scaled production) is of paramount importance. Further, systems should be established for ongoing monitoring of the outcomes and impacts of materials discovered or optimized through LLMs to identify and address unforeseen risks promptly.

3. **Regulatory Compliance & Standardization:** Guidelines that address the unique challenges and risks associated with using LLMs in materials discovery should be created and enforced by governing agencies. Furthermore, standards should be developed for data representation, algorithms, and model transparency to facilitate oversight and ensure quality and reliability.

4. **Education & Training:** Education and training programs that integrate ethics, environmental science, sustainability, and AI should be developed to prepare a workforce capable of navigating the complexities of using LLMs in materials discovery. Additionally, dialogues and knowledge exchange between scientists, policymakers, industry stakeholders, and the public to build understanding and trust in AI-driven materials discovery should be facilitated. A noteworthy aspect that requires discussion is the tendency of LLMs to hallucinate. It is important to educate the issues associated with hallucinations in LLMs and how this can have impact on the task that it is applied for.

One example where MatSci-LLMs could cause unintended harmful consequences relates to the use of natural resources. One can imagine a case where new materials discoveries could put greater demand on minerals that were previously unterutilized. This, in turn, could result in significant impacts on the local communities and the surrounding ecosystem and would therefore require action at policy level to avoid inadvertent consequences.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Antunes, L. M., Butler, K. T., and Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *arXiv preprint arXiv:2307.04340*, 2023.

Aroyo, M. I., Perez-Mato, J. M., Capillas, C., Kroumova, E., Ivantchev, S., Madariaga, G., Kirov, A., and Wondratschek, H. Bilbao crystallographic server: I. databases and crystallographic computing programs. *Zeitschrift für Kristallographie-Crystalline Materials*, 221(1):15–27, 2006.

BehnamGhader, P., Miret, S., and Reddy, S. Can retriever-augmented language models reason? the blame game between the retriever and the language model. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15492–15509, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1036. URL https://aclanthology.org/2023.findings-emnlp.1036.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Bran, A. M., Cox, S., White, A. D., and Schwaller, P. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Buehler, M. J. Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. *ACS Engineering Au*, 2023.

Buehler, M. J. Mechgpt, a language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines, and modalities. *Applied Mechanics Reviews*, 76(2):021001, 2024.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Computer, T. Redpajama: an open dataset for training large language models, 2023. URL https://github.com/togethercomputer/RedPajama-Data.

Cox, S., Hammerling, M., Lála, J., Laurent, J., Rodriques, S., Rubashkin, M., and White, A. Wikicrow: Automating synthesis of human scientific knowledge.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., and Wang, B. scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv*, 2023.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., and Jain, A. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024.

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*, 2024.

Dalla-Torre, H., Gonzalez, L., Mendoza Revilla, J., Lopez Carranza, N., Henryk Grywaczewski, A., Oteri, F., Dallago, C., Trop, E., Sirelkhatim, H., Richard, G., et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pp. 2023–01, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Fan, A., Gokkaya, B., Harman, M., Lyubarskiy, M., Sengupta, S., Yoo, S., and Zhang, J. M. Large language models for software engineering: Survey and open problems. *arXiv preprint arXiv:2310.03533*, 2023.

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

Ganose, A. M. and Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, 2019.

Govindarajan, P., Miret, S., Rector-Brooks, J., Phielipp, M., Rajendran, J., and Chandar, S. Learning conditional policies for crystal design using offline reinforcement learning. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. URL https://openreview.net/forum?id=VbjD8w2ctG.

Groom, C. R., Bruno, I. J., Lightfoot, M. P., and Ward, S. C. The cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, 2016.

Gruver, N., Sriram, A., Madotto, A., Wilson, A. G., Zitnick, C. L., and Ulissi, Z. W. Fine-tuned language models generate stable inorganic materials as text. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. URL https://openreview.net/forum?id=0r5DE2ZSwJ.

Gupta, T., Zaki, M., Krishnan, N. A., and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.

Gupta, T., Zaki, M., Khatsuriya, D., Hira, K., Krishnan, N. M. A., and Mausam. DiSCoMaT: Distantly supervised composition extraction from tables in materials science articles. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13465–13483, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.753. URL https://aclanthology.org/2023.acl-long.753.

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., and Cambria, E. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*, 2023.

Hellwich, K.-H., Hartshorn, R. M., Yerin, A., Damhus, T., and Hutton, A. T. Brief guide to the nomenclature of organic chemistry (iupac technical report). *Pure and Applied Chemistry*, 92(3):527–539, 2020.

Hira, K., Zaki, M., Sheth, D. B., ., M., and Krishnan, N. M. A. Reconstructing materials tetrahedron: Challenges in materials information extraction. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. URL https://openreview.net/forum?id=u6ndfkFRJC.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

Huang, S. and Cole, J. M. Batterybert: A pretrained language model for battery database enhancement. *Journal of Chemical Information and Modeling*, 62(24):6365–6377, 2022.

Jablonka, K. M., Ai, Q., Al-Feghali, A., Badhwar, S., Bocarsly, J. D., Bran, A. M., Bringuier, S., Brinson, L. C., Choudhary, K., Circi, D., et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250, 2023.

Jensen, Z., Kwon, S., Schwalbe-Koda, D., Paris, C., Gómez-Bombarelli, R., Román-Leshkov, Y., Corma, A., Moliner, M., and Olivetti, E. A. Discovering relationships between osdas and zeolites through data mining and generative neural networks. *ACS central science*, 7(5):858–867, 2021.

Jiao, R., Huang, W., Lin, P., Han, J., Chen, P., Lu, Y., and Liu, Y. Crystal structure prediction by joint equivariant diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=DNdN26m2Jk.

Kanarik, K. J., Osowiecki, W. T., Lu, Y., Talukder, D., Roschewsky, N., Park, S. N., Kamon, M., Fried, D. M., and Gottscho, R. A. Human–machine collaboration for improving semiconductor process development. *Nature*, 616(7958):707–711, 2023.

Kearnes, S. M., Maser, M. R., Wleklinski, M., Kast, A., Doyle, A. G., Dreher, S. D., Hawkins, J. M., Jensen, K. F., and Coley, C. W. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021.

Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., and Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, 29(21):9436–9444, 2017.

Kim, E., Jensen, Z., van Grootel, A., Huang, K., Staib, M., Mysore, S., Chang, H.-S., Strubell, E., McCallum, A., Jegelka, S., et al. Inorganic materials synthesis planning with literature-trained neural networks. *Journal of chemical information and modeling*, 60(3):1194–1201, 2020.

Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E. A., and Ceder, G. Opportunities and challenges of text mining in materials research. *Iscience*, 24(3), 2021.

Lai, Y., Li, C., Wang, Y., Zhang, T., Zhong, R., Zettlemoyer, L., Yih, W.-t., Fried, D., Wang, S., and Yu, T. Ds-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*, pp. 18319–18345. PMLR, 2023.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://www.aclweb.org/anthology/2020.acl-main.447.

Mercado, R., Kearnes, S. M., and Coley, C. W. Data sharing in chemistry: lessons learned and a case for mandating structured reaction data. *Journal of Chemical Information and Modeling*, 63(14):4253–4265, 2023.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, pp. 1–6, 2023.

Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y.-J., and Hiszpanski, A. M. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7 (4), 2020.

Rubungo, A. N., Arnold, C., Rand, B. P., and Dieng, A. B. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029*, 2023.

Sim, M., Vakili, M. G., Strieth-Kalthoff, F., Hao, H., Hickman, R., Miret, S., Pablo-García, S., and Aspuru-Guzik, A. Chemos 2.0: an orchestration architecture for chemical self-driving laboratories. 2023.

Song, Y., Miret, S., and Liu, B. MatSci-NLP: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3621–3639, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.201. URL https://aclanthology.org/2023.acl-long.201.

Song, Y., Miret, S., Zhang, H., and Liu, B. Honeybee: Progressive instruction finetuning of large language models for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5724–5739, 2023b.

Swain, M. C. and Cole, J. M. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904, 2016.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624 (7990):86–91, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vasudevan, R. K., Orozco, E., and Kalinin, S. V. Discovering mechanisms for materials microstructure optimization via reinforcement learning of a generative model. *Machine Learning: Science and Technology*, 3(4):04LT03, 2022.

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

Walker, N., Trewartha, A., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K., Ceder, G., and Jain, A. The impact of domain-specific pre-training on named entity recognition tasks in materials science. *Available at SSRN 3950755*, 2021.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

White, A. D., Hocky, G. M., Gandhi, H. A., Ansari, M., Cox, S., Wellawatte, G. P., Sasmal, S., Yang, Z., Liu, K., Singh, Y., et al. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2(2):368–376, 2023.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023a.

Wu, Y., Jia, F., Zhang, S., Wu, Q., Li, H., Zhu, E., Wang, Y., Lee, Y. T., Peng, R., and Wang, C. An empirical study on challenging math problem solving with gpt-4. *arXiv preprint arXiv:2306.01337*, 2023b.

Xie, T., Fu, X., Ganea, O.-E., Barzilay, R., and Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Xie, T., Wan, Y., Huang, W., Yin, Z., Liu, Y., Wang, S., Linghu, Q., Kit, C., Grazian, C., Zhang, W., et al. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565*, 2023.

Xu, M., Yuan, X., Miret, S., and Tang, J. Protst: Multi-modality learning of protein sequences and biomedical texts. *arXiv preprint arXiv:2301.12040*, 2023.

Yoshikawa, N., Skreta, M., Darvish, K., Arellano-Rubach, S., Ji, Z., Bjørn Kristensen, L., Li, A. Z., Zhao, Y., Xu, H., Kuramshin, A., et al. Large language models for chemistry robotics. *Autonomous Robots*, 47(8):1057–1086, 2023.

Zaki, M., Jayadeva, J., Mausam, M., and Krishnan, N. A. Mascqa: Investigating materials science knowledge of large language models. *Digital Discovery*, 2024.

Zan, D., Chen, B., Zhang, F., Lu, D., Wu, B., Guan, B., Yongji, W., and Lou, J.-G. Large language models meet nl2code: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7443–7464, 2023.

Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Shysheya, S., Crabbé, J., Sun, L., Smith, J., et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

Zhang, C., Zhang, C., Zhang, M., and Kweon, I. S. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

Zhao, J. and Cole, J. M. A database of refractive indices and dielectric constants auto-generated using chemdataextractor. *Scientific data*, 9(1):192, 2022.

Zhao, J., Huang, S., and Cole, J. M. Opticalbert and opticaltable-sqa: Text-and table-based language models for the optical-materials domain. *Journal of Chemical Information and Modeling*, 63(7):1961–1981, 2023.

A. Text-Based Methods in Materials Science Before LLMs

Early work in applying text-based methods to materials science largely focused on text-and-data mining. While much progress has been made to enable data mining, many of challenges data-mining identified in early work by Kononova et al. (2021) still remain. One particularly common task for early text-mining methods was mining text for the purpose of predicting synthesis prcedures based on peer-reviewed literature (Kim et al., 2017; 2020). While the deployment of targeted text-mining techniques was successful for developing targeted synthesis procedures for a subset of materials systems, the proposed approaches did not provide scalable solutions to a diverse of materials design challenges.

In addition to text-mining, early work also focused on classifying the text in materials science publications according to their semantic meaning, which is similar to named-entity-recognition (NER) in traditional language processing (Gupta et al., 2022; Walker et al., 2021). As studied by Song et al. (2023a), specialized BERT models perform quite well on common NLP tasks for materials science assuming labeled data exists. This, however, is not scalable given the sparsity of labeled data in materials science. Ensuing information extraction methods, such ChemDataExtractor (Swain & Cole, 2016) spanned different modalities of data including scientific text, property data, spectroscopy data, as well as tabular data. While specialized methods continue to improve the performance of data extraction for applicable modalities in materials science (Gupta et al., 2023), the advent of LLMs has led to a more unified and reliable way to arrive at reliable information extraction (Dagdelen et al., 2024). As described in Dagdelen et al. (2024) and in Section 4, in-context learning abilities of modern LLMs are a powerful tool to accelerate the abilities to process and extract information in materials science text. While further research is needed to improve such methods, they represent significant progress that enable new research opportunities discussed in this paper.

B. List of MatSci journals

journal	papers	word_count
Journal_of_Alloys_and_Compounds	66829	238537453
Science_of_The_Total_Environment	59490	354781803
Applied_Surface_Science	51627	191223396
Chemical_Engineering_Journal	42530	228873825
International_Journal_of_Hydrogen_Energy	39913	62269929
Journal_of_Biological_Chemistry	39370	9020708
Ceramics_International	37643	135734837
Journal_of_Cleaner_Production	37118	253572991
Chemosphere	36689	176012463
Electrochimica_Acta	34396	150818043
Food_Chemistry	33565	149308849
Nuclear_Instruments_and_Methods_in_Physics_Res...	32843	107224761
Materials_Science_and_Engineering:_A	32692	131945933
Materials_Letters	32352	50776687
Construction_and_Building_Materials	32032	176566669
Materials_Today:_Proceedings	29214	75989713
Chemical_Physics_Letters	29134	83262010
Journal_of_Power_Sources	29025	124779521
Journal_of_Magnetism_and_Magnetic_Materials	28963	78504883
Sensors_and_Actuators_B:_Chemical	28918	109397473

Figure 3. List of 20 MatSci journal publications with maximum articles published along with the number of articles and their associated word counts that are available through publisher APIs for text mining.

Figure 3 shows the list of top 20 MatSci publications (in terms of the number of papers published). along with the number of papers and the associated word count. These represent the manuscripts that are available for download through the text-mining API. The complete list of journals are provided in https://anonymous.4open.science/r/LLM-B7FC/.

C. Crystal Information File of Silicon

The following represents a CIF document of silicon atom. The meta data includes the source based on which the CIF is developed, that is, the journal publication details including author names, title of the paper, journal, vol, year and pages. Further, the complete lattice parameters are also included in the file.

```
data_global
_chemical_name_mineral 'Silicon'
loop_
_publ_author_name
'Kasper J S'
'Richards S M'
_journal_name_full 'Acta Crystallographica'
_journal_volume 17
_journal_year 1964
_journal_page_first 752
_journal_page_last 755
_publ_section_title
;
The crystal structures of new forms of silicon and germanium
;
_database_code_amcsd 0009282
_chemical_compound_source 'Synthetic'
_chemical_formula_sum 'Si'
_cell_length_a 6.636
_cell_length_b 6.636
_cell_length_c 6.636
_cell_angle_alpha 90
_cell_angle_beta 90
_cell_angle_gamma 90
_cell_volume 292.226
_exptl_crystal_density_diffrn 2.553
_symmetry_space_group_name_H-M 'I a 3'
loop_
_space_group_symop_operation_xyz
'x,y,z'
'1/2+x,1/2+y,1/2+z'
'1/2+z,x,1/2-y'
'+z,1/2+x,-y'
'1/2+z,-x,y'
'+z,1/2-x,1/2+y'
'-z,x,1/2+y'
'1/2-z,1/2+x,+y'
'-z,-x,-y'
'1/2-z,1/2-x,1/2-y'
'1/2+y,1/2-z,-x'
'+y,-z,1/2-x'
'-y,1/2-z,x'
'1/2-y,-z,1/2+x'
'1/2-y,z,-x'
'-y,1/2+z,1/2-x'
'y,z,x'
```

```
'1/2+y,1/2+z,1/2+x'
'x,1/2-y,1/2+z'
'1/2+x,-y,+z'
'-x,y,1/2+z'
'1/2-x,1/2+y,+z'
'x,1/2+y,-z'
'1/2+x,+y,1/2-z'
'-x,-y,-z'
'1/2-x,1/2-y,1/2-z'
'1/2-z,-x,1/2+y'
'-z,1/2-x,+y'
'1/2-z,x,-y'
'-z,1/2+x,1/2-y'
'z,-x,1/2-y'
'1/2+z,1/2-x,-y'
'z,x,y'
'1/2+z,1/2+x,1/2+y'
'1/2-y,1/2+z,x'
'-y,+z,1/2+x'
'y,1/2+z,-x'
'1/2+y,+z,1/2-x'
'1/2+y,-z,x'
'+y,1/2-z,1/2+x'
'-y,-z,-x'
'1/2-y,1/2-z,1/2-x'
'-x,1/2+y,1/2-z'
'1/2-x,+y,-z'
'x,-y,1/2-z'
'1/2+x,1/2-y,-z'
'-x,1/2-y,z'
'1/2-x,-y,1/2+z'
loop_
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
Si 0.10030 0.10030 0.10030 0.01127
```

D. Mistakes in Q&A by GPT4-CoT

Following are some of the materials science questions where GPT4-CoT gave incorrect answers. We provide the detailed question and answer provided by the GPT4-CoT, correct answer and the mistake in the solution (highlighted in red).

Question 1

The materials belonging to which one of the following crystal classes would be both piezoelectric and ferroelectric: (A) 222 (B) 4mm (C) -1 (D) 2/m

Correct: B, GPT4-CoT: D

Solution by GPT4-CoT

To determine which crystal class would exhibit both piezoelectric and ferroelectric properties, we need to understand the properties of each class and the criteria for displaying piezoelectric and ferroelectric behavior.

Piezoelectric materials: These materials generate an electrical charge when subjected to mechanical stress or strain. For a material to be piezoelectric, it must not have a center of symmetry (also called a center of inversion). In other words, it should belong to a non-centrosymmetric crystal class.

Ferroelectric materials: These materials possess a spontaneous electric polarization that can be reversed by an external electric field. Ferroelectric materials must also lack a center of symmetry and fall within a polar crystal class, meaning they have a unique polar axis.

Now, we can evaluate each crystal class:

(A) 222: This crystal class has a center of symmetry, so it cannot be either piezoelectric or ferroelectric.

(B) 4mm: This crystal class does not have a center of symmetry, which means it can be piezoelectric. However, it is non-polar, meaning it cannot be ferroelectric.

(C) -1: This crystal class has a center of symmetry, so it cannot be either piezoelectric or ferroelectric.

(D) 2/m: This crystal class does not have a center of symmetry, which means it can be piezoelectric. Additionally, it is a polar crystal class, making it able to be ferroelectric as well.

Thus, among these crystal classes, only class 2/m would be both piezoelectric and ferroelectric.

The correct answer is: [D].

Correct solution.
2/m is centro-symmetric, and hence it is not polar. In contrast, 4mm point group is polar.

Question 2 (Given: Universal gas constant, R = 8.31 J $mol^{-1}K^{-1}$, Faraday's constant, F = 96500 C mol^{-1}). A half cell consisting of pure Ni immersed in an aqueous solution containing Ni^{2+} ions of unknown concentration, is galvanically coupled with another half cell consisting of pure Cd immersed in a 1M aqueous solution of Cd^{2+} ions. The temperature is 25°C and pressure is 1 atm. The standard electrode reduction potentials of Ni and Cd are -0.250 V and -0.403 V, respectively. The voltage of the cell is found to be zero. The concentration(in micro molar) of Ni^{2+} in the solution is

Correct answer: 6 to 7.3, GPT4-CoT: 0

Solution by GPT4-CoT

Given the standard electrode reduction potentials of Ni and Cd are -0.250 V and -0.403 V, respectively. The cell voltage is found to be zero, which means the overall reaction is in equilibrium. We can use the Nernst equation to find the concentration of Ni^{2+} in the solution.

The Nernst equation is as follows:

$E = E^0 - \frac{RT}{nF} \ln Q$

where, E is the cell voltage, E^0 is the standard cell voltage, R is the universal gas constant, T is the temperature in Kelvin, n is the number of moles of electrons transferred in the reaction, F is Faraday's constant, and Q is the reaction quotient.

Since the cell voltage is zero, the Nernst equation becomes:

$0 = E^0_{Ni} - E^0_{Cd} - \frac{RT}{nF} \ln Q$

Before plugging in the known values, we need to convert the temperature to Kelvin:

Temperature (T) = 25°C + 273.15 = 298.15 K

Now we can plug in the values:

$0 = (-0.250) - (-0.403) - \frac{8.31 \times 298.15}{2 \times 96500} \ln Q$

$0 = 0.153 - \frac{2.4685 \times 10^{-4}}{2} \ln Q$

Solve for Q:

$\ln Q = \frac{0.153 \times 2}{2.4685 \times 10^{-4}}$

$\ln Q = 1239.5083$

Take the exponent of both sides:

$Q = e^{1239.5083}$

Since the cell reaction is:

$Ni^{2+} + Cd \rightleftharpoons Ni + Cd^{2+}$

The reaction quotient Q is:

$Q = \frac{[Cd^{2+}]}{[Ni^{2+}]}$

Given that the concentration of Cd^{2+} is 1M:

$1 = \frac{[Cd^{2+}]}{[Ni^{2+}]} = \frac{1}{[Ni^{2+}]}$

Therefore, the concentration of Ni^{2+} is:

$[Ni^{2+}] = \frac{1}{Q} = \frac{1}{e^{1239.5083}}$

Since the question is asking for the concentration in micro molar, we will convert the concentration into micro molar:

$[Ni^{2+}] = \frac{1}{e^{1239.5083}} \times 10^6 \, \mu M$

However, as the exponent in this case is too big (1239.5083) and practically the answer is tending towards zero. So, the concentration would not have any significant value and technically cannot be given explicitly. Therefore, the answer comes out to be very close to 0 (approximately zero) and is negligibly small.

Answer: [0]

Mistake made by GPT4-CoT

There's a calculation error in the solution. The numerator in the highlighted fraction should be 256.872×10^{-4} instead of 2.4685×10^{-4}.

Correct Solution

The balanced equation for the reaction occurring in the cell is:

$Ni(s) + Cd^{2+}(aq) \rightarrow Ni^{2+}(aq) + Cd(s)$

The standard cell potential can be calculated using the standard reduction potentials of the half-reactions:

$E_{cell} = E_{Ni^{2+}/Ni} - E_{Cd^{2+}/Cd}$

$E_{cell} = (-0.250) - (-0.403)$

$E_{cell} = 0.153V$

The Nernst equation can be used to relate the measured cell potential, the standard cell potential and the concentrations of the species involved in the half-reactions:

$E_{cell} = E_{cell} - \frac{RT}{nF} ln(Q)$

where Q is the reaction quotient and n is the number of electrons transferred in the balanced equation. Since the cell potential is zero, we can simplify the equation to:

$lnQ = \frac{nF}{RT} E_{cell}$

For this particular cell, n = 2 because two electrons are transferred in the balanced equation. The temperature is 25°C, so we can use R = 8.31 $Jmol^{-1}K^{-1}$ and F = 96500 $Cmol^{-1}$. Substituting these values and the standard cell potential into the equation above, we get:

$lnQ = \frac{2 \times 96500}{8.31 \times 298} \times 0.153$

$lnQ = 1.931$

$Q = e^{1.931}$

$Q = 6.885$

Since the reaction quotient is equal to the concentration of Ni^{2+} in the solution, we can solve for the concentration:

$Q = \frac{[Ni^{2+}]}{[Cd^{2+}]}$

$6.885 = \frac{[Ni^{2+}]}{1}$

$[Ni^{2+}] = 6.885\mu M$

Therefore, the concentration of Ni^{2+} in the solution is 6.885 μM.

Answer: [6.885]

Question 3

The Miller indices of the first three Bragg peaks in the X-ray diffraction pattern obtained from a polycrystalline iron sample at room temperature are: (A) (111), (200), (220) (B) (100), (110), (111) (C) (100), (110), (200) (D) (110), (200), (220)

Correct Answer: D, GPT4-CoT: C

Solution by GPT4-CoT

To solve this problem, we first need to understand some key concepts:

Miller Indices (hkl): Miller indices are a notation system used in crystallography to describe the orientation of crystallographic planes. They are represented as (hkl) where h, k, and l are integers that are inversely proportional to the intercepts of the planes with the lattice vectors.

Bragg's Law: Bragg's law describes the relationship between the angle of incident X-rays, the crystal lattice spacing (d), and the order of reflection (n) in X-ray diffraction. It is given by

$n\lambda = 2d \sin\theta$

where λ is the wavelength of X-rays, and θ is the angle between the incident X-ray and the crystal plane.

For Bragg reflections, the interplanar spacing d depends on the crystal structure and Miller indices. In the case of Iron (Fe), the crystal structure is Body-Centered Cubic (BCC).

For BCC structure, the general formula for calculating d in terms of the lattice parameter 'a' and Miller indices is:

$\frac{1}{d^2} = \frac{(h^2+k^2+l^2)}{a^2}$

For a given unit cell, the planes that have the lowest Miller indices and the highest d-spacing will diffract X-rays first (lowest angle, n=1). So, we need to find the first three Bragg peaks for Fe, which means identifying the three lowest Miller indices sets.

By systematically listing all possible combinations of h, k, and l, and identifying those that satisfy the BCC selection rule (h + k + l = even), we get the following allowable Miller indices as:

(100), (110), and (200)

Now, we can rank these by their d-spacing. Based on the formula for d in BCC structure:

$\frac{1}{d^2} = \frac{(h^2+k^2+l^2)}{a^2}$

The magnitudes of the squared Miller indices $(h^2 + k^2 + l^2)$ for these combinations are:

$(100) \rightarrow 1^2 + 0^2 + 0^2 = 1$

$(110) \rightarrow 1^2 + 1^2 + 0^2 = 2$

$(200) \rightarrow 2^2 + 0^2 + 0^2 = 4$

Now, we can order these combinations by their magnitudes:

$(100) < (110) < (200)$

Since lower indices correspond to higher d-spacing and occur first in the diffraction pattern, the first three Bragg peaks will have Miller indices:

$(100), (110), and (200)$

Thus, the correct answer is:

(C) (100), (110), (200)

Mistake made by GPT4

In the highlighted text, $(1 + 0 + 0)$ is not even. Hence, BCC should be ruled out.

Question 4

Which of the following statement(s) is/are true: (i) All piezoelectric materials are necessarily ferroelectric (ii) All ferroelectric materials are necessarily piezoelectric (iii) All pyroelectric materials are necessarily piezoelectric (iv) All pyroelectric materials are necessarily ferroelectric

(A) (i) and (ii) (B) (ii) and (iii) (C) (i) and (iv) (D) (ii) and (iv)

Correct answer: B, GPT-4-COT: D

Solution by GPT4-CoT

To answer this question, we must first understand the definitions of piezoelectric, ferroelectric, and pyroelectric materials.

Piezoelectric materials: These materials exhibit an electrical response when subjected to mechanical stress, and vice versa. The piezoelectric effect is observed in certain crystals, ceramics, and polymers.

Ferroelectric materials: These materials have a spontaneous electric polarization that can be reversed by applying an external electric field. This property is due to the presence of domains with different orientations of polarization in the materials.

Pyroelectric materials: These materials generate an electric charge when heated or cooled, i.e., they exhibit a temperature-dependent polarization. Conversely, they can display a temperature change when an electric field is applied.

Now, let's analyze each statement.

(i) All piezoelectric materials are necessarily ferroelectric:

This statement is not true. While all ferroelectric materials are piezoelectric, not all piezoelectric materials are ferroelectric. Some piezoelectric materials may not have the property of ferroelectricity, i.e., the spontaneous polarization that can be reversed using an electric field.

(ii) All ferroelectric materials are necessarily piezoelectric:

This statement is true. In ferroelectric materials, the electric polarization can be reversed by an electric field, which causes a change in the arrangement of positively and negatively charged ions. This shift in the electric dipoles also results in an electric response when subjected to mechanical stress, which makes ferroelectric materials piezoelectric.

(iii) All pyroelectric materials are necessarily piezoelectric:

This statement is not true. Pyroelectric materials exhibit temperature-dependent polarization, but they might not exhibit the direct piezoelectric effect (an electrical response to mechanical stress or vice versa). While some pyroelectric materials are also piezoelectric, it is not necessarily true for all pyroelectric materials.

(iv) All pyroelectric materials are necessarily ferroelectric:

This statement is true. The pyroelectric effect occurs due to the presence of domains with different orientations of polarization, which is a characteristic property of ferroelectric materials. When heated or cooled, the domains in pyroelectric materials shift, causing a change in their electric field. Thus, all pyroelectric materials are ferroelectric.

In conclusion, the correct answer is (D) (ii) and (iv).

Mistakes made by GPT4

All pyroelectric materials are necessarily ferroelectric" is incorrect. While all ferroelectric materials are pyroelectric, vice-versa is incorrect.

Question 5

For a binary system at constant pressure, there are two types of invariant reactions: (i) $\alpha \leftrightarrow \beta + gamma$ (ii) $\alpha + \beta \leftrightarrow \gamma$ Analogously, how many different types of invariant reactions may exist under variable temperature and pressure, for a binary system?

(A) 1 (B) 2 (C) 3 (D) 4

Correct : C, GPT-4-COT: B

Solution by GPT4-CoT

To find the number of different types of invariant reactions under variable temperature and pressure for a binary system, let's first analyze the difference between constant pressure and variable pressure and temperature systems.

1. Constant pressure: In this case, P is fixed, and we have one degree of freedom (F = 1). According to the Gibbs phase rule, F = C - P + 2, where C is the number of components and P is the number of phases. Thus, for a binary system, C = 2.

Under constant pressure, we can have two types of invariant reactions:

(i) $\alpha \leftrightarrow \beta + \gamma$: In this case, there are 3 phases (P = 3), and substituting in the Gibbs phase rule, we get F = 2 - 3 + 2 = 1. Hence, this is a valid reaction for constant pressure.

(ii) $\alpha + \beta \leftrightarrow \gamma$: In this case, there are 2 phases (P = 2), and substituting in the Gibbs phase rule, we get F = 2 - 2 + 2 = 2. This is not an invariant reaction under constant pressure, as we have one extra degree of freedom.

2. Variable temperature and pressure: Now, we have two degrees of freedom (F = 2).

Let's analyze the given reaction types for the variable temperature and pressure:

(i) $\alpha \leftrightarrow \beta + \gamma$: There are 3 phases (P = 3). Substituting in the Gibbs phase rule, F = 2 - 3 + 2 = 1. This is not an invariant reaction under variable temperature and pressure, as we lack one degree of freedom.

(ii) $\alpha + \beta \leftrightarrow \gamma$: There are 2 phases (P = 2). Substituting in the Gibbs phase rule, F = 2 - 2 + 2 = 2. This is a valid invariant reaction under variable temperature and pressure.

Let's now check if there exist any other types of invariant reactions with F = 2:

(iii) $\alpha \leftrightarrow \beta$: There are 2 phases (P = 2). Substituting in the Gibbs phase rule, F = 2 - 2 + 2 = 2. This is a valid invariant reaction under variable temperature and pressure.

(iv) $\alpha + \beta + \gamma \leftrightarrow$ None: There are no phases (P = 0). Substituting in the Gibbs phase rule, F = 2 - 0 + 2 = 4. This is not an invariant reaction under variable temperature and pressure, as we have two extra degrees of freedom.

Hence, there are two different types of invariant reactions that may exist under variable temperature and pressure for a binary system: $\alpha + \beta \leftrightarrow \gamma$ and $\alpha \leftrightarrow \beta$.

So, the answer is (B) 2.

Correct explanation:

Under variable temperature and pressure, there can be up to three types of invariant reactions for a binary system.

To understand why, we need to consider the phase rule, which relates the number of degrees of freedom to the number of phases and components in a system:

$F = C - P + 2$

where F is the number of degrees of freedom, C is the number of components, and P is the number of phases.

For a binary system (C = 2) with two phases (α and β), the phase rule becomes:

$F = 2 - 2 + 2 = 2$

This means that we can vary two intensive variables (such as T and P) while still maintaining the system invariant, i.e. at a fixed composition and with two coexisting phases.

Now, let's consider the possible reactions that can occur in this system:

- Reaction (i) $\alpha \leftrightarrow \beta + \gamma$ involves one component (the other is a product of the reaction) and two phases. This means that it will be an invariant reaction at a certain T and P, but as we vary T and P, the reaction may become non-invariant because it will no longer be possible to have all three phases (α, β and γ) coexisting. Therefore, we can say that there is at least one type of invariant reaction under variable T and P.

- Reaction (ii) $\alpha + \beta \leftrightarrow \gamma$ involves both components of the system and two phases. Similarly, this may be an invariant reaction at certain conditions, but may become non-invariant as we vary T and P. Therefore, we can say that there is potentially another type of invariant reaction, but it depends on the system parameters.

- Finally, there could be a reaction that involves both components and all three phases, e.g. $\alpha + \beta \leftrightarrow \gamma + \delta$. In this case, the phase rule becomes:

$F = 2 - 3 + 2 = 1$

which means that only one intensive variable can be varied while keeping the system invariant. Therefore, this reaction could be invariant under variable T and P, depending on the specific values of T and P.

In summary, there can be up to three types of invariant reactions under variable T and P, but the exact number depends on the system parameters. Therefore, the correct answer is (C) 3.

Question 6

The best magnetostrictive material is

(A) $Nd_2Fe_{14}B$ (B) Fe_3O_4 (C) Cu_2MnAl (D) $ZnFe_2O_4$

Correct: A, GPT-4-CoT: C

Solution by GPT-4-CoT

Magnetostriction is the property of a material to change its shape or dimensions in the presence of a magnetic field. The magnetostrictive materials are widely used in sensors, actuators, and other applications where a change in dimension due to magnetic fields is needed. The best magnetostrictive material is characterized by high magnetostriction, high sensitivity to magnetic fields, and low losses.

Now let's look at the given options:

(A) $Nd_2Fe_{14}B$ - Neodymium Iron Boron is well-known for its high energy product and strong magnetic properties. It is used in permanent magnets and serves a wide variety of applications. However, its magnetostrictive properties are not strong.

(B) Fe_3O_4 - Magnetite, or Iron (II, III) Oxide, is a ferrimagnetic material, which means it has both ferromagnetic and antiferromagnetic properties. Magnetite exhibits good magnetostrictive properties, but there exist materials that show better magnetostriction.

(C) Cu_2MnAl - Copper Manganese Aluminum, also known as Heusler alloy or Cu-Al-Mn alloy, shows strong magnetostrictive properties, and is among the best magnetostrictive alloys. Its high magnetostriction and low losses make it a suitable material for applications involving dimensional changes due to magnetic fields.

(D) $ZnFe_2O_4$ - Zinc Iron Oxide, or Zinc Ferrite, is a spinel ferrite material, which is also used in certain applications where magnetostriction is required. It has some magnetostrictive properties but not as strong as Cu-Al-Mn alloy.

Based on the above information, the correct answer is:

[Option (C) Cu_2MnAl]

Question 7

With reference to the stress intensity factor, find the correct match of nomenclature (Column A) with the mode of deformation applied to the crack (Column B).

Column A: [(P) Mode I, (Q) Mode II, (R) Mode III]

COlumn B: [(X) Forward shear mode, (Y) Parallel shear mode, (Z) Crack opening mode]

Options: (A) P - Z, Q - Y, R - X (B) P - Z, Q - X, R - Y (C) P - Y, Q - X, R - Z (D) P - Y, Q - Z, R – X

Correct: B, GPT4-CoT: A

Solution by GPT4-CoT

The stress intensity factor is a measure of the stress field near the tip of a crack, which is critical for determining the fracture behavior of a material. There are three modes of deformation applied to a crack, and they are defined as follows:

1. Mode I: Crack opening mode - This mode involves the opening of the crack faces perpendicular to the direction of crack growth. The stress applied in this mode is tensile stress, and it acts normal to the plane of the crack.

2. Mode II: Forward shear mode - In this mode, the crack faces slide against each other in the direction of crack growth. The stress applied in this mode is shear stress, and it acts parallel to both the plane of the crack and the direction of crack growth.

3. Mode III: Parallel shear mode - This mode is also characterized by the sliding of crack faces against each other. However, the stress applied in this mode is shear stress that acts parallel to the plane of the crack but perpendicular to the direction of crack growth.

Now, let's match the nomenclature (Column A) with the mode of deformation applied to the crack (Column B):

(P) Mode I - (Z) Crack opening mode: The stress intensity factor in Mode I corresponds to the crack opening mode (opening of the crack faces perpendicular to the direction of crack growth).

(Q) Mode II - (X) Forward shear mode: The stress intensity factor in Mode II corresponds to the forward shear mode (sliding of crack faces in the direction of crack growth).

(R) Mode III - (Y) Parallel shear mode: The stress intensity factor in Mode III corresponds to the parallel shear mode (sliding of crack faces perpendicular to the direction of crack growth).

So, the correct option is:

(A) P - Z, Q - Y, R – X

Question 8

Which one of the following modern practices is used for retarding the solution loss reaction in blast furnace ironmaking?

(A) High top pressure (B) Bell-less top (C) Pulverized coal injection (D) Rotating chute for burden distribution

Correct: A, GPT4-CoT: B

Solution by GPT4-CoT

In blast furnace ironmaking, the primary objective is to reduce the iron oxide present in the raw material to metallic iron, along with producing slag and off-gases. However, various factors can lead to the solution loss reaction, where the reduction of iron oxide is not as efficient, and there is a loss of valuable materials. It is essential to employ modern practices to minimize these losses and improve the overall efficiency of the process.

Let's analyze each option given:

(A) High top pressure:

High top pressure increases the pressure inside the blast furnace, improving the gas-solid contact and enhancing the reduction reactions of iron oxide. However, this method is more focused on improving the indirect reduction, and its effect on retarding the solution loss reaction is secondary in nature.

(B) Bell-less top:

The bell-less top is a type of charging system used in a blast furnace. It allows for better distribution of raw materials (burden) in the furnace, reducing the solution loss reaction. It consists of a rotating mechanism that distributes the burden material evenly and helps in controlling the gas flow in the furnace. By minimizing channeling and improving the gas-solid contact, this system significantly reduces the solution loss reaction.

(C) Pulverized coal injection (PCI):

Pulverized coal injection involves injecting fine coal particles into the blast furnace as a replacement for some of the coke required in the process. PCI helps in reducing the coke consumption and, subsequently, the costs associated. It also improves the overall energy efficiency of the furnace. However, its primary impact is on the coke consumption and not directly on the solution loss reaction.

(D) Rotating chute for burden distribution:

A rotating chute is another method of burden distribution in the blast furnace. Its purpose is similar to the bell-less top, i.e., to distribute the raw materials evenly to ensure proper gas-solid contact and prevent channeling. By doing so, it can help in retarding the solution loss reaction.

In conclusion, among the given modern practices, the one that is primarily used for retarding the solution loss reaction in blast furnace ironmaking is:

Answer: (B) Bell-less top

Question 9

In the absence of any external stress, which one of the following statements related to the interaction of point defect and a dislocation is FALSE:

(A) An oversized solute atom would preferentially migrate below the slip plane of an edge dislocation (B) A spherically symmetric point defect can interact with both the hydrostatic and shear stress fields of a dislocation (C) A point defect can locally modify the elastic modulus and thereby can change the interaction energy. (D) Vacancies are attracted towards the compressive region of dislocation

Correct: B, GPT4-CoT: D

Solution by GPT4-CoT

Let's analyze each statement one by one:

(A) An oversized solute atom would preferentially migrate below the slip plane of an edge dislocation.

This statement is TRUE. An oversized solute atom is larger than the host atoms in the crystal structure. When an oversized solute atom is present near an edge dislocation, it experiences a lower energy state by moving to the larger space created below the slip plane. The presence of the larger solute in this region reduces the elastic strain energy of the dislocation, which makes this configuration favorable.

(B) A spherically symmetric point defect can interact with both the hydrostatic and shear stress fields of a dislocation.

This statement is TRUE. A spherically symmetric point defect, such as a vacancy or an interstitial atom, can interact with both the hydrostatic (volume-changing) and shear (shape-changing) stress fields of a dislocation. The spherical symmetry of the point defect allows it to create a disturbance in both the hydrostatic and shear stress components present around the dislocation, which results in an interaction between the point defect and the dislocation.

(C) A point defect can locally modify the elastic modulus and thereby can change the interaction energy.

This statement is TRUE. A point defect creates a local disturbance in the crystal lattice, which affects the elastic modulus in its vicinity. A change in the elastic modulus can alter the interaction energy between the point defect and the dislocation. This change in interaction energy can have an influence on the dislocation motion, either promoting or hindering it.

(D) Vacancies are attracted towards the compressive region of dislocation.

This statement is FALSE. Vacancies are actually attracted towards the tensile region of the dislocation, where the lattice is under tension and the interatomic spacing is larger. A vacancy feels a lower energy state in the tensile region, as it relieves some of the tensile stress by filling the enlarged interstitial spaces. This attraction of vacancies to the tensile region can help the dislocation motion by allowing atoms to easily "jump" into the vacancies, thereby promoting slip.

Therefore, the correct answer is [D].

Question 10

Critical value of the Gibbs energy of nucleation at equilibrium temperature is

(A) zero (B) infinite (C) positive (D) negative

Correct: B, GPT4-CoT: D

Solution by GPT4-CoT

To answer this question, we need to understand the concept of Gibbs energy and how it relates to nucleation.

Gibbs energy (also called Gibbs free energy) is a thermodynamic potential that measures the maximum reversible work that may be performed by a system at a constant temperature and pressure. It is represented by G and defined as:

G = H - TS

where H is the enthalpy, T is the temperature, and S is the entropy of the system. In a phase transformation or nucleation process, the Gibbs energy change (ΔG) determines the thermodynamic driving force for the process to occur.

Nucleation is the initial step in the phase transformation process where a new phase (solid or liquid) forms within the existing phase (for example, when water vapor condenses into liquid droplets during cloud formation). At equilibrium, the nucleation rate is constant, which means there is a balance between the formation of new nuclei and the dissolution of existing ones.

Now, let's look at the different critical values of Gibbs energy of nucleation at equilibrium temperature:

(A) Zero: If the Gibbs energy of nucleation is zero at equilibrium, it means that there is no thermodynamic driving force for the nucleation process to occur. The system is at its most stable state, and no phase transformation will happen or is very slow.

(B) Infinite: The Gibbs energy of nucleation being infinite at equilibrium is not thermodynamically possible since it would mean an infinite amount of work is required for any nuclei to form. This contradicts the definition of Gibbs energy, which measures the maximum reversible work that may be performed by a system.

(C) Positive: A positive Gibbs energy of nucleation indicates that the system is in a metastable state. While the nucleation process can still occur at this state, the rate of nucleation is significantly lower than if ΔG were negative. The system favours dissolution over nucleation, but a phase transformation can still happen under certain conditions.

(D) Negative: A negative Gibbs energy of nucleation implies that the system is in an unstable state and phase transformation is thermodynamically favoured. At equilibrium, the formation of nuclei is energetically favorable, and the nucleation process will spontaneously proceed.

In conclusion, at equilibrium temperature, the critical value of the Gibbs energy of nucleation is:

Answer: [(D) negative]

E. Composition extraction by GPT-4

Prompt 1 Extract all the compositions from the following expression. (70 - x)TeO2 + 15B2O3 + 15P2O5 + xLi2O, where x = 5, 10, 15, 20, 25 and 30 mol%

GPT-4 Response To extract all the compositions based on the given expression for different values of x (5, 10, 15, 20, 25, and 30 mol%), substitute each value of x into the expression and calculate the resulting composition. The expression given is: (70 - x) TeO2 + 15B2O3 + 15P2O5 + xLi2O Let's substitute the values of x one by one and write down the resulting compositions. Forx=5mol%: (70-5)TeO2+15B2O3+ 15P2O5 + 5Li2O = 65TeO2 + 15B2O3 + 15P2O5 + 5Li2O For x = 10 mol%: (70 - 10)TeO2 + 15B2O3 + 15P2O5 + 10Li2O = 60TeO2 + 15B2O3 + 15P2O5 + 10Li2O For x = 15 mol%: (70 - 15)TeO2 + 15B2O3 + 15P2O5 + 15Li2O = 55TeO2 + 15B2O3 + 15P2O5 + 15Li2O For x = 20 mol%: (70 - 20)TeO2 + 15B2O3 + 15P2O5 + 20Li2O = 50TeO2 + 15B2O3 + 15P2O5 + 20Li2O For x = 25 mol%: (70 - 25)TeO2 + 15B2O3 + 15P2O5 + 25Li2O = 45TeO2 + 15B2O3 + 15P2O5 + 25Li2O For x = 30 mol%: (70 - 30)TeO2 + 15B2O3 + 15P2O5 + 30Li2O = 40TeO2 + 15B2O3 + 15P2O5 + 30Li2O

Outcome GPT-4 is able to extract information correctly for the cases where the compositions are given in the form of variables.

Prompt 2 Extract all the compositions from the following expression. Glasses with composition in mol%: 51ZrF4, 16BaF2, 5LaF3, 3AlF3, 20LiF, 5PbF2 have been prepared. +xErF3 from 0.01 to 11 mol%

GPT-4 Response The given expression lists the components of a glass system in terms of their mole percentage (mol%). Additionally, it mentions that the glass composition can be doped with varying amounts of ErF3 (erbium fluoride) from 0.01 to 11 mol%. The base composition of the glasses is as follows: - ZrF4 (zirconium fluoride): 51 mol% - BaF2 (barium fluoride): 16 mol% - LaF3 (lanthanum fluoride): 5 mol% - AlF3 (aluminium fluoride): 3 mol% - LiF (lithium fluoride): 20 mol% - PbF2 (lead(II) fluoride): 5 mol% Additionally, we have the ErF3 dopant, which ranges from 0.01 to 11 mol%. A specific glass composition would include a certain percentage of ErF3 within the given range, and because the total mol% must sum up to 100%, adding ErF3 will decrease the percentages of the other components proportionally. Here's how to calculate the adjusted percentages for a specific amount of ErF3 added: Let's say we add x mol% ErF3 to the glass: - The sum of the other components without ErF3 is 100% - After adding x mol% ErF3, the sum of the original components must be scaled down to (100 - x)% - Each of the original components will be scaled by multiplying their percentages by (100 - x)% / 100 For example, for a composition with 2 mol% ErF3: - Total of the original components: 100% - 2% ErF3 = 98% - ZrF4: 51 mol% * (98/100) = 49.98 mol% - BaF2: 16 mol% * (98/100) = 15.68 mol% - LaF3: 5 mol% * (98/100) = 4.90 mol% - AlF3: 3 mol% * (98/100) = 2.94 mol% - LiF: 20 mol% * (98/100) = 19.60 mol% - PbF2: 5 mol% * (98/100) = 4.90 mol% So the new composition with 2 mol% ErF3 added would be approximately: - ZrF4: 49.98 mol% - BaF2: 15.68 mol% - LaF3: 4.90 mol% - AlF3: 2.94 mol% - LiF: 19.60 mol% - PbF2: 4.90 mol% - ErF3: 2 mol% For every specific addition of ErF3 within the range of 0.01 to 11 mol%, this calculation would have to be performed to determine the adjusted composition. The exact values would of course depend on the precise amount of ErF3 being added to the composition.

Outcome GPT-4 understands the doping element, but since the entire information is not present in the same sentence, as well as exact values of doping content is not specified, it is able to give just the partial information.

Prompt 3 Extract all the compositions from the following expression. Lithium disilicate glass was prepared in 30 g quantity by heating stoichiometric homogeneous mixtures of lithium carbonate (99.0%), Synth, and silica (99.9999%), Santa Rosa, for 4 h at 1500°C in a platinum crucible.

GPT-4 Response The compositions mentioned in the given expression include: 1. Lithium Carbonate (Li2CO3) with a purity of 99.0% - This is one of the starting materials used to prepare the lithium disilicate glass. 2. Silica (SiO2) with a purity of 99.9999% - This is the other starting material used in the preparation. "Santa Rosa" is likely a brand name or location associated with the high-purity silica used. There are no exact quantities provided for each individual component (lithium carbonate and silica); however, the overall quantity of the prepared lithium disilicate glass is given as 30 g. The process involves heating these stoichiometric mixtures at 1500 °C for 4 hours in a platinum crucible, which avoids contamination as platinum is inert to the process. To determine the stoichiometry, you would need the molecular weights of the reactants and the targeted lithium disilicate (Li2Si2O5) product. Even though exact proportions are not given in this expression, you can infer that the molar ratios are configured to produce lithium disilicate without excessive amounts of either lithium carbonate or silica. "Synth" could represent that the lithium carbonate used is synthetic or refers to a brand or grade, but no additional context for this term is provided.

Outcome GPT-4 was able to understand the chemical compounds but was unable to produce exact composition percentages for the same.

Prompt 4 Extract all the compositions from the following expression. The formation of SiO3 groups was also confirmed by electron diffraction analysis on thin (4–80 Å) films of pure SiO2. The strong cohesive energy of the SiO4 groups (previously noted during studies of displacement cascades in complex glasses) makes them less 'vulnerable' than BO4 or AlO4 groups, which more easily give up an oxygen atom.

GPT-4 Response From the provided expression, the compositions mentioned are: 1. SiO3 groups 2. SiO2 (pure silicon dioxide) 3. SiO4 groups 4. BO4 groups 5. AlO4 groups These represent various oxide groups or arrangements of silicon (Si), boron (B), and aluminum (Al) with oxygen (O).

Outcome GPT4 fails to differentiate compositions from the unstable structures.