

# How Does a Generative Large Language Model Perform on Domain-Specific Information Extraction?—A Comparison between GPT-4 and a Rule-Based Method on Band Gap Extraction

Published as part of *Journal of Chemical Information and Modeling* special issue “Harnessing the Power of Large Language Model-Based Chatbots for Scientific Discovery”.

Xin Wang, Liangliang Huang, Shuozi Xu, and Kun Lu\*



Cite This: *J. Chem. Inf. Model.* 2024, 64, 7895–7904



Read Online

ACCESS |

Metrics & More

Article Recommendations



**ABSTRACT:** The advent of generative Large Language Models (LLMs) has greatly impacted the field of Natural Language Processing. However, it is inconclusive how generative LLMs perform on domain-specific information extraction tasks. This study compares the performance of GPT-4 and a rule-based information extraction method based on ChemDataExtractor on band gap information extraction, a task that has important implications for the materials science domain. No training data is required for either method, which is desirable because there is a lack of training data in the materials science domain compared with a variety of material information that is of interest. Manual evaluation on 415 randomly selected articles showed that the GPT-4 model achieved a higher level of accuracy in extracting materials' band gap information than the rule-based method (Correctness 87.95% vs 51.08%, Partial correctness 11.33% vs 36.87%, incorrectness 0.72% vs 12.05%). Further analysis of the errors reveals the strengths and weaknesses of the GPT-4 model compared to the rule-based method. The GPT-4 model shows stronger performance in interdependency resolution and complicated material name recognition, while it also has weaknesses in hallucination, identifying band gap values, and identifying band gap types. Revised prompt based on the error analysis leads to improved accuracy for GPT-4. To the best of our knowledge, this study is the first to compare the GPT-4 model and ChemDataExtractor for the band gap extraction task. This study provides evidence to support using generative LLMs for domain-specific information extraction tasks.

## INTRODUCTION

The field of Natural Language Processing (NLP) has been greatly influenced by the advent of Large Language Models (LLMs) such as BERT<sup>1</sup> and GPT<sup>2</sup> in recent years. What precedes the LLMs is the transformer architecture<sup>3</sup> that circumvents the sequential nature of recurrent models such as long short-term memory,<sup>4</sup> which facilitates parallelization during training. This paves the way for large-scale pretraining on existing textual corpora without human supervision. Results from this large-scale pretraining process, generally represented as text embeddings in a vector format, have been proved very useful for downstream NLP tasks. The pretraining and fine-tuning paradigm proposed by BERT has shown very impressive performance on a variety of NLP tasks.<sup>1,5,6</sup> With the introduction of ChatGPT in November 2022, the generative LLMs have started to gather more attention as the prompt

interface is widely applicable and has the potential to unify different NLP tasks.

Scientific literature holds a large volume of human knowledge in a textual format that is not amenable to traditional statistical analysis methods. It has been a longstanding application context of NLP techniques because a large-scale analysis of scientific knowledge from literature has the potential to guide scientific research and accelerate scientific discoveries.<sup>7</sup> Due to differences in terminology systems, language expression traditions, and

Received: May 20, 2024

Revised: September 27, 2024

Accepted: September 27, 2024

Published: October 8, 2024



concept structures, among others, NLP techniques need to be customized to different domains to be most effective. However, the application of NLP to scientific literature varies greatly across disciplines. The biomedical domain is likely the spearhead of this area with numerous infrastructures such as publication databases, ontologies, manually annotated training data, etc. On the other hand, the materials science domain, also a very important scientific field, has only started to examine this opportunity more recently.<sup>8</sup> One of the most common needs from materials science for NLP is information extraction. This is because lots of useful information about materials' composition, structures, and properties have been reported in the literature, while the sheer volume of publications presents a challenge for scientists to keep up to date on all developments. An information extraction system can automatically extract relevant information from text, aggregate the scattered information across articles, and organize the extracted information in a structured format that is easy to digest, query, and analyze. Extracting information from literature can also retain the provenance of the information if there is a need to view original articles. To achieve information extraction, the traditional NLP approach generally needs to accomplish two tasks: named-entity recognition (NER)<sup>9</sup> and relationship extraction (RE),<sup>10</sup> where NER recognizes entity mentions in the text and RE identifies the semantic relationship between entities. In the context of materials science, this often corresponds to recognizing material names and properties in the text (generally an NER task) and relating the correct material names with their corresponding properties (a RE task). This has not included extracting property values that are usually not a part of the NER task because numeric values are not generally considered entities. Before the advent of LLMs, supervised machine learning (ML) was the best approach to NER and RE in general NLP applications. However, materials science is a low-resource domain for NLP in that it has few well-established training data sets for supervised ML compared to various types of property and structure information that is of interest. Not to mention that the problem of extracting property values is not included in NER and RE. A rule-based approach has been used to tackle the problem of chemical information extraction.<sup>11</sup> By defining grammatical rules on how relevant information is generally reported in the literature, the ChemDataExtractor toolkit can extract chemical names, properties, and values from text and tables without the need of much training data. In a previous pilot study,<sup>12</sup> the ChemDataExtractor toolkit is extended by Ghosh and Lu to extract information about materials and their band gap information because band gap information is important to many material applications such as solar cells, light emitting diodes, and laser diodes. Their evaluation on a random sample of 415 articles showed 51.08% correctness, 36.87% partial correctness, and 12.05% incorrectness in the information extraction task. On the other hand, generative LLMs, such as GPT, have shown impressive performance in information extraction without the need for any training data.<sup>13–15</sup> However, their performance on domain-specific information extraction tasks is still under-researched, especially in the materials science domain, and therefore, it is not conclusive. It is of interest to understand how the generative LLM performs on a domain-specific information extraction task—band gap extraction. This has important implications for the materials science domain that generates a large amount of literature but lacks manually annotated training data for NLP. Therefore, this study aims to compare the performance of Ghosh and Lu's traditional NLP method<sup>12</sup> and GPT-4 on the

band gap extraction task to understand how applicable generative LLM is to domain-specific information extraction. This study sets out to address the following research questions:

1. How does a generative LLM (GPT-4) perform on a domain-specific information extraction—band gap extraction in the materials science domain? How does it compare with a more traditional rule-based NLP approach?
2. What are the strengths and weaknesses of the generative LLM in this specific information extraction task?

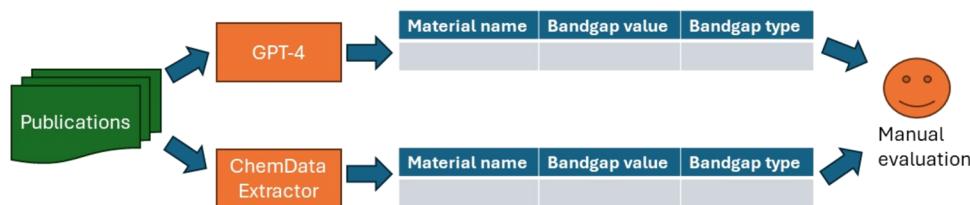
This study aims to compare the performance of GPT-4 and the rule-based approach in a more realistic setting for materials science where training data is not required and complicated fine-tuning is not applied. This offers a more realistic expectation of their performance in real world applications. To the best of our knowledge, this study is the first to compare GPT-4 and a rule-based information extraction method on band gap extraction. The findings from this study provide evidence on using generative LLMs for domain-specific information extraction tasks.

## ■ RELATED WORK

Our study uses an LLM for a domain-specific NLP task: band gap extraction. The most relevant work to this study can be categorized into two groups: information extraction in materials science and LLMs for information extraction. With generative LLMs, it is also possible to directly interrogate the LLMs with relevant scientific questions to assist experimental design or hypothesis generation.<sup>16,17</sup> This represents an emerging area of LLM applications and is summarized as well. However, it should be noted they are different from using LLMs for NLP tasks.

**Information Extraction in Materials Science.** Early information extraction research in materials science focused on chemical named entity recognition (CNER) task.<sup>18</sup> There are dictionary-based, rule-based, and ML-based approaches to the CNER task. The dictionary-based methods use predefined dictionaries (e.g., DrugBank<sup>19</sup>) to match text tokens, often yielding high precision but low recall due to the static nature of entity collections and the labor-intensive maintenance required.<sup>19–21</sup> Rule-based methods, while slightly more generalizable, rely on complex pattern matching rules from conventions like IUPAC,<sup>22</sup> facing challenges in portability and scalability with changing naming conventions. Contrastingly, ML-based methods leverage domain-specific corpora (e.g., CHEMDNER<sup>23</sup>) for training algorithms such as conditional random field (CRF) and support vector machine (SVM), demonstrating improved adaptability and effectiveness in chemical named entity detection.<sup>23–26</sup> The major drawback of ML-based approach is the dependency on large amounts of annotated training data, which are often not available in specialized domains such as materials science.

In the field of materials science, ML methods have been used to extract information from the literature, such as chemical formulas, material properties, and synthesis methods, which can significantly aid research and development. An NER model<sup>9</sup> was developed to extract key information from over 3 million materials science abstracts with an 87% accuracy rate. This approach identified 80 million entities and transformed abstract content into a structured database, facilitating simplified queries on complex literature questions. While NER forms the basis for identifying entities, RE extends this by identifying relationships between these entities, such as the association between materials

**Figure 1.** Overview of the research method.

and their properties or applications. A pipeline for extracting material property data from materials science papers<sup>27</sup> is built by using a custom language mode, trained on 2.4 million abstracts, to identify trends and phenomena in polymers and other materials.

Compared with ML methods, rule-based NER methods do not require large sets of training data, which makes them readily deployable with less preparatory work. ChemDataExtractor is a comprehensive toolkit for automatically extracting chemical entities and properties from scientific texts to populate chemical databases. It demonstrates enhanced NER and effective phrase parsing using rule-based grammar.<sup>11</sup> The overall precision of the NLP pipeline with ChemDataExtractor is at 66–73% for a magnetic materials database and around 80% for a battery materials database.<sup>28,29</sup> Then ChemDataExtractor is extended to extract band gap information from scholarly texts, aiding materials scientists in designing photovoltaic cells.<sup>12</sup> Moreover, almost at the same time, the modified Snowball algorithm of ChemDataExtractor was designed to autogenerate a database of semiconductor band gaps, achieving a weighted precision of 84%.<sup>30</sup>

A variety of methods employed by ChemDataExtractor utilize rule-based grammar to convert textual sentences into structured tree models, from which relevant information is subsequently extracted. This approach benefits from not needing training data, though its effectiveness is constrained by the finite number of grammar rules it can support. Natural language, inherently flexible and ambiguous, presents additional challenges. The development in generative LLM and zero-shot performance motivates us to evaluate the performance of ChatGPT on this task with no training data.

**LLMs for Information Extraction.** LLMs are advanced artificial intelligence systems designed to understand and generate human language, and therefore can be used to perform NLP tasks, such as information extraction. Among these, BERT is designed for understanding context in text.<sup>1</sup> This model and its variants have been broadly applied to NER and RE tasks in materials science, highlighting their performance in extracting structured information from unstructured data. A BERT model is customized to recognize named entities in materials science literature, achieving an 85% F-score.<sup>31</sup>

In the generative approach, models like GPT<sup>32</sup> have been pivotal in crafting coherent and contextually appropriate text-based responses. Building on these capabilities, a method is built which is initially trained an LLM using JSON-formatted abstracts and then is iteratively refined through manual corrections of preannotations.<sup>33</sup> This process significantly enhances the model's information extraction capabilities. Similarly, a toolkit named ChatExtract is a method using conversational LLMs for automated and accurate data extraction from material texts.<sup>15</sup> Their approach demonstrates near 90% precision and recall.

The primary advantage of generative models such as GPT is their ability to perform zero-shot learning, which means they can function without the need for additional, specialized training data. This characteristic significantly simplifies their implementation compared to other models like BERT that require deep expertise and extensive fine-tuning. Furthermore, zero-shot large language models (LLMs) are highly versatile: they can be used to extract information from both generic and specialized texts. However, it is inconclusive how the generative LLMs perform on domain-specific information extraction tasks, especially in the materials science domain. This study aims to fill in this gap.

**LLMs for Materials Science.** In addition to NLP tasks such as information extraction, LLMs have also been directly used to answer materials questions to assist scientific discoveries in materials science owing to their capabilities of understanding human language and generating relevant responses. For example, 14 examples of LLM applications relevant to materials science and chemistry is reported, including predicting properties of molecules and materials, designing novel interfaces, knowledge extraction, and developing educational applications.<sup>17</sup> Also, GPT-4 is tested for generating reasonable scientific hypotheses in materials science and found the model can make interesting hypotheses worthy of testing.<sup>16</sup> Beyond general purpose LLMs, fine-tuned LLMs on specific domains<sup>34,35</sup> have been developed and shown better performance on domain-specific tasks such as knowledge recall, hypothesis generation, and research assistance. The fine-tuned specialized LLMs can also be integrated using X-LoRA model<sup>36</sup> or the multiagent model<sup>37–39</sup> where multiple LLMs interact autonomously with each other playing different roles to complete complex tasks. LLMs can also be applied to construct structured knowledge graphs from text,<sup>40</sup> which can support downstream tasks such as graph reasoning.<sup>41</sup> These innovative uses of LLMs in materials science show great promise, especially as LLMs continue to improve in their ability in understanding language and the knowledge behind language.

## METHODS

The methods of this research is summarized in Figure 1.

**Data and Prompt.** To be comparable with Ghosh and Lu's rule-based method,<sup>12</sup> this study used the same 415 articles and asked ChatGPT to extract materials' band gap information from them. The 415 articles were a random sample of 1.44 million articles with filtering effort to ensure the articles contain band gap information.<sup>12</sup> The GPT-4 model was used as is with all default parameters. For each of the articles, we developed the following prompt to ask ChatGPT (GPT-4 model) to extract material names, their band gap values, and types from the title and abstract of the article:

*Extract material names and their corresponding band gap values and types from the text and list them in a tabular format.*

*Text: "{<title> + < abstract>}”*

Extract material names and their corresponding band gap values and types from the text and list them in a tabular format.

Text: "Evolution of structural and optical properties of rutile TiO<sub>2</sub> thin films synthesized at room temperature by chemical bath deposition method

Nanocrystalline thin films of TiO<sub>2</sub> were prepared on glass substrates from an aqueous solution of TiCl<sub>3</sub> and NH<sub>4</sub>OH at room temperature using the simple and cost-effective chemical bath deposition (CBD) method. The influence of deposition time on structural, morphological and optical properties was systematically investigated. TiO<sub>2</sub> transition from a mixed anatase–rutile phase to a pure rutile phase was revealed by low-angle XRD and Raman spectroscopy. Rutile phase formation was confirmed by FTIR spectroscopy. Scanning electron micrographs revealed that the multigrain structure of as-deposited TiO<sub>2</sub> thin films was completely converted into semi-spherical nanoparticles. Optical studies showed that rutile thin films had a high absorption coefficient and a direct bandgap. The optical bandgap decreased slightly (3.29–3.07 eV) with increasing deposition time. The ease of deposition of rutile thin films at low temperature is useful for the fabrication of extremely thin absorber (ETA) solar cells, dye-sensitized solar cells, and gas sensors."

**ChatGPT**

Here is the extracted information presented in a tabular format:

Material	Phase	Band Gap Type	Band Gap Value (eV)
TiO <sub>2</sub>	Rutile	Direct	3.29–3.07

This table summarizes the properties of rutile TiO<sub>2</sub> thin films, including the phase, band gap type, and band gap value as mentioned in the text. [-]

**Figure 2.** Example conversation with ChatGPT on band gap extraction.**Table 1. Evaluation Criteria on Band Gap Extraction are the Same as in Ghosh and Lu's Method<sup>12</sup>**

labels	definition
correct	The title and abstract have information about a material and its associated band gap value. In this scenario, both are extracted correctly and fully.
partially correct	The title and abstract have information about a film material and its associated band gap value. In this case, the extracted information includes correct material or band gap value, but not both. This category also includes the cases where both material and band gap value are extracted, but the value is not associated with the correct material. In addition, when multiple band gap values are reported, only a subset of them are extracted.
incorrect	The title and abstract have information about a material and its associated band gap value. In this case, neither information is correctly extracted.

Before the real experiment started, there were a few pilot tests to gauge ChatGPT's ability for this task. For ease of evaluation, we also asked ChatGPT to present the results in a tabular format. No training examples were provided to ChatGPT. Therefore, this is considered as a zero-shot performance. Below is an example of one conversation (Figure 2):

**Rule-Based Method.** Ghosh and Lu's rule-based method<sup>12</sup> extends ChemDataExtractor to extract material mentions along with their reported band gap values and types. A rule-based band gap parser is created to combine with the built-in compound parser from ChemDataExtractor to create a record of material name, band gap type, and band gap value. The band gap parser looks for patterns of how band gap information is reported in text. This includes three basic elements of the information, including cem (chemical entity mentions), affix (words that are usually used before and after the band gap value to describe it) and value\_andSpecifier (band gap value, unit, and band gap type), and the various orders of the elements:

1. Cem + affix + value\_andSpecifier,
2. Cem + value\_andSpecifier + affix,
3. Affix + value\_andSpecifier + cem,
4. Affix + cem + value\_andSpecifier,
5. Value\_andSpecifier + cem + affix,
6. Value\_andSpecifier + affix + cem

These orders aim to capture different ways the band gap information is reported in text. The rule-based method extracts the band gap information if the text matches any of the patterns.

More detailed descriptions of this method can be found in Ghosh and Lu's rule-based method.<sup>12</sup>

**Evaluation.** The results from ChatGPT were manually evaluated for their correctness. We used the same evaluation criteria as in Ghosh and Lu's method<sup>12</sup> which is summarized in the following table (Table 1)

The reliability of this annotation scheme has been validated with a high Cohen's Kappa value at 0.815.<sup>12</sup> When there is doubt, the annotator discussed with domain experts among the coauthors to resolve the issue.

## RESULTS

The comparison of accuracy between the GPT-4 model performance and that reported in Ghosh and Lu's method<sup>12</sup> is presented in Table 2. For the 415 articles, the GPT-4 model achieved 87.95% correctness, 11.33% partial correctness, and 0.72% incorrectness. It is worth mentioning that when we reviewed the ChatGPT results for evaluation and error analysis,

**Table 2. Comparison of Accuracy between Rule-Based Method<sup>12</sup> with Zero-Shot GPT-4 Performance**

	Ghosh & Lu (2023)	GPT-4
correct	212 (51.08%)	365 (87.95%)
partially correct	153 (36.87%)	47 (11.33%)
incorrect	50 (12.05%)	3 (0.72%)
total	415	415

**Table 3.** Breakdown “Partially Correct” Category for GPT-4

error category	subcategory	occurrence	percentage (%)	explanation
material name error	missed ratio value	1	2.13	“GaNAs” extracted, but should be “GaN <sub>1-x</sub> As $x$ ”
	missed material name	1	2.13	“topological systems” extracted without specific material names
	material names are not full name	6	12.77	“polymer” extracted without more specific information about what polymer it is
value error	missed some band gap values	12	25.53	when multiple band gap values are reported, some are missing
	missed band gap value range	12	25.53	when a range of band gap value is reported, but the range was not extracted
	extracted nonband-gap information	4	8.51	e.g., extracted “excitonic emission”
	extracted expressions rather than values	3	6.38	e.g., extracted “0.45 ≤ EC – E ≤ 1.05”
	missed value expression	1	2.13	“1.15” extracted when “1.15 ± 0.01” reported
interdependency error		7	14.89	mismatched the correct material names with their corresponding band gap values
total		47	100	

Material	Band Gap Value (eV)	Band Gap Type
Cu <sub>2</sub> ZnSnS <sub>4</sub> (CZTS, x=1)	1.63	Not specified
Cu <sub>2</sub> Zn0.75Ni0.25SnS <sub>4</sub>	Reduced as x decreases	Not specified
Cu <sub>2</sub> Zn0.5Ni0.5SnS <sub>4</sub>	Reduced as x decreases	Not specified
Cu <sub>2</sub> Zn0.25Ni0.75SnS <sub>4</sub>	Reduced as x decreases	Not specified
Cu <sub>2</sub> NiSnS <sub>4</sub> (x=0)	1.36	Not specified

**Figure 3.** GPT-4 results on the interdependency resolution example.

we noticed that sometimes ChatGPT automatically combines different prompts into one although our prompts were separately submitted for each article. In this case, we resubmitted the prompts and evaluated the performance of ChatGPT on the new results if the extracted results from the first attempt were from the combined articles. There were also a few cases where the extracted information could not be found in the combined articles. We considered these cases as incorrect.

## ■ ERROR ANALYSIS

To understand the strengths and weaknesses of the GPT-4 model, we conducted error analysis on where the model was partially correct or incorrect. We also compared it with the Ghosh and Lu’s error analysis<sup>12</sup> to understand how it compares to a rule-based method.

The three incorrect cases for the GPT-4 model are where it extracted a long list of materials and band gap values that are not in the text and did not extract the materials and band gap values from the text provided. These can be considered hallucinations for LLMs, a well-known problem of generative LLMs.<sup>42</sup> This problem does not exist in the rule-based method because it cannot extract information that does not exist in the provided text. In our experiment, it happens in 0.72% of the cases.

For the 47 partially correct cases, we manually categorized the error into three groups that are comparable to Ghosh and Lu’s method:<sup>12</sup> Material name error, value error, and interdependency error. Each of the three groups is further categorized into subcategories to provide a more detailed understanding of the errors (Table 3). Generally, the material name error corresponds to the cases where the names of the materials were not correctly identified. There were 8 such cases for GPT-4 compared to 50 in Ghosh and Lu’s method.<sup>12</sup> The value error corresponds to the cases where the band gap values were not correctly identified. It can be further categorized into missing some band gap values

(e.g., multiple values were reported but only some were extracted), missing band gap value range (e.g., a range was reported, but not extracted), extracted nonband-gap information, extracted expressions rather than values, or missing value expression etc. In total, there were 32 cases with value errors accounting for the largest percentage of errors in GPT-4 (68.09% of all GPT-4 errors). While in Ghosh and Lu’s method,<sup>12</sup> there were 40 cases with band gap value errors accounting for 19.70% of all errors in their rule-based method. The interdependency errors refer to the cases where after correctly identifying the materials and band gap values, the material names were related to the wrong band gap values in the text. This was the primary error that accounts for 55.67% of all errors observed in Ghosh and Lu’s method.<sup>12</sup> However, in the GPT-4 model, this error was only observed in 7 cases accounting for 15.56% of all errors.

## ■ DISCUSSION

The comparison between the GPT-4 model and a rule-based method allows us to observe the pros and cons of the generative LLM in this domain-specific task. Several strengths are observed in the GPT-4 model, along with some weaknesses as well.

### Strengths of the GPT-4 Model on Bandgap Extraction.

**Interdependency Resolution.** Interdependency errors were found to be the most common (55.67% of all errors) in Ghosh and Lu’s method<sup>12</sup> in their rule-based method. In fact, interdependency resolution is a more sophisticated NLP task because it requires the understanding of the context in addition to correctly identifying chemical entities and their property values. Sometimes, the context required to correctly extract band gap information spans across sentences. GPT-4 is very impressive in correctly identifying the interdependency between chemical entities and their reported properties. This shows an impressive ability to understand the context of the text. In fact, in

a few relatively complicated examples, GPT-4 was able to correctly link the band gap values with the correct materials. For example, for the following article:

*"Low-cost hydrothermal synthesis and characterization of pentanary Cu<sub>2</sub>Zn<sub>x</sub>Ni<sub>1-x</sub>SnS<sub>4</sub> nanoparticle inks for thin film solar cell applications"*

In the present investigation, Cu<sub>2</sub>(Zn<sub>x</sub>Ni<sub>1-x</sub>)SnS<sub>4</sub> nanoparticles were successfully synthesized by facile and low-cost hydrothermal method for different concentration of  $x = 1, 0.75, 0.50, 0.25, 0$  at 190 °C for 24 h. The probable reaction mechanism for the formation of Cu<sub>2</sub>Zn<sub>x</sub>Ni<sub>1-x</sub>SnS<sub>4</sub> nanoparticles is reported. The synthesized nanoparticles were characterized by X-ray diffraction (XRD), Raman spectroscopy, field emission scanning electron microscope (FESEM), transmission electron microscope (TEM), x-ray photoelectron spectroscopy (XPS) and ultraviolet-visible (UV-vis) spectrometric studies. The phase pure kesterite and stannite structures of synthesized nanoparticles were confirmed by XRD and Raman spectroscopy analysis. The structural transition from tetragonal kesterite copper zinc tin sulfide (Cu<sub>2</sub>ZnSnS<sub>4</sub>—CZTS) to stannite Cu<sub>2</sub>Zn<sub>x</sub>Ni<sub>1-x</sub>SnS<sub>4</sub> ( $x = 0$ ) occurs in the Cu<sub>2</sub>Zn<sub>x</sub>Ni<sub>1-x</sub>SnS<sub>4</sub> samples with nickel concentration ( $x$ ) in the range of 0.25–1. The results of SEM and TEM show the morphological features of the prepared Cu<sub>2</sub>Zn<sub>x</sub>Ni<sub>1-x</sub>SnS<sub>4</sub> nanoparticles. The optical band gaps of the prepared Cu<sub>2</sub>Zn<sub>x</sub>Ni<sub>1-x</sub>SnS<sub>4</sub> nanoparticles were reduced from 1.63 eV ( $x = 1$ ) to 1.36 eV ( $x = 0$ ). It falls well within the suitable band gap range of absorber materials to be useful for thin film solar cell applications. Photoresponse behavior of Cu<sub>2</sub>Zn<sub>x</sub>Ni<sub>1-x</sub>SnS<sub>4</sub> ( $x = 1$  and 0.75) thin films under dark and light illumination indicates the potential of Cu<sub>2</sub>Zn<sub>x</sub>Ni<sub>1-x</sub>SnS<sub>4</sub> nanoparticles as an alternate absorber material in low-cost thin film solar cells applications."

The GPT-4 model was able to successfully identify the materials with correct composition, and their corresponding band gap values are as follows (Figure 3):

This example shows that the GPT-4 model was able to understand the context where the  $x$  value varies for the different ratios of atoms and correctly links the  $x = 1$  to the bandgap value of 1.63 eV and  $x = 0$  to the band gap value of 1.36 eV. In addition, it understood the context that as  $x$  decreases, the band gap value decreases. It even substitutes the  $x$  values into the chemical formula for the different ratios. These kinds of cases are generally more challenging for the rule-based method. In Ghosh and Lu's rule-based method,<sup>12</sup> ChemDataExtractor only extracted the general material name "pentanary Cu<sub>2</sub>Zn<sub>x</sub>Ni<sub>1-x</sub>SnS<sub>4</sub>" and one of the band gap values 1.63 eV. It fails to understand the nuances of the context where  $x$  value varies and the band gap value decreases as  $x$  values. The ability of interdependency resolution reduced the number of errors in this category from 113 to 7 compared to Ghosh and Lu's.<sup>12</sup>

**Material Name Recognition.** Another strength we observed in the GPT-4 model is its ability to recognize complex material names without any fine-tuning. This demonstrates its domain-specific NER ability. ChemDataExtractor sometimes has difficulties in correctly identifying complex chemical names such as "Ge<sub>1-x-y</sub>Si<sub>x</sub>Sn<sub>y</sub>" (where  $x$  and  $y$  are used for the ratio of different elements).<sup>12</sup> However, the GPT-4 model was able to identify this entity correctly. Overall, GPT-4 reduced the number of compound name errors from 50 in the Ghosh and Lu's rule-based method<sup>12</sup> to 8. Out of the eight name errors, six

are for the cases of not using full material names, which we will discuss below.

### Weaknesses of the GPT-4 Model on Bandgap Extraction.

We also observed some weaknesses in the current GPT-4 model.

**Hallucination.** All 3 incorrect cases from GPT-4 are due to hallucinations where the extracted information does not exist in the given text. This is a problem unique to the generative LLM compared to the rule-based method. There is currently not a perfect solution to this problem although some mitigation methods exist.<sup>41</sup> In our experience, we only observed this in 0.72% of cases, but this shows the GPT-4 model is not always reliable.

**Band Gap Type.** Although our prompt asked ChatGPT to extract band gap types, we observed that the GPT-4 model only extracts band gap types that are "direct/indirect". Many articles reported optical band gap types in the text, but these were not extracted by the GPT-4 model (68 cases out of 415). We counted these cases as correct in our study. However, strictly speaking, it is beneficial to also extract the optical band gap type information because there is a distinction between optical band gap versus electronic band gap, and the former is usually lower. This may be addressed by appropriate prompt engineering, but it was an issue observed from our prompt.

**Band Gap Value Errors.** Band gap value errors account for the largest portion of errors by the GPT-4 model. Within this error category, the "Missed some band gap values" error and "Missed band gap value range" error are the most common. The former refers to the cases where multiple band gap values were reported in the text, but only some were extracted. This is also a common type of error observed in Ghosh and Lu's rule-based method<sup>12</sup> as the ways of reporting band gap values are very diverse, which poses challenges to information extraction. For the "Missed band gap value range," we found that the GPT-4 model seems to have difficulties in understanding range values. For example, for the article:

### "Magnetic and Optical Properties of Mn<sub>1-x</sub>Zn<sub>x</sub>Fe<sub>2</sub>O<sub>4</sub> Nanoparticles"

Mn<sub>1-x</sub>Zn<sub>x</sub>Fe<sub>2</sub>O<sub>4</sub> ( $x = 0.0\text{--}1.0$ ) NPs (MZF NPs) were synthesized by a citric acid assisted sol-gel process. MZF NPs show superparamagnetic characteristics at room temperature. Saturation magnetization ( $M_s$ ) of MnFe<sub>2</sub>O<sub>4</sub> NPs is 70.52 emu/g is very close to the bulk saturation magnetization value of 80 emu/g. The observed  $M_s = 35.90$  emu/g value for ZnFe<sub>2</sub>O<sub>4</sub> particles is much greater than the bulk  $M_s$  value of 5 emu/g. This case is attributed to cation distribution change from normal spinel to mixed structure. The small  $M_p/M_s$  ratios (the maximum 0.147) specify uniaxial anisotropy in the Mn<sub>1-x</sub>Zn<sub>x</sub>Fe<sub>2</sub>O<sub>4</sub> NPs. The average crystallite diameter ( $D_{mag}$ ) was evaluated from magnetic analyses. The obtained  $D_{mag}$  values are between 27.67 and 33.60 nm and this range is in great accordance with the results calculated from XRD measurements. Among the NPs, the samples with more zinc content show higher diffuse reflectance. The optical direct band gap of MZF NPs is found to decrease from 2.1 to 1.90 eV as the zinc content rises."

The GPT-4 model extracted (Figure 4):

It was only the start and end values of the range 2.1 to 1.90 eV. On the contrary, Ghosh and Lu's rule-based method<sup>12</sup> was able to extract the range correctly. We found that the GPT-4 model tends to extract only the start and end values of a range rather

Material	Band Gap Value (eV)	Band Gap Type
Mn <sub>1-x</sub> ZnxFe <sub>2</sub> O <sub>4</sub> Nanoparticles ( $x=0.0$ )	2.1 ( $x=0.0$ )	Direct
Mn <sub>1-x</sub> ZnxFe <sub>2</sub> O <sub>4</sub> Nanoparticles ( $x=1.0$ )	1.90 ( $x=1.0$ )	Direct

**Figure 4.** GPT-4 results on the band gap value example.

Material Type	Band Gap Value (eV)	Type	Notes
BiFeO <sub>3</sub> (BFO)	2.0	Not specified	Pure phase
Bi <sub>1-x</sub> Ba <sub>x</sub> FeO <sub>3</sub> ( $x=0.01$ )	Not specified	Not specified	Band gap decreases with increase in Ba content
Bi <sub>1-x</sub> Ba <sub>x</sub> FeO <sub>3</sub> ( $x=0.03$ )	1.68	Capable of responding to visible light	Higher photocatalytic activity than pure BFO
Bi <sub>1-x</sub> Ba <sub>x</sub> FeO <sub>3</sub> ( $x=0.05$ )	Not specified	Not specified	Band gap decreases with increase in Ba content
Bi <sub>1-x</sub> Ba <sub>x</sub> FeO <sub>3</sub> ( $x=0.10$ )	Not specified	Not specified	Band gap decreases with increase in Ba content

**Figure 5.** GPT-4 results on the overinterpretation example.

than the range itself from the text. This indicates that the model may have some weaknesses in understanding the range values.

There were two cases that stood out that may fit in both “Missed band gap value range” and “Interdependency resolution” errors, but they were counted as the former in our error analysis. In these two cases, we found that the GPT-4 model attempted to overinterpret the context in the text rather than directly extracting from the text. For example,

*Enhanced Photocatalytic Activity in Bi<sub>1-x</sub>Ba<sub>x</sub>FeO<sub>3</sub> Prepared by a PEG400 Assisted Sol–Gel Method*

*Ferroelectric Bi<sub>1-x</sub>Ba<sub>x</sub>FeO<sub>3</sub> nanoparticles for  $x = 0, 0.01, 0.03, 0.05$ , and 0.10 were synthesized by a polyethylene glycol 400 (PEG400) assisted sol–gel method. X-ray diffraction reveals that Bi<sub>1-x</sub>Ba<sub>x</sub>FeO<sub>3</sub> nanoparticles exhibit a distorted rhombohedral structure with the R3c space group, and the diffraction peaks shift upon incorporation of Ba. Transmission electron microscope analysis shows that the particle size of Bi<sub>1-x</sub>Ba<sub>x</sub>FeO<sub>3</sub> nanoparticles is in the range of 30–60 nm, decreasing with an increase in Ba content. Bi<sub>1-x</sub>Ba<sub>x</sub>FeO<sub>3</sub> nanoparticles have band gaps in the range of 1.68–2.0 eV, which are capable of responding to visible light irradiation. The rate of the photocatalytic degradation of Bi<sub>1-x</sub>Ba<sub>x</sub>FeO<sub>3</sub> nanoparticles for  $x = 0.03$  to methyl orange (MO) dye achieves about 81% under visible light irradiation for 3 h, which is higher than that of 66% for pure phase BiFeO<sub>3</sub>(BFO). Moreover, the effects of Ba<sup>2+</sup> modification on the band gap of BFO crystallites have been investigated and discussed.”*

The GPT-4 model extracted the following (Figure 5):

The text only mentioned the material Bi<sub>1-x</sub>Ba<sub>x</sub>FeO<sub>3</sub> has a band gap value in the range of 1.68–2.0 eV, but did not specify how the band gap value changes with different values of  $x$ . The GPT-4 model seems to overinterpret the context and associate BiFeO<sub>3</sub> (when  $x = 0$ ) to 2.0 eV and Bi<sub>1-x</sub>Ba<sub>x</sub>FeO<sub>3</sub> (when  $x = 0.03$ ) to 1.68 eV, even making notes on “Band gap decreases with increase in Ba content.” However, the information provided in the text does not support this. Simply extracting the range value from the text would have led to the correct result. In comparison, Ghosh and Lu’s rule-based method<sup>12</sup> was able to correctly extract the range value 1.68–2.0 eV although it

identified a wrong material name in this case (polyethylene glycol 400).

**Impact of Prompts.** It has been recognized that prompt engineering can improve the performance of generative LLMs. With what we learned from the error analysis, we attempted to generate new prompts using ChatGPT4’s web interface and tested their performance on previously problematic abstracts. We told ChatGPT what needed improvements based on observed errors in the initial experiment and iteratively refine the prompt. Once we finalized the prompt, we utilized the GPT-4-turbo API, inputting only the title and abstract of each article with the following new prompt and examined its impact on the performance:

*Read the following text and organize the extracted information into a table format. The table should list all mentioned materials and material groups along with any specific conditions under which different band gaps or band gap ranges are specified. Include the type of band gap if mentioned. If certain details like specific band gaps or band gap types are missing, use “none” to indicate missing information.*

*Text: [Title + Abstracts]*

*Instructions:*

1. Identify and list all the materials and material groups mentioned in the text.
2. For each material, include any specific band gaps or band gap ranges mentioned along with any specific conditions where the band gaps vary.
3. Determine and record both the electronic and optical band gap types if specified.
4. Organize the extracted information into a table with columns for Materials (including conditions), Band Gaps, Electronic Band Gap Types, and Optical Band Gap Types. If any column is missing information for a material, fill in “none”.

The results from the new prompt were then evaluated manually following the same protocol. The accuracy of the new prompt improves to 93.98%, with 0 incorrect, 25 partially correct, and 390 correct. This shows prompt engineering can

**Table 4.** Breakdown “Partially Correct” Category for Improved Prompt

error category	subcategory	occurrence	percentage (%)	explanation
material name error	material names are not full name	3	12	“glass” extracted, but should be “glass and glass ceramic samples with composition $55\text{SiO}_2 - 30\text{B}_2\text{O}_3 - (x)\text{Li}_2\text{O} - (15 - x)\text{Y}_2\text{O}_3$ ”
value error	missed band gap value range	2	8	when a range of band gap value is reported, but the range was not extracted
	extracted nonband-gap information	16	64	e.g., extracted “excitonic emission”
	right material names but wrong conditions	2	8	e.g., “ $\text{In}_x\text{S}_3$ films (thickness: 800–450 nm)” and “ $\text{In}_x\text{S}_3$ films (thickness: 50–30 nm)” Should be $\text{In}_x\text{S}_3$ films (varying thickness from 800–450 nm to 50–30 nm) 2.0 to 3.6 eV
interdependency error		2	8	mismatched the correct material names with their corresponding band gap values
total		25	100	

improve the performance of GPT-4 in the band gap extraction task.

Table 4 listed the breakdown of the “Partially Correct” category for the improved prompt. Compared to the original prompt, issues such as “Extracted information not existing in the text” were eliminated. Of the initial errors, 51.06% were attributed to missing band gap values and ranges. Among these specific errors, 58.33% occurred because the ChatGPT API response failed to include all band gap values for certain materials mentioned in the abstracts under specific conditions. Additionally, 20.83% of these errors involved cases where multiple band gap values would have been more accurately represented as a range. Our refined prompt emphasized extracting all mentioned materials/material groups and their conditions, listing corresponding band gaps/ranges. Consequently, there were no errors in missing band gap values and a reduced incidence of missing band gap ranges from 25.53 to 8%. However, this approach introduced a new drawback: the responses to the new prompt included 16 cases for extracting nonband-gap information (accounting for 64% of errors, a significant increase from the previous 8.89%). This likely stems from ChatGPT’s insufficient knowledge base about band gaps, merely extracting all possible related information without secondary validation. In the future, enhancing ChatGPT’s ability to discern band gap information could address this issue.

**Other LLM Applications in Science.** This study applies an LLM to a domain-specific NLP task: Band gap extraction. This application is in the realm of using LLMs for NLP tasks, of which the results are useful for the material science domain. Language is a carrier of knowledge. LLMs, trained on a vast volume of text, have shown impressive understanding of the underlying meaning and have been used to directly answer domain questions. For example, fine-tuned LLMs have shown superior capability in addressing domain tasks such as knowledge recall, hypothesis generation, research assistance and material design and property prediction than general LLMs<sup>34</sup> and conventional machine learning models.<sup>17</sup> In addition to fine-tuning, techniques such as Retrieval Augmented Generation, Prompt Engineering, and Multiagents methods have been used to further improve the performance of LLMs in addressing domain questions.<sup>37,38,40</sup> The multiagent methods are particularly useful for complicated and open-ended problems that may benefit from multiple perspectives and specialized roles.<sup>43</sup> Multiple LLMs are given different roles and profiles. Then, these LLMs autonomously communicate and interact with each other to progressively solve the given task. The multiagent methods allow the integration of multiple specialized LLMs to solve complex problems. In fact, not all agents have to be LLMs. LLMs can also

interact with other tools, systems, functions, APIs, data sources in the multiagent methods,<sup>39,44</sup> which makes this approach very powerful in integrating diverse resources. These applications of LLMs have shown promise. The continuous improvement of LLMs will only strengthen them in the future. It should be noted that this line of study is related but different from ours since our use of LLM is directly for NLP tasks and indirectly for domain questions.

## CONCLUSIONS

This study compares the performance of the GPT-4 model and a rule-based information extraction method on the band gap information extraction, a domain-specific information extraction task. The GPT-4 model shows impressive performance with no training data, prompt engineering, and parameter fine-tuning. Further analysis of the errors by the GPT-4 model and the rule-based information extraction method reveals the strengths and weaknesses of the generative LLM. Revised prompt based on the initial error analysis can further improve the accuracy of the GPT-4 model. To the best of our knowledge, this is the first study that compares the GPT-4 model and ChemDataExtractor for the band gap extraction task. This study adds to the evidence of using generative LLMs for domain-specific information extraction. Generative LLMs are progressing every day, and numerous efforts are underway to improve their performance. This study shows that with very little effort and training, GPT-4 can already perform at a very high level for domain-specific tasks. It is very likely that with increasing improvements, generative LLMs will become the first choice due to its ease-of-use conversational interface, no required fine-tuning/training, and versatility in performing different NLP tasks using the same interface.

**Limitation.** It is worth mentioning that this study does not intend to conclude that ChemDataExtractor cannot perform as well as the GPT-4 model for the band gap extraction task. In fact, ChemDataExtractor can be tuned and refined for better performance with snowball algorithm and manual fine-tuning.<sup>30</sup> In a study that is independent of Ghosh and Lu’s study,<sup>12</sup> Dong and Cole achieved a weighted precision of 84% and a weighted recall of 65% in band gap extraction using optimized ChemDataExtractor. However, their method takes much more effort and requires annotated training data for supervised machine learning. In a more recent study,<sup>15</sup> ChemDataExtractor shows 57% precision and 31% recall when used to extract bulk modulus data from text. On the other hand, the weaknesses of the generative LLM can also be mitigated by prompt engineering, in-context learning, and fine-tuning. It is not the goal of this study to seek the best possible performance for each

of the methods. We are trying to compare these two in a more realistic setting that does not require too much effort for real-world applications. The rule-based method used by ChemDataExtractor naturally requires defining grammar rules before information extraction is possible. And the quality of the extraction depends heavily on the precision and completeness of the rules. The generative LLM seems to demonstrate very good performance without much training and fine-tuning. This finding has implications for using the generative LLM for domain-specific information extraction tasks.

## ASSOCIATED CONTENT

### Data Availability Statement

The data used in this study is available at [https://www.dropbox.com/scl/fi/be43gcyjcuvirhbmuu5d2/results\\_chatGPT\\_2024.csv.csv?rlkey=3nsdcmhmsivrsvppr853zpyb5&st=k066jlxc&dl=0](https://www.dropbox.com/scl/fi/be43gcyjcuvirhbmuu5d2/results_chatGPT_2024.csv.csv?rlkey=3nsdcmhmsivrsvppr853zpyb5&st=k066jlxc&dl=0). The source code of the rule-based information extraction method is available at <https://github.com/kun-ou-projects/SolarCell>

## AUTHOR INFORMATION

### Corresponding Author

Kun Lu – School of Library and Information Studies, The University of Oklahoma, Norman, Oklahoma 73019, United States;  [orcid.org/0000-0001-5614-7042](https://orcid.org/0000-0001-5614-7042); Email: [kunlu@ou.edu](mailto:kunlu@ou.edu)

### Authors

Xin Wang – School of Library and Information Studies, The University of Oklahoma, Norman, Oklahoma 73019, United States;  [orcid.org/0009-0003-5454-859X](https://orcid.org/0009-0003-5454-859X)

Liangliang Huang – School of Sustainable Chemical, Biological and Materials Engineering, The University of Oklahoma, Norman, Oklahoma 73019, United States;  [orcid.org/0000-0003-2358-9375](https://orcid.org/0000-0003-2358-9375)

Shuozhi Xu – School of Aerospace and Mechanical Engineering, The University of Oklahoma, Norman, Oklahoma 73019, United States

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jcim.4c00882>

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

We acknowledge the support of the Vice President for Research and Partnerships of the University of Oklahoma (OU) and the Data Institute for Societal Challenges. S.X. was supported by a grant from the Research Council of the OU Norman Campus. Some of the computing for this project was performed at the OU Supercomputing Center for Education & Research (OSCER) at OU.

## REFERENCES

- (1) Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805. arXiv.org e-Print archive, 2018 <https://arXiv.org/abs/1810.04805>.
- (2) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language Understanding by Generative Pre-Training, 2024. <https://www.mikecaptain.com/resources/pdf/GPT-1.pdf>. Retrieved on April 20th.
- (3) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Polosukhin, I. In *Attention is All You Need*, Advances in Neural Information Processing Systems; NeurIPS, 2017.
- (4) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9* (8), 1735–1780.
- (5) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36* (4), 1234–1240.
- (6) Gupta, T.; Zaki, M.; Krishnan, N. A.; Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput. Mater.* **2022**, *8* (1), No. 102.
- (7) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Jain, A.; et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571* (7763), 95–98.
- (8) Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G. Opportunities and challenges of text mining in materials research. *iScience* **2021**, *24* (3), No. 102155, DOI: [10.1016/j.isci.2021.102155](https://doi.org/10.1016/j.isci.2021.102155).
- (9) Weston, L.; Tshitoyan, V.; Dagdelen, J.; Kononova, O.; Trewartha, A.; Persson, K. A.; Ceder, G.; Jain, A. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.* **2019**, *59* (9), 3692–3702.
- (10) Wang, H.; Qin, K.; Zakari, R. Y.; Lu, G.; Yin, J. Deep neural network-based relation extraction: an overview. *Neural Comput. Appl.* **2022**, *34*, 4781–4801.
- (11) Swain, M. C.; Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **2016**, *56* (10), 1894–1904.
- (12) Ghosh, S.; Lu, K. Band gap information extraction from materials science literature – a pilot study. *Aslib J. Inf. Manage.* **2023**, *75* (3), 438–454.
- (13) Wei, X.; Cui, X.; Cheng, N.; Wang, X.; Zhang, X.; Huang, S.; Han, W. Zero-shot information extraction via chatting with chatgpt, arXiv:2302.10205. arXiv.org e-Print archive, 2023 <https://arXiv.org/abs/2302.10205>.
- (14) Hu, D.; Liu, B.; Zhu, X.; Lu, X.; Wu, N. Zero-shot information extraction from radiological reports using ChatGPT. *Int. J. Med. Inf.* **2024**, *183*, No. 105321.
- (15) Polak, M. P.; Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **2024**, *15* (1), No. 1569.
- (16) Park, Y. J.; Kaplan, D.; Ren, Z.; Hsu, C. W.; Li, C.; Xu, H.; Li, J.; Li, J. Can ChatGPT be used to generate scientific hypotheses? *J. Materomics* **2024**, *10* (3), 578–584.
- (17) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Blaiszik, B.; et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery* **2023**, *2* (5), 1233–1250.
- (18) Eltyeb, S.; Salim, N. Chemical named entities recognition: a review on approaches and applications. *J. Cheminf.* **2014**, *6*, No. 17.
- (19) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. Drugbank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2017**, *46* (D1), D1074–D1082, DOI: [10.1093/nar/gkx1037](https://doi.org/10.1093/nar/gkx1037).
- (20) Hettne, K. M.; Stierum, R. H.; Schuemie, M. J.; Hendriksen, P. J.; Schijvenaar, B. J.; van Mulligen, E. M.; Kleijnjans, J.; Kors, J. A. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **2009**, *25* (22), 2983–2991.
- (21) Rebholz-Schuhmann, D.; Kirsch, H.; Arregui, M.; Gaudan, S.; Riethoven, M.; Stoehr, P. EBIMed—text crunching to gather facts for proteins from Medline. *Bioinformatics* **2007**, *23* (2), e237–e244.
- (22) Eaborn, C. *Compendium of Chemical Terminology: IUPAC Recommendations*; Gold, V.; Loening, K. L.; McNaught, A. D.; Sehmi, P., Eds.; Blackwell: Oxford, 1988.

- (23) Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Valencia, A.; et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminf.* **2015**, *7*, No. S2.
- (24) Kim, J. D.; Ohta, T.; Tateisi, Y.; Tsujii, J. I. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* **2003**, *19* (suppl\_1), i180–i182.
- (25) Bada, M.; Eckert, M.; Palmer, M.; Hunter, L. In *An Overview of the CRAFT Concept Annotation Guidelines*, Proceedings of the Fourth Linguistic Annotation Workshop; ACL, 2010; pp 207–211.
- (26) Leaman, R.; Wei, C. H.; Lu, Z. tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminf.* **2015**, *7*, No. S3.
- (27) Shetty, P.; Rajan, A. C.; Kuenneth, C.; Gupta, S.; Panchumartti, L. P.; Holm, L.; Zhang, C.; Ramprasad, R. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Comput. Mater.* **2023**, *9* (1), No. 52.
- (28) Court, C. J.; Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **2018**, *5* (1), No. 180111.
- (29) Huang, S.; Cole, J. M. A database of battery materials auto-generated using ChemDataExtractor. *Sci. Data* **2020**, *7* (1), No. 260.
- (30) Dong, Q.; Cole, J. M. Auto-generated database of semiconductor band gaps using ChemDataExtractor. *Sci. Data* **2022**, *9* (1), No. 193.
- (31) Zhao, X.; Greenberg, J.; An, Y.; Hu, X. T. In *Fine-Tuning BERT Model for Materials Named Entity Recognition*, 2021 IEEE International Conference on Big Data (Big Data); IEEE, 2021; pp 3717–3720.
- (32) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Amodei, D. *Language Models are Few-Shot Learners*, Advances in Neural Information Processing Systems; NeurIPS, 2020; pp 1877–1901.
- (33) Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A. S.; Ceder, G.; Persson, K. A.; Jain, A. Structured information extraction from scientific text with large language models. *Nat. Commun.* **2024**, *15* (1), No. 1418.
- (34) Luu, R. K.; Buehler, M. J. BioinspiredLLM: Conversational Large Language Model for the Mechanics of Biological and Bio-Inspired Materials. *Adv. Sci.* **2024**, *11* (10), No. 2306724.
- (35) Buehler, M. J. MechGPT, a Language-Based Strategy for Mechanics and Materials Modeling That Connects Knowledge Across Scales, Disciplines, and Modalities. *Appl. Mech. Rev.* **2024**, *76* (2), No. 021001.
- (36) Buehler, E. L.; Buehler, M. J. X-LoRA: Mixture of Low-Rank Adapter Experts, a Flexible Framework for Large Language Models with Applications in Protein Mechanics and Molecular Design. *APL Mach. Learn.* **2024**, *2* (2), No. 026119, DOI: [10.1063/5.0203126](https://doi.org/10.1063/5.0203126).
- (37) Ghafarollahi, A.; Buehler, M. J. ProtAgents: Protein Discovery via Large Language Model Multi-Agent Collaborations Combining Physics and Machine Learning. *Digital Discovery* **2024**, *3*, 1389–1409, DOI: [10.1039/D4DD00013G](https://doi.org/10.1039/D4DD00013G).
- (38) Ni, B.; Buehler, M. J. MechAgents: Large Language Model Multi-Agent Collaborations Can Solve Mechanics Problems, Generate New Data, and Integrate Knowledge. *Extreme Mech. Lett.* **2024**, *67*, No. 102131.
- (39) Ghafarollahi, A.; Buehler, M. J. AtomAgents: Alloy Design and Discovery through Physics-Aware Multi-Modal Multi-Agent Artificial Intelligence, arXiv:2407.10022. arXiv.org e-Print archive, 2024. <https://arXiv.org/abs/2407.10022>.
- (40) Buehler, M. J. Generative Retrieval-Augmented Ontologic Graph and Multiagent Strategies for Interpretive Large Language Model-Based Materials Design. *ACS Eng. Au* **2024**, *4* (2), 241–277.
- (41) Buehler, M. J. Accelerating Scientific Discovery with Generative Knowledge Extraction, Graph-Based Representation, and Multimodal Intelligent Graph Reasoning, arXiv:2403.11996. arXiv.org e-Print archive, 2024. <https://arXiv.org/abs/2403.11996>.
- (42) Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Fung, P.; et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55* (12), 1–38.
- (43) Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Zhang, X. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges, arXiv:2402.01680. arXiv.org e-Print archive, 2024. <https://arXiv.org/abs/2402.01680>.
- (44) Xia, Y.; Dittler, D.; Jazdi, N.; Chen, H.; Weyrich, M. LLM Experiments with Simulation: Large Language Model Multi-Agent System for Process Simulation Parametrization in Digital Twins, arXiv:2405.18092. arXiv.org e-Print archive, 2024. <https://arXiv.org/abs/2405.18092>.