Patterns

Article

# Creation of a structured solar cell material dataset and performance prediction using large language models

## Highlights

- This study updates solar cell data and predicts material performance with large language models
- A new NLP task called structured information inference (SII) was introduced
- The authors proposed a one-step approach to summarize literature into structured data
- The broad accessibility of the approach for various domains in science

## Authors

Tong Xie, Yuwei Wan, Yufei Zhou, ...,
Clara Grazian, Wenjie Zhang,
Bram Hoex

## Correspondence

ctckit@cityu.edu.hk (C.K.),
b.hoex@unsw.edu.au (B.H.)

## In brief

This study explores the transformative power of big data in materials science, tackling the long-standing issue of data harnessability. The authors introduce a one-step approach that condenses unstructured data from publications into structured formats. By leveraging large language models, this method not only automates the enrichment of existing solar cell datasets but also offers predictive insights into material performance. These advancements underscore the role large language models play in the acquisition of scientific knowledge and the evolution of material science.

Xie et al., 2024, Patterns *5*, 100955
May 10, 2024 © 2024 The Author(s).
https://doi.org/10.1016/j.patter.2024.100955

CellPress

Patterns

CellPress
OPEN ACCESS

Article

# Creation of a structured solar cell material dataset and performance prediction using large language models

Tong Xie,[1,2] Yuwei Wan,[2,3] Yufei Zhou,[3] Wei Huang,[6] Yixuan Liu,[2] Qingyuan Linghu,[2,6] Shaozhou Wang,[1,2] Chunyu Kit,[3,*] Clara Grazian,[4,5] Wenjie Zhang,[6] and Bram Hoex[1,7,*]

[1]School of Photovoltaic and Renewable Energy Engineering, University of New South Wales, Kensington, NSW, Australia
[2]GreenDynamics Pty. Ltd, Kensington, NSW, Australia
[3]Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, China
[4]School of Mathematics and Statistics, University of Sydney, Camperdown, NSW, Australia
[5]DARE ARC Training Centre in Data Analytics for Resources and Environments, South Eveleigh, NSW, Australia
[6]School of Computer Science and Engineering, University of New South Wales, Kensington, NSW, Australia
[7]Lead contact
*Correspondence: ctckit@cityu.edu.hk (C.K.), b.hoex@unsw.edu.au (B.H.)
https://doi.org/10.1016/j.patter.2024.100955

**THE BIGGER PICTURE** Big data's importance in materials science is clear, yet its effective use is challenging due to the sheer volume and complexity of the data. Natural language processing (NLP) offers a solution by transforming unstructured text into structured formats, facilitating tasks such as extraction and summarization. In materials science, this means converting information from scientific papers into structured datasets, a process often slowed by the continuous influx of new data. To circumvent the inefficiencies of multi-step NLP workflows, there is a growing need for streamlined, one-step NLP methods. Employing fine-tuned large language models could be key, allowing for the rapid updating of datasets and providing valuable training data for further model development. This approach not only expedites research but also accelerates material prediction, leading to faster scientific breakthroughs.

## SUMMARY

Materials scientists usually collect experimental data to summarize experiences and predict improved materials. However, a crucial issue is how to proficiently utilize unstructured data to update existing structured data, particularly in applied disciplines. This study introduces a new natural language processing (NLP) task called structured information inference (SII) to address this problem. We propose an end-to-end approach to summarize and organize the multi-layered device-level information from the literature into structured data. After comparing different methods, we fine-tuned LLaMA with an F1 score of 87.14% to update an existing perovskite solar cell dataset with articles published since its release, allowing its direct use in subsequent data analysis. Using structured information, we developed regression tasks to predict the electrical performance of solar cells. Our results demonstrate comparable performance to traditional machine-learning methods without feature selection and highlight the potential of large language models for scientific knowledge acquisition and material development.

## INTRODUCTION

Data have long been the cornerstone of empirical science and serve as the basis for discoveries and our understanding of the world. In recent years, big data have become an indispensable resource for various industries, especially the technology sector. Materials science is no exception to this trend. It has revolutionized the research and development of advanced materials for a wide range of applications, including catalysts,[1] thermoelectrics,[2] and batteries.[3,4] These initiatives underscore the growing significance of data in materials research and pave the way for ground-breaking innovations in this field.

Despite the widespread recognition of the importance of data and ongoing initiatives to exploit their potential, experimental materials science continues to encounter difficulties in effectively leveraging the abundance of available data.[5] This problem is particularly evident in applied disciplines, where materials are frequently assessed primarily based on their device performance rather than through a thorough understanding of their inherent properties and behavior.[6] A crucial question in this context is how to utilize relevant information from the vast, unstructured scientific literature into a format suitable for materials scientists. This challenge not only makes it difficult to gain a comprehensive understanding of material candidates and their properties but also hinders the identification of future applications. It further adds to the bottleneck in the materials discovery pipeline, given the laborious and time-consuming nature of experimental synthesis.

Named entity recognition (NER), which aims to identify and classify named entities in unstructured text, is a commonly used natural language processing (NLP) technique for the automatic construction of domain-specific datasets. However, these NER-extracted datasets typically require a large amount of annotation, along with additional pre-processing or post-processing steps.[7–9] They also differ from findable, accessible, interoperable, reusable (FAIR)[10] datasets created by materials scientists in several ways, including the lack of entity correspondences and the single-label format, which limits content diversity. This limitation can make it challenging for materials scientists to query or utilize them effectively, resulting in reduced utility. In this study, we introduce a new NLP task called structured information inference (SII) to leverage pre-existing FAIR datasets in materials science. This task is at the discourse level and, in practice, covers mainstream tasks such as NER, entity normalization (EN), relation extraction (RE), and information inference (II). We accomplished this by fine-tuning LLaMA[11] on the Perovskite Database (www.perovskitedatabase.com), a manually summarized perovskite solar cell FAIR dataset published in February 2021 in a single step.[12] Our method achieved good performance on the SII task and is applicable for updating other FAIR datasets derived from scientific literature. We applied this approach to the highly dynamic field of perovskite solar cells and successfully extracted intricate relationships constructing an updated FAIR dataset for other perovskite solar cells published from March 2021 to March 2023. Additionally, we designed a regression task to predict the electrical performance of solar cells and to facilitate the design of materials or devices with targeted parameters.

Our approach provides evidence that large language model (LLM) can autonomously learn complex knowledge data frames and construct output according to predefined schemas from unstructured scientific text without requiring additional manual annotation. The produced dataset is formatted and normalized, enabling its direct utilization as input in subsequent data analysis, such as machine learning, without additional processing steps. This feature will enable materials scientists to update existing FAIR datasets or create new ones within their domains by developing their own models, fine-tuned on high-quality FAIR datasets and source papers. Even in cases where no FAIR dataset exists in a specific domain, our proposed approach allows for the rapid construction of a new dataset with minimal annotation, significantly faster than previous methods. The results of our regression task predicting device performance also demonstrate the potential of the fine-tuned LLM to handle various intricate materials informatics tasks, thereby reducing the cost of trial and error.

## RESULTS

### Issues of traditional annotation mechanism

In material science, significant effort has been devoted to extracting entities such as chemical terminologies, properties, and synthesis parameters from relevant scientific literature. The related NER methods used in materials science can be broadly categorized as rule-based[13] (relying on dictionaries or regex rules), recurrent neural network (RNN),[8,14,15] and transformer-based LLMs. In recent years, the emergence of LLMs such as bidirectional encoder representations from transformers (BERT)[16] have become the state-of-the-art for numerous NLP tasks, including NER. Both fine-tuned BERT[17] and domain-specific pre-trained BERT[3,18] have shown significant improvement in material-science NER tasks compared to RNN methods. Several datasets[19–21] also utilize the NER tool[13] to automatically generate tabular databases of material property data aggregated from textual entries. These NER-extracted datasets usually link material names with their co-occurring entities to analyze potential relations.

Extracting relationships between entities in materials science has been a challenge but this RE task received less attention than the NER task. Mysore et al.[22] built a dataset of 230 synthesis procedures with labeled graphs where nodes represent synthesis operations and their typed arguments, and labeled edges specify relations between the nodes. MatSciBERT[23] yields the best performance of RE on this dataset. Most existing research treats RE as a classification step following NER in an information-retrieval pipeline and usually focuses on intra-sentence binary relationships.[24–26] Nonetheless, real-world situations are considerably more intricate. Current approaches simplify the relations too much and result in significant information loss. It is worth noting that N-ary relations (involving N entities) have received increasing attention due to their additional challenges.[24] Recently, Dunn et al.[27] proposed a sequence-to-sequence LLM approach capable of addressing complex interrelations without the need to enumerate all possible N-ary relations.

The research mentioned above used a word-by-word traditional annotation mechanism (Figure 1), which does not align with the needs of material scientists. In label-based NER tasks, the output is extractive information that requires further processing, such as merging abbreviations and their full forms. Sometimes, certain implicit information cannot be integrated based on individually identified entities, leading to information loss. For example, in the FAIR dataset used in this paper, the value of attribute Perovskite_composition_long_form does not directly appear in the source text but needs to be inferred from perovskite information such as coefficients. On the one side, word-by-word mechanism typically demands significant effort from both NLP and materials science experts in several aspects: (1) creation of NER categories, (2) development of a labeling interface, (3) learning costs associated with NER/RE labeling rules, and (4) time costs of NER/RE labeling. Conversely, scientific information often goes beyond simple pairwise relationships between entities. For instance, a compound's properties are influenced by multiple factors, such as material name, phase structure, morphology, and synthesis methods. This complexity is exemplified in the distinction between plasma-enhanced chemical vapor deposition (PECVD) Al-doped $TiO_2$ film and

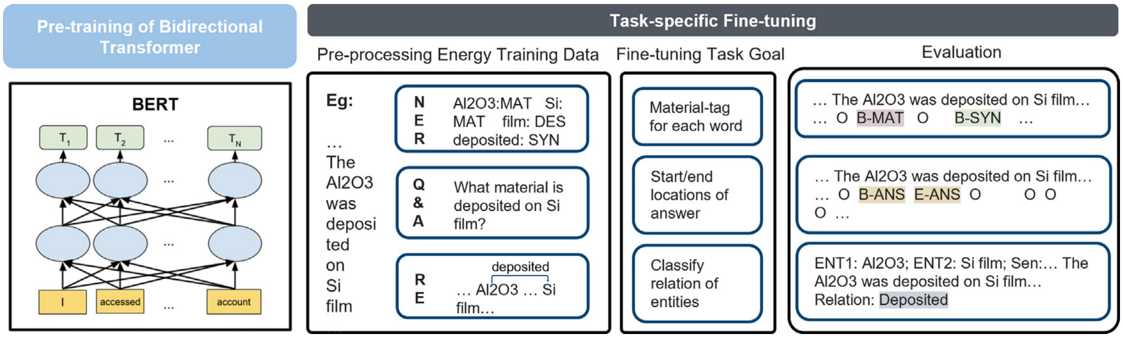

Pre-training of Bidirectional Transformer
BERT
T1
T2
...
TN
I
accessed
account
Task-specific Fine-tuning
Pre-processing Energy Training Data
Fine-tuning Task Goal
Evaluation
Eg:
… The Al2O3 was deposited on Si film
NER
Al2O3:MAT Si: MAT film: DES deposited: SYN
Q&A
What material is deposited on Si film?
RE
deposited
… Al2O3 … Si film…
Material-tag for each word
Start/end locations of answer
Classify relation of entities
… The Al2O3 was deposited on Si film…
… O B-MAT O B-SYN …
… The Al2O3 was deposited on Si film…
… O B-ANS E-ANS O O O O …
ENT1: Al2O3; ENT2: Si film; Sen:… The Al2O3 was deposited on Si film…
Relation: Deposited

**Figure 1. Traditional annotation mechanism**

Common NLP methods used for constructing domain-specific datasets include NER, and other tasks built upon NER (such as RE). Taking the example of using the BERT model for an NER task, the typical workflow, following the formulation of domain-specific labeling standards by domain experts, involves several steps: text pre-processing, tokenization, manual annotation, fine-tuning on the base model, and subsequent evaluation. To enhance performance, domain-specific text is sometimes used to continue pre-training on the base model to obtain a domain-specific base model. The entire process is labor intensive and time consuming.

atomic layer deposition (ALD) Al-doped $TiO_2$, which exhibit different properties. Additionally, materials knowledge is often hierarchical, with relations that may only be valid between one entity type and a compound entity comprising multiple entities and relationships. Theoretically, such relations can be modeled as N-ary, but comprehensively enumerating all possible variations is both impractical and unsuitable for traditional RE methods, as each relation type requires a sufficient number of training examples.

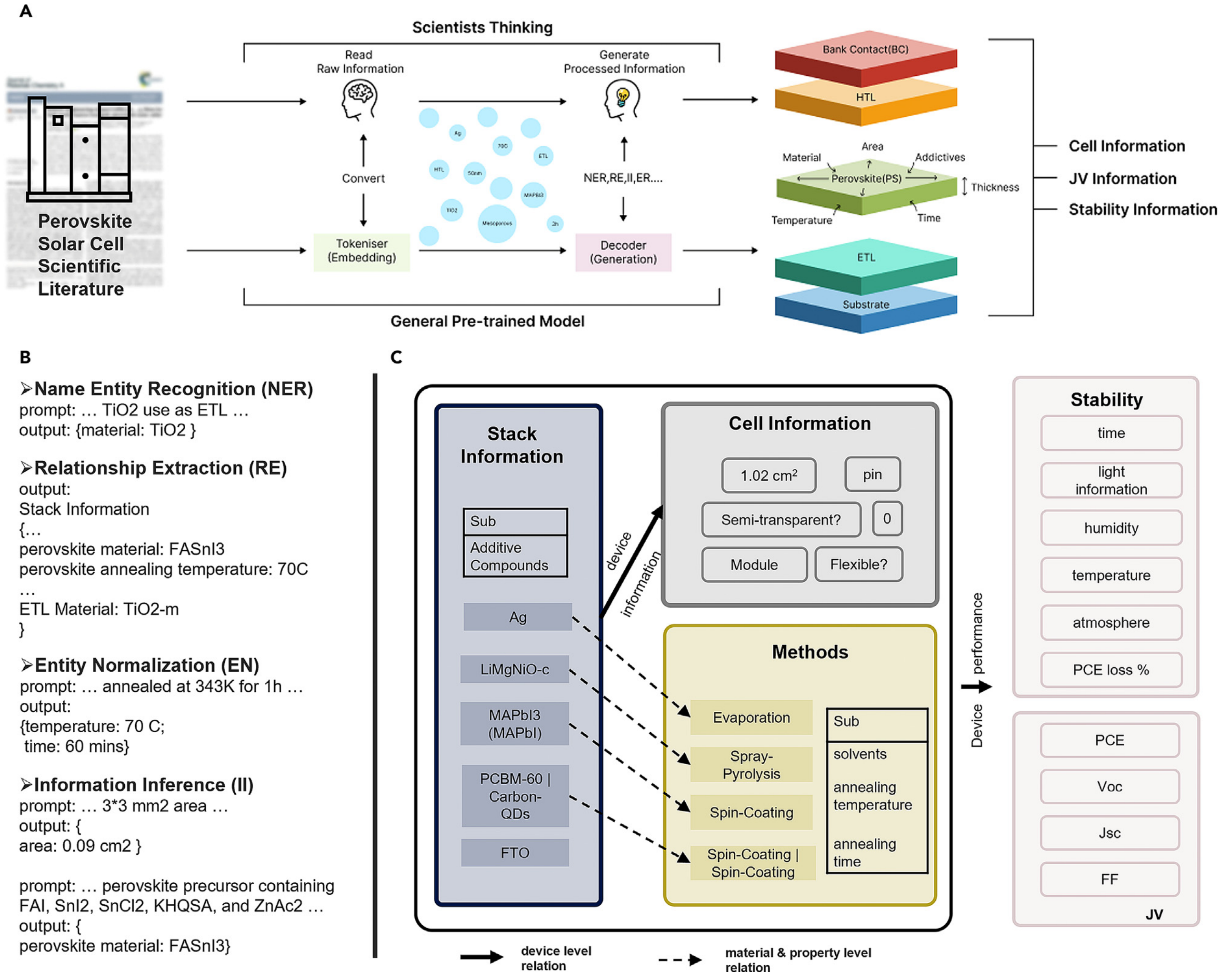
Due to the difficulty in annotation and simulation of material information, high-quality annotated data are limited in material science, prompting us to utilize existing review-paper databases. A review article represents a scholarly publication that amalgamates and evaluates prior research papers on a specific topic. Such articles investigate distinct research questions or theoretical or practical approaches, providing readers a comprehensive and current understanding of the research area. These articles contain natural, high-quality summaries and intricate relationships in domain-specific subjects, materials and properties, and device information. We endeavored to trace back the summarized information in the review paper, as provided by other scientists, to the original text through entity and relation annotation. However, these efforts did not yield a perfect match with the corresponding sections. According to statistics, exact match demonstrates a low matching rate of 44.7% on average. The unmatched parts require manual annotation, with an estimated annotation time of 20 s per material entry.[27] Thus, converting this dataset into traditional word-by-word NER and RE data annotation proved challenging.

## Opportunities and new NLP task: SII

As discussed in previous sections, existing research primarily concentrates on identifying entities and their relationships. However, the practical process of extracting information by materials scientists is considerably more complex, which can be explained in two aspects:

First, it refers to the complexity of the data themselves, as illustrated in Figures 2A and 2C, where device information is multi-layered, and each layer contains similar elements, making it prone to confusion (e.g., various deposition procedures). The expressions found in research articles are also intricate and varied. At the device level, not only do complex relationships between materials need to be considered but units may also differ across publications. For example, 0.3 $mm^2$ is equivalent to 0.003 $cm^2$. Entity definitions can be flexible, and, occasionally, their meanings depend on words in separate paragraphs. A paper might only mention Al-doped $TiO_2$, leaving scientists to infer whether it is ALD or PECVD grown. Furthermore, field-specific vocabulary may introduce ambiguity, complicating matters further; for instance, ''Al-doped $TiO_2$ film'' could be synonymous with ''$Al_xTi_yO$ film.'' Second, it pertains to the inherent complexity of the task itself. Transforming unstructured text into a FAIR dataset demands advanced NLP capabilities. Previously, this was achieved through multiple sequential steps, including NER, RE, normalization, and ultimately structuring the data. We examined various activities that scientists employ to summarize and infer information from materials science articles and discovered that they mainly need four types of ability (the first three are also existing NLP tasks), as demonstrated by the examples in Figure 2B:

- NER: the fundamental task involves identifying and classifying named entities within text, such as material names and associated properties such as temperature. In materials science, NER is crucial for cataloging and organizing information about various materials, which serves further analysis of co-occurring relationship of entities and high-level visualization.
- RE: RE involves discerning and uncovering connections and associations between individual entries or groups of entries within a text. In materials science, this task can be used to identify the relationships between materials, their properties, and applications, providing queries and valuable insights for researchers.
- EN: EN is the process of standardizing the expression format, units, abbreviations, and other variations in the information extracted from text. In materials science, EN ensures the inner consistency of data, making it easier to integrate information from different sources and facilitating meaningful cross-referencing.



➢Name Entity Recognition (NER)
prompt: … TiO2 use as ETL …
output: {material: TiO2 }

➢Relationship Extraction (RE)
output:
Stack Information
{…
perovskite material: FASnI3
perovskite annealing temperature: 70C
…
ETL Material: TiO2-m
}

➢Entity Normalization (EN)
prompt: … annealed at 343K for 1h …
output:
{temperature: 70 C;
time: 60 mins}

➢Information Inference (II)
prompt: … 3*3 mm2 area …
output: {
area: 0.09 cm2 }

prompt: … perovskite precursor containing FAI, SnI2, SnCl2, KHQSA, and ZnAc2 …
output: {
perovskite material: FASnI3}

**Figure 2. A new NLP task SII**

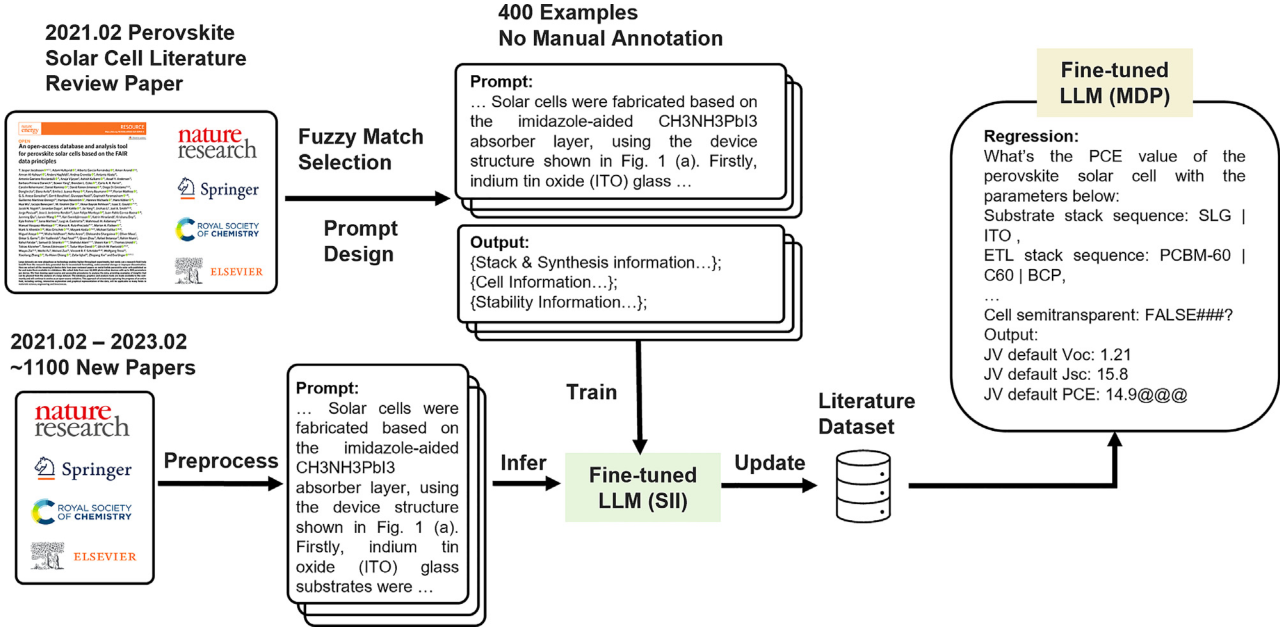
(A) An overview of SII through multi-task learning is presented, with the decoder responsible for comprehending tasks and generating the corresponding outputs.

(B) Examples of abilities in creating FAIR datasets, which include named entity recognition (NER), relationship extraction (RE), entity normalization (EN), and information inference (II).

(C) A diagram of multi-layered device information of perovskite solar cells in this study.

- II: in contrast to conventional information extraction, II further involves advanced capabilities such as computational analysis and component inference. In materials science, II is of utmost importance for materials scientists when creating FAIR datasets. This is because they need to establish a clear schema from the outset and align information that may not have appeared explicitly in the data to this schema during the manual curation process.

Extracting information from scientific texts can be more challenging than the processes applied to general texts. Moreover, a piece of material knowledge might be inferred through multiple NLP tasks with multiple entities. For instance, the Al-doped $TiO_2$ compact layer could be inferred as ALD c-$Al_xTi_yO$ layer in a review paper or FAIR dataset when the deposition method is mentioned in another paragraph. These complexities make annotating related training datasets particularly demanding, especially at the documents level, as they represent an accumulation of materials science knowledge spanning centuries. To simulate the process of scientists extracting information from domain-specific texts, we propose a new NLP task designed for the scientific field called SII. This task aims to jointly perform II (or extraction) and RE. The relationship could be hierarchical or listed as multiple items without enumerating all possible n-tuple relationships. Initially, we attempted to use BERT-based approaches; however, the need to determine specific tasks for each piece of information complicated the problem, rendering the original BERT or domain-specific BERT unsuitable for fine-tuning. However, the advent of Generative Pre-trained Transformers 3 (GPT-3)[28] and its related applications offers new opportunities. As depicted in Figure 2A, GPT-3 and other generative language models employ a decoder structure, well suited for sequence-to-sequence tasks (i.e., input text generates output text) and aligns with the



Prompt:
… Solar cells were fabricated based on the imidazole-aided CH3NH3PbI3 absorber layer, using the device structure shown in Fig. 1 (a). Firstly, indium tin oxide (ITO) glass …

Output:
{Stack & Synthesis information…};
{Cell Information…};
{Stability Information…};

Regression:
What's the PCE value of the perovskite solar cell with the parameters below:
Substrate stack sequence: SLG | ITO ,
ETL stack sequence: PCBM-60 | C60 | BCP,
…
Cell semitransparent: FALSE###?
Output:
JV default Voc: 1.21
JV default Jsc: 15.8
JV default PCE: 14.9@@@

Prompt:
… Solar cells were fabricated based on the imidazole-aided CH3NH3PbI3 absorber layer, using the device structure shown in Fig. 1 (a). Firstly, indium tin oxide (ITO) glass substrates were …

**Figure 3. Workflow of updating an existing FAIR dataset**

Our workflow begins by extracting information-rich attributes from records in the FAIR dataset and preparing corresponding source text using DOIs. Then, we design prompts for GPT-3.5 prompting and LLaMA fine-tuning (where the input is the source text, and the output is a schema containing multiple attributes in a fixed format). After obtaining the fine-tuned model, we search for the most recent articles to input into the model, generating corresponding schemas. These schemas are then transformed into CSV (comma-separated values) format records to update the FAIR dataset. Additionally, we explored fine-tuning LLaMA model for MDP tasks (the input is schema, and the output is prediction of performance metrics).

generating logic of material scientists in collecting data from the literature. Consequently, we propose fine-tuning LLaMA, an open-source alternative, to infer key information from original papers directly. This approach not only saves significant time and cost but also leads to more accurate and comprehensive information summarization. The LLM can capture high-dimensional information and relationships within a paper that traditional word-by-word labeling mechanisms may overlook.

## Using LLMs for SII task

As shown in Figure 3, the practical process of this study involves data preparation, prompt design, training the model to complete SII task, making inferences on new data, and training the model for material and device prediction (MDP) tasks. The FAIR dataset we used has data from over 42,400 photovoltaic devices with up to 100 attributes per device.[12] We associated these data with more than 15,000 corresponding articles using our paper downloading tool SciCrawler (https://github.com/MasterAI-EAM/SciCrawler). The attributes cover stack information, system-level data, and performance metrics. Our SII and MDP study focuses on those with abundant information and widespread usage (about 35 attributes). We ranked records from the FAIR dataset by using a fuzzy match mechanism and selected the top 400 records as our dataset (see section "experimental procedures"). For each record, there is a source text and corresponding schema containing multiple attributes. The information of each schema was organized into two levels.

- Material- and property-level information: stack information (set A) and methods information (set B) for each layer, encompassing substrate, electron transport layer (ETL), perovskite, hole transport layer (HTL), and back contact.
- Device-level information: stability (set C) and electrical (J-V) performance data (set D).

Each attribute in the schema has an attribute name and corresponding value that can be inferred from the source text. These schemas serve as the structured information our model will learn to extract. To aid the model in understanding the task, we implemented both prompting (directly use model to infer without fine-tuning) and fine-tuning (supervised training on base model) on GPT-3/3.5[29] and the LLaMA model. In the following sections, we primarily compare the results of prompting GPT-3.5 and fine-tuned LLaMA. The results of prompting LLaMA were too poor to parse or calculate metrics. With the fine-tuned LLaMA model, we updated the FAIR dataset using perovskite solar cell papers published from March 2021 to March 2023. Additionally, we explored fine-tuning the LLaMA model for MDP tasks.

In the following section, we report the results of SII task in three parts: NER results, RE results, as well as the II and EN results. The results show that our fine-tuned model outperformed prompting GPT-3.5 in both NER and RE tasks, and it also obtained high accuracy in unique EN and II tasks. Fine-tuning can significantly reduce the gap between open-source models and proprietary models, enabling open-source models to achieve commendable task performance. The results of fine-tuned GPT-3 for the SII task are available in Tables S2 and S3.

**Table 1. Results of NER in SII task**

| Model | Set | Precision | Recall | F1 score | Manual |
|---|---|---|---|---|---|
| GPT-3.5 | A | 19.9 | 44.9 | 27.6 | 71.0 |
| | B | 13.1 | 27.1 | 17.7 | 72.1 |
| | C | 18.9 | 57.3 | 28.4 | 82.5 |
| | D | 27.0 | 43.6 | 33.3 | 59.3 |
| | total | 22.6 | 43.0 | 28.7 | 72.1 |
| Fine-tuned LLaMA | A | 83.54 | 81.23 | 82.1 | – |
| | B | 90.2 | 90.2 | 90.2 | – |
| | C | 82.8 | 78.31 | 80.14 | – |
| | D | 97.96 | 97.96 | 97.96 | – |
| | total | 88.34 | 86.11 | 87.14 | – |

## Results of SII task

Table 1 shows the results for the schema attributes matching computed using metrics described in "evaluation" in the "experimental procedures" section, along with human evaluation. We provided manual metrics only for prompting GPT-3.5. Despite its poor performance in SII automatic evaluation metrics, it could extract relevant content based solely on prompts. The lack of standardized formatting made manual evaluation beneficial to offer a more comprehensive assessment of GPT-3.5’s performance. On the other hand, the fine-tuned LLaMA model has already learned the implicit data transformations and formatting requirements present in the training data. Even if manual evaluation were used, it would not yield significantly different results, and there would be no substantial improvement compared to the current automatic metrics. If a more stringent automated verification method is used (the first three metrics), GPT-3.5 performs poorly on all sets but shows significant improvement in human evaluation. This disparity indicates that, although the powerful GPT-3.5 can extract some correct information based on the provided schema prompts in the absence of fine-tuning, it fails to conform to the requirements of the FAIR dataset format in terms of expression. Particularly, the identification of set D remains extremely challenging for the GPT-3.5 (F1-score = 59.3). We speculate that this is due to the deterioration of identification performance when the schema prompts themselves are not explicitly mentioned in the text (set D prompts are generally not explicitly stated in the text). In contrast, the fine-tuned LLaMA achieves F1 score exceeding 80 on all sets, with set D reaching nearly 98%. This indicates that the fine-tuned LLaMA accurately extracts schema-relevant information and adheres to the requirements of the FAIR dataset format and expression.

We also look into details: the attributes generated by GPT-3.5 are longer than the target answers, especially in procedure-related attributes (even though we have attempted to impose length restrictions during prompt design). According to statistics in manual evaluation, about 15% of correct predictions produced by GPT-3.5 contain significant unrelated information, while this ratio is only 4% for our fine-tuned model. Consequently, the former had a lower precision and higher recall considering the averaged length of the output. However, its recall is still significantly lower than that of a LLaMA fine-tuned on material scientific knowledge datasets, which implies that GPT-3.5 cannot accurately summarize hidden information in input paragraphs directly. In contrast, a fine-tuned model not only finds the corresponding parts accurately but also learns to summarize, normalize, or even deduce. Fine-tuned LLaMA achieved about 94% of the performance of fine-tuned GPT-3.

**Table 2. Results of RE in SII task**

| Model | Relation | Precision | Recall | F1 score | Manual |
|---|---|---|---|---|---|
| GPT-3.5 | A-B | 5.02 | 11.96 | 6.67 | 43.4 |
| | A-C | 7.23 | 29.51 | 10.3 | 66.5 |
| | ABC-D | 2.76 | 10.73 | 3.95 | 49.38 |
| Fine-tuned LLaMA | A-B | 78.86 | 75.31 | 76.81 | – |
| | A-C | 72.63 | 67.22 | 69.42 | – |
| | ABC-D | 71.97 | 65.84 | 68.23 | – |

Table 2 reports the RE scores, which evaluate the consistency of inner attribute sets of output schema. Since a proper relation must be based on the correct extracted entities, the performance of the RE task is influenced by the performance of the NER task. Thus, the RE scores here can be seen as a reflection of the NER-RE task, not just RE.

Overall, the fine-tuned model significantly outperforms GPT-3.5 in all three types of RE (about 15 points in average). Specifically, the performance degradation of the fine-tuned model (about 10%) from the NER to NER-RE task is much smaller than that of GPT-3.5 (about 30%). The averaged difference between precision and recall is also smaller for the fine-tuned model. In comparison, the fine-tuned model has a more balanced performance among the three types of relations.

We further analyze the results of SII in detail. We introduce the concept of EN and II to measure the model performance. EN in our SII task reflects in the normalization of different units and forms of terms, while II reflects in the inference of implicit information, which does not appear directly in the source text. Table 3 shows the support number and accuracy of II, entity normalization for units (EN-U) and entity normalization for terms (EN-T) on our fine-tuned model, respectively. To help to understand, we also give their example prompts and outputs. We did not display the results of the GPT-3.5 model for these tasks because the accuracy of each task is 0% or close to 0%. In comparison, the high accuracy achieved by fine-tuning the model indicates that our model greatly enhances the ability of language modeling to comprehend data formats and fill in missing information, simulating the process by which scientists extract and process data from research papers.

## MDP with LLM

Upon further investigation of the SII results, we identify a phenomenon known as hallucination within the LLM outputs in set D. In this context, hallucination refers to instances where no stability test is mentioned in the input but the fine-tuned SII model is employed. We devise a regression task for predicting device performance to quantify the model’s performance. However, as only 11% of the training device data have undergone stability tests, the sample size is insufficient to generate adequate training and test sets. Consequently, we opt for electrical performance data, as all data points possess associated values, including open-circuit voltage ($V_{oc}$), short-circuit current ($J_{sc}$), and power conversion efficiency (PCE). Notably, the model

Table 3. Results of II and EN in SII task

| Task | Example prompt | Example completion | Support | Accuracy |
|---|---|---|---|---|
| II | … perovskite precursor containing FAI, SnI2 … | perovskite material: FASnI3 | 70 | 75.71 |
| EN-U | … annealed at 343 K for 1 h … | temperature: 70°C; time: 60 min | 41 | 70.73 |
| EN-T | … mesoporous $TiO_2$ … | material: $TiO_2$-m | 154 | 75.32 |

only predicts data points of JV_light_spectra under AM1.5 and JV_light_intensity equal to 1,000 W/$m^2$.

The fine-tuning of the regression model used the same method as the SII task. We employed plain-text schema with corresponding values as input to predict values of three properties (the values are continuous numbers), which are $V_{oc}$, $J_{sc}$, and PCE, for perovskite solar cells employing specific synthesis methods. We use mean absolute error (MAE) and root-mean-square error (RMSE) to measure the difference between the predicted value and the true value:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \qquad \text{(Equation 1)}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n}} \qquad \text{(Equation 2)}$$

Although LLMs cannot predict real numbers in highly precise regression tasks, they can still produce predictions of acceptable accuracy by employing rounded values during training. A precision of two decimal points is deemed sufficient for electrical performance data. The prompt question and a detailed example schema are available in the supplemental information. Figure 4 visualizes the experimental results and prediction values of $V_{oc}$, $J_{sc}$, and PCE for comparison. Table 4 demonstrates the MAE metrics of the fine-tuned model in MDP for the regression task with RMSE metrics for $V_{oc}$, $J_{sc}$, and PCE being 0.12, 4.48, 4.71, respectively.

The fine-tuned GPT-3 (Table S4) has much better performance than fine-tuned LLaMA on the MDP task, and the level of randomness in predicting PCE has been significantly reduced. It can be observed that, in fine-tuned GPT-3 (Figure S1) and fine-tuned LLaMA (Figure 4), both $J_{sc}$ and $V_{oc}$ look slightly more promising than PCE. We believe the relatively poor performance of PCE may be due to the fact that PCE values are calculated from $J_{sc}$ and $V_{oc}$ using the formula

$$\mathrm{PCE} = \frac{FF V\mathrm{oc}J\mathrm{sc}}{P_{in}} \qquad \text{(Equation 3)}$$

where FF is fill factor and Pin is the input power. Therefore, the PCE values accumulate errors from $J_{sc}$ and $V_{oc}$, and these errors present in the training dataset are also propagated to the predictions, making the prediction of PCE more challenging. On one hand, the composition of the devices is quite complex, and, on the other hand, there is inherent experimental data error obtained from the papers, making the prediction of device performance itself a challenging task. Taking the experimental dataset HOPV15[30] as an example, performance of machine-learning models trained on device information in predicting PCE is around 3.6 ± 0.8 (MAE).[31] Even for devices prepared from the same batch of experiments, there is a variation in performance of 2%–5%. Therefore, in this study, we are merely exploring the capabilities of LLMs and find that they can indeed learn some correlation between certain device parameters and their performance during training.

We also depict the effect of the training dataset size in Figure 5. An epoch refers to one complete pass through the entire training dataset during the training. As we have set the number of epochs to be equal to three, the examples beyond 360 are repeated. It can be observed that there is a sharp reduction of training loss during the first 180 examples, but, after one epoch, the decrease is relatively slow and marginal.

## DISCUSSION

Based on the predicted schemas, we summarize the issues of direct use of GPT-3.5: (1)the predicted schema may occasionally miss one or two attributes (considered as incorrect answers during evaluation). (2) The suggested schema can alter the expression of the generated answer. For example, "Backcontact additives compounds" becomes "Backcontact additives/compounds." (3) There can be multiple expressions for the same answer, such as "not mentioned," "N/A," "none mentioned," and "not specified," which causes difficulties in parsing and predicting a unified format. (4) The generated answer's length is not fixed and can sometimes be very long, even if the prompt design limits the length (the limit not always works). (5) Correct answers may undergo unnecessary changes in expressions, such as converting "60 min" to "1 h." (6) Repetitive answers with similar content. (7) Sometimes, hallucinations occur (details in section "MDP with LLM").

In comparison, fine-tuning exhibits significant advantages: (1) top experts in the field design the framework and it aligns more closely with the domain-specific experimental thinking. (2) It saves the cost, time, and effort of annotation. (3) Professionals in the relevant field can directly use the results of the same framework without any additional learning costs.

Our study demonstrates that an LLM, even without prior training in materials science, can predict device performance data that may not be explicitly stated in the literature. Although the generated hallucinated information is not completely accurate, it remains valuable for researchers using the amassed scientific knowledge. In contrast to the recent advancements in perovskite solar cell prediction by Liu et al.,[32] who manually collected 814 data points from 2,735 publications and built machine-learning models for J-V performance prediction using only 13 features, LLMs are capable of automatically generating higher-dimensional datasets. This ability allows LLMs to guide subsequent device design at the material level, accounting for

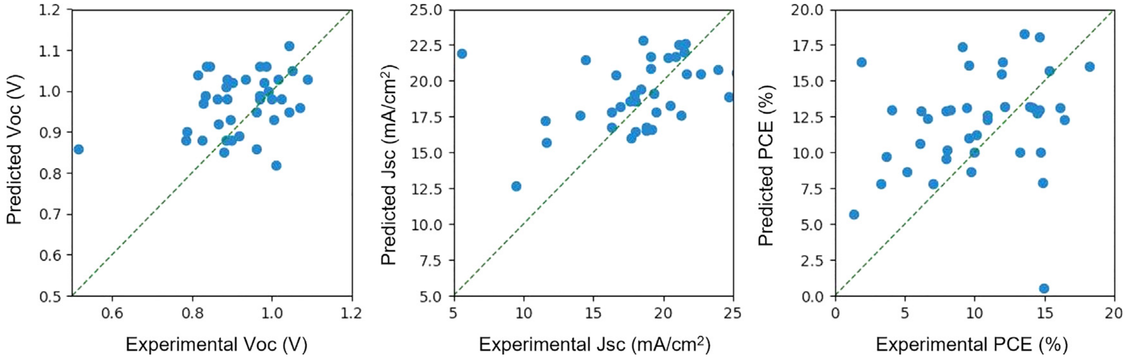


**Figure 4. Performance of fine-tuned LLaMA on predicting $V_{oc}$, $J_{sc}$, and PCE**

We compare the experimental values and fine-tuned LLaMA prediction values. The horizontal x axis displays experimental results, while the vertical y axis displays predicted values. Compared with $V_{oc}$ and $J_{sc}$, the prediction of PCE exhibits higher randomness.

parameters such as annealing time, annealing temperature, material thickness, and area. The models offer greater flexibility in feature selection, and feature values are not confined to numerical data, providing readily obtainable information for scientists. Jablonka et al.[33] also demonstrated that GPT-3 performs comparably with or outperforms traditional techniques when confronted with limited data, particularly for organic compounds with unique line encodings such as SMILES[34] or SELFIES.[35] Similarly, we devised a schema for predicting the organic photovoltaic devices (OPVs) PCE (density functional theory [DFT], calculated) based on the Harvard Photovoltaic (HOPV15: https://doi.org/10.1038/sdata.2016.86)[30] Dataset. Compared to the Bayesian regularized artificial neural network with Laplacian prior (BRANNLP) method employed by Meftahi et al.,[36] fine-tuned GPT achieves comparable performance with a simple schema design (see supplemental information).

LLMs, such as the LLaMA used in the study, have demonstrated proficiency for identifying structural and property-related similarities between novel materials and those previously investigated, akin to the expertise of seasoned materials scientists. This ability to identify similarities enables the investigation of variations in these novel materials, consequently opening up opportunities for innovative applications. Moreover, LLMs exhibit the potential to design cutting-edge devices by harnessing detailed material information. While these general-purpose LLMs were not initially tailored for scientific fields, their performance in this domain suggests a promising future in scientific applications. By augmenting LLMs with further training in relevant scientific literature, they may potentially be empowered to guide experimental design and significantly expand their scope of applications in materials science and beyond.

In this study, we introduce a new NLP task called SII, which aims to obtain hierarchical, domain-specific material and device information within a structured FAIR format from unstructured scientific texts. After analyzing traditional annotation mechanism and characteristics of LLM, we proposed to solve this task by fine-tuning one of the LLMs, LLaMA. Remarkably, this approach does not necessitate manual annotation, instead relying on review papers or FAIR datasets for training purposes. By employing this method, LLaMA can effectively predict material properties and device performance, as well as generate innovative materials or devices tailored to meet specialized requirements. On the most important NER metrics, fine-tuned LLaMA achieves 94% of the performance of fine-tuned GPT-3, while prompting LLaMA get almost no usable results. This indicates that fine-tuning can significantly reduce the gap between open-source models and proprietary models, enabling open-source models to achieve commendable task performance. We recognize that open-source models may not perform as well as mature commercial models due to factors such as parameter scale and training strategies. However, we chose to utilize popular open-source models to ensure that the trained models can be openly shared, used, and maintained by the academic community. This helps drive the development of this task or paradigm and garners more attention.

**Table 4. MAE of regression tasks on performance prediction of perovskite solar cell**

| Sample | 10 | 90 | 180 | 360 |
|---|---|---|---|---|
| $V_{oc}$ | – | 0.203 | 0.123 | 0.098 |
| $J_{sc}$ | 18.10 | 6.91 | 4.52 | 3.42 |
| PCE | 10.22 | 6.15 | 4.73 | 3.99 |

Demonstrating exceptional flexibility, the approach readily adapts to various challenges within scientific fields and exhibits outstanding performance in both SII and MDP tasks, particularly for perovskite and organic photovoltaic devices. Existing LLMs can leverage this method to extract structured relational datasets, thereby guiding material development. We will continue exploring multimodal models to further utilize table and figure data, as demonstrated in our convolutional neural network (CNN) framework.[37] This end-to-end approach ultimately seeks to empower scientists with the ability to swiftly generate material knowledge and design novel materials or chemicals for research purposes. To showcase our method, an online demonstration can be accessed at http://www.masterai.com.au.

## Limitations of the study

In this study, we observed that failures predominantly occur when a sample surpasses LLaMA's prompt-completion token limit, which was set at 2,048 during the investigation. This limitation implies that paragraphs characterized by considerable

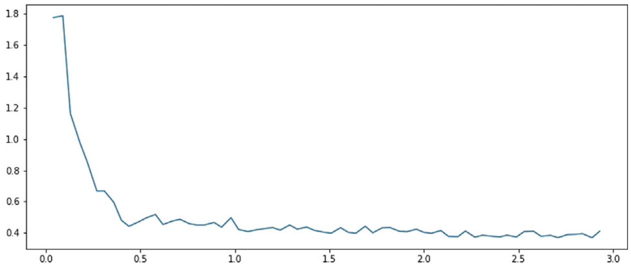


**Figure 5. Relationship between epoch and training loss on regression task**

We visualize the decreasing curve of training loss during fine-tuning LLaMA on MDP task. The horizontal axis represents the epoch, while the vertical axis represents the training loss. The curve indicates that 180 examples can enable the model to grasp the underlying patterns of predicting device performance based on device information.

length or high information density are fundamentally incompatible with the current approach. A significant proportion of unparsable completions can be attributed to instances where the passage and partial completion extend beyond the imposed token limit, consequently leading to premature truncation and hindering the generation of a comprehensive output.

## EXPERIMENTAL PROCEDURES

### Resource availability

#### *Lead contact*

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Bram Hoex (b.hoex@unsw.edu.au).

#### *Materials availability*

This study did not generate physical materials.

#### *Data and code availability*

The fine-tuned models and the code used for fine-tuning and inference are publicly available at GitHub (https://github.com/masterAI-EAM/MATGPT) and Original data have been deposited to figshare: https://doi.org/10.6084/m9.figshare.24972699.v1.[38] We also provide related datasets and the DOIs of the articles we used in the SII task. To showcase our method, an online demonstration can be accessed at http://www.masterai.com.au.

### Dataset preparation (fuzzy match mechanism)

Each record we extracted from the FAIR dataset is formed by a doi of source paper and one or more schema (most source papers have one schema and each schema stores information of one device) with multiple attributes. Each attribute in the schema is formed by a pair of name and value; for example, in the attribute “Substrate_stack_sequence: SLG — ITO,” “Substrate_stack_sequence” is the name and “SLG — ITO” is the value. For the dataset preparation, we first downloaded and processed the full text of source papers. To meet the 2,048 token limit requirement (about 1,500 words) of LLaMA, we only extracted the most informative sections in the papers. The condition is that the header of the section should contain keywords “experimental,” “materials,” “methods,” or “experiment.” We joined these extracted sections by space and call it source text.

We proposed a fuzzy matching mechanism to figure out how well a schema matched with its source text. The match rate of a schema and its source text is the ratio of matched value to all values in the schema. The schemas and source text were converted to lowercase, and the value of each attribute in a given schema is split into a list of pieces by delimiters (e.g., |, ; ,:). Each attribute had a given matching rule. By default, as long as one piece of the split values appears in the source text, we consider the entire value and source text to be a match. For the name “ETL stack sequence,” if the whole string or the substring before “-” in a given value appears in the source text, it is a matched value. For the names “perovskite composition long form” and “perovskite composition short form,” if a given value is a subset of a single word within the source text, it is a matched value. For the names with the value “unknown,” we always count them as a match. For source text with more than one schema, we ranked the schemas by the match rate from high to low and only select the top one. We then ranked the source text by the match rate of only schema and select the top 400 source text and their schema as the training samples. The match rates of schemas in these selected samples ranged from 100% to about 85%.

### Schema design

In fine-tuning, we transformed the original tabular data into a plain text schema to facilitate the model’s understanding. Each schema is presented as a dictionary, where the keys represent the attribute names and the values represent the corresponding attributes. For each relevant paragraph, we aimed to enable the model to learn how to automatically and accurately summarize a corresponding schema. We conducted prompt design to obtain results that are as close to the desired format as possible (we opted not to utilize the original LLaMA for comparison, as, even with well-designed prompts, the model produced results that were exceedingly difficult to decipher, often containing numerous repetitions and nonsensical outputs). After multiple attempts, we designed a prompt using the original paragraph with the prefix “Read the following paragraphs and extract the information below:” and list of attribute names attached. For comparison, we use similar prompts on GPT-3.5. We removed underscores in attribute names and added some requirements to limit the length or content of the attribute names. For example, we added “(only name, not details)” after the attribute name “HTL deposition procedure.” For attribute names that require boolean answers, we changed the attribute names into general questions. For example, we changed “Module” to “Any Module test?.”

### Fine-tuning details

We choose llama-7b-hf (7B parameters)[11] as our base model since it is one of the most capable open-source LLMs available for fine-tuning, considering our limited computing resources. And for GPT-3 fine-tuning, we used Davinci via OpenAI API. Each data sample has an instruction, an input, and an output. Specifically, the instruction is a short sentence describing the task. The input is the text extracted from scientific papers with several paragraphs having schema information. The output act as the answer to those schemas, including 31 name-value pairs in the form of “schema name: answer.” For parsing convenience, our dataset is in .json format, where \n is inserted among each schema, $< s >$ at the beginning of the output, and $< /s >$ at the end of the output. Then the dataset is split into a training set and a test set containing 360 and 40 samples separately. The model is trained for three epochs at a batch size of 1.

### Evaluation

We evaluated the performance of the fine-tuned model and GPT-3.5 on SII task using four decomposed sub-tasks: NER, RE, EN, and II. The metrics of the NER task evaluated how likely an output schema (prediction) was matched with the target schema (answer). Instead of BiLingual Evaluation Understudy (BLEU) or Recall-Oriented Understudy for Gisting Evaluation (ROGUE) scores (common metrics for natural language generation tasks), we opted for custom word tokenization due to the special delimiters in some values of the FAIR dataset. Each value of attribute in the output schema can be seen as an entity $E^p$ and the corresponding value in the target schema can be seen as an entity $E^a$. We design a word-basis measurement by separating an entity $E$ into a set of words $S = \{w_1, w_2, w_3, \ldots, w_k\}$ and comparing the difference between $S^p$ and $S^a$. The separators include ;, |, :, and $\gg$. After separating both entities, the number of matching words in both sets is counted as true positives ($S^p \cap S^a$) and the set difference is counted as false positives ($S^p \backslash S^a$) or false negatives ($S^a \backslash S^p$). For example, an attribute in the output schema is “70.0 $\gg$ 120.0” and the corresponding answer was “70.0 $\gg$ Unknown,” and we recorded one true positive “70.0,” one false positive “120.0,” and one false negative “Unknown.” With true positives (*tp*), false positives (*fp*) and false negatives (*fn*) identified, metrics of each pair of entities were calculated as:

$$\text{precision} = \frac{tp}{tp+fp} \qquad \text{(Equation 4)}$$

$$\text{recall} = \frac{tp}{tp+fn} \qquad \text{(Equation 5)}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision}+\text{recall}} \quad \text{(Equation 6)}$$

The metrics of the RE task evaluate how likely the output schema caught the inner relationship between related attribute sets. According to the nature and internal relation of attributes in the schema described in "using LLMs for SII task," we construct three types of relations: A-B, A-C, and ABC-D. The relationships are also scored by a word-basis measurement similar to the one NER uses, using a number of correct collocations. Each collocation is an n-tuple relating words $w_n^m$ of each involved entities $E_n$ in relation $r$. For each type of relation, we can summarize collocations in the predicted schema into a predicted relation set ($R^p$) and those in the answer schema into an answer relation set ($R^a$). The number of matching collocations in both relation sets is counted as true positives ($R^p \cap R^a$) and the collocation difference is counted as false positives ($R^p \backslash R^a$) or false negatives ($R^a \backslash R^p$). After all kinds of collocations were identified, metrics of RE were calculated with the same Equations 1, 2, and 3 described above.

In addition to word-basis measurement, we also manually evaluated the performance of NER and RE. Two experts with domain knowledge of material science were invited to manually judge the prediction of models by their quality. For each prediction of the attribute, they need to give a score of 0 (incorrect), 1 (correct but with unrelated information), or 2 (correct). When they have different opinions on the same prediction, they negotiate with each other and give a final decision. Both 1 and 2 were counted as correct to calculate the accuracy of manual evaluation. However, there is a discrepancy in the evaluation scores: when using the exact match to evaluate, it is too strict for GPT-3.5 without format training, while manual evaluation ignores the form differences and may not be fair to the fine-tuned model (not counting its ability of EN and II). Thus, we further evaluate the performance of II, EN-U, and EN-T. The attributes selected do not appear in the original text, which means their target answers have changes in form, scale, or expression compared to corresponding parts in the original text. Only an exact match with the target answer is counted as correct.

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2024.100955.

## ACKNOWLEDGMENTS

T.X. expresses gratitude for the PhD scholarship provided by UNSW Future Manufacturing Institute. This study was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government, the department of education, and through the UNSW-NCI partner trial scheme. Microsoft Research also provides resources and support to this study through the Acceleration Foundational Models Research Grant. We thank C. Siu from the UNSW Library for assisting our team in obtaining text and data mining agreements with publishers. Additionally, we acknowledge H. Wen from Carnegie Mellon University for valuable discussions on developing the SII task, M. He and Y. Wang from University of New South Wales for SII schema design, and L. Duan from Australian National University for MDP schema design.

## AUTHOR CONTRIBUTIONS

T.X. and Y.W. proposed the question. T.X., Y.W., Q.L., and S.W. discussed and designed the experiments. T.X., Y.W., Y.Z., W.H., and Y.L. conducted the experiments, evaluation, and visualization. T.X., Y.W., Y.Z., and W.H. wrote the paper. C.G. and W.Z. gave advice on method and writing. C.K. and B.H. supervised the work.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 3, 2023
Revised: February 26, 2024
Accepted: February 27, 2024
Published: March 22, 2024

## REFERENCES

1. Jensen, Z., Kim, E., Kwon, S., Gani, T.Z.H., Román-Leshkov, Y., Moliner, M., Corma, A., and Olivetti, E. (2019). A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. ACS Cent. Sci. *5*, 892–899. https://doi.org/10.1021/acscentsci.9b00193.
2. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature *571*, 95–98. https://doi.org/10.1038/s41586-019-1335-8.
3. Huang, S., and Cole, J.M. (2022a). BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. J. Chem. Inf. Model. *62*, 6365–6377. https://doi.org/10.1021/acs.jcim.2c00035.
4. Huang, S., and Cole, J.M. (2022b). BatteryDataExtractor: battery-aware text-mining software embedded with BERT models. Chem. Sci. *13*, 11487–11495. https://doi.org/10.1039/d2sc04322j.
5. (2017). Empty rhetoric over data sharing slows science. Nature *546*, 327. https://doi.org/10.1038/546327a.
6. Olivetti, E.A., Cole, J.M., Kim, E., Kononova, O., Ceder, G., Han, T.Y.J., and Hiszpanski, A.M. (2020). Data-driven materials research enabled by natural language processing and information extraction. Appl. Phys. Rev. *7*. https://doi.org/10.1063/5.0021106.
7. Wang, L., Gao, Y., Chen, X., Cui, W., Zhou, Y., Luo, X., Xu, S., Du, Y., and Wang, B. (2023). A corpus of CO2 electrocatalytic reduction process extracted from the scientific literature. Sci. Data *10*, 175. https://doi.org/10.1038/s41597-023-02089-z.
8. Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V., and Ceder, G. (2019). Text-mined dataset of inorganic materials synthesis recipes. Sci. Data *6*, 203. https://doi.org/10.1038/s41597-019-0224-1.
9. Gao, Y., Wang, L., Chen, X., Du, Y., and Wang, B. (2023). Revisiting electrocatalyst design by a knowledge graph of cu-based catalysts for co2 reduction. ACS Catal. *13*, 8525–8534. https://doi.org/10.1021/acscatal.3c00759.
10. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The fair guiding principles for scientific data management and stewardship. Sci. Data *3*, 160018–160019. https://doi.org/10.1038/sdata.2016.18.
11. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Open and efficient foundation language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2302.13971.
12. Jacobsson, T.J., Hultqvist, A., García-Fernández, A., Anand, A., Al-Ashouri, A., Hagfeldt, A., Crovetto, A., Abate, A., Ricciardulli, A.G., Vijayan, A., et al. (2021). An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nat. Energy *7*, 107–115. https://doi.org/10.1038/s41560-021-00941-3.
13. Swain, M.C., and Cole, J.M. (2016). ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. J. Chem. Inf. Model. *56*, 1894–1904. https://doi.org/10.1021/acs.jcim.6b00207.
14. Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K.A., Ceder, G., and Jain, A. (2019). Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. J. Chem. Inf. Model. *59*, 3692–3702. https://doi.org/10.1021/acs.jcim.9b00470.
15. He, T., Huo, H., Bartel, C.J., Wang, Z., Cruse, K., and Ceder, G. (2023). Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. Sci. Adv. *9*, eadg8180. https://doi.org/10.1126/sciadv.adg8180.
16. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.B. (2019). Pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.

17. Zhao, X., Greenberg, J., An, Y., and Hu, X.T. (2021). Fine-Tuning BERT Model for Materials Named Entity Recognition. In 2021 IEEE International Conference on Big Data (Big Data) (IEEE), pp. 3717–3720. https://doi.org/10.1109/BigData52589.2021.9671697.

18. Trewartha, A., Walker, N., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K.A., Ceder, G., and Jain, A. (2022). Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns *3*, 100488. https://doi.org/10.1016/j.patter.2022.100488.

19. Sierepeklis, O., and Cole, J.M. (2022). A thermoelectric materials database auto-generated from the scientific literature using chemdataextractor. Sci. Data *9*, 648. https://doi.org/10.1038/s41597-022-01752-1.

20. Dong, Q., and Cole, J.M. (2022). Auto-generated database of semiconductor band gaps using ChemDataExtractor. Sci. Data *9*, 193–211. https://doi.org/10.1038/s41597-022-01294-6.

21. Beard, E.J., and Cole, J.M. (2022). Perovskite- and Dye-Sensitized Solar-Cell Device Databases Auto-generated Using ChemDataExtractor. Sci. Data *9*, 329–419. https://doi.org/10.1038/s41597-022-01355-w.

22. Mysore, S., Jensen, Z., Kim, E., Huang, K., Chang, H.-S., Strubell, E., Flanigan, J., McCallum, A., and Olivetti, E. (2019). The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In Proceedings of the 13th Linguistic Annotation Workshop, A. Friedrich, D. Zeyrek, and J. Hoek, eds. (Association for Computational Linguistics), pp. 56–64. https://doi.org/10.18653/v1/W19-4007.

23. Gupta, T., Zaki, M., Krishnan, N.M.A., and Mausam. (2022). MatSciBERT: A materials domain language model for text mining and information extraction. npj Comput. Mater. *8*, 102. https://doi.org/10.1038/s41524-022-00784-w.

24. Pawar, S., Palshikar, G.K., and Bhattacharyya, P. (2017). Relation extraction : A survey. Preprint at arXiv. https://doi.org/10.48550/arXiv.1712.05191.

25. Song, M., Kim, W.C., Lee, D., Heo, G.E., and Kang, K.Y. (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. J. Biomed. Inform. *57*, 320–332. https://doi.org/10.1016/j.jbi.2015.08.008.

26. Cejuela, J.M., Vinchurkar, S., Goldberg, T., Prabhu Shankar, M.S., Baghudana, A., Bojchevski, A., Uhlig, C., Ofner, A., Raharja-Liu, P., Jensen, L.J., and Rost, B. (2018). LocText: relation extraction of protein localizations to assist database curation. BMC Bioinf. *19*. 15-11. https://doi.org/10.1186/s12859-018-2021-9.

27. Dunn, A., Dagdelen, J., Walker, N., Lee, S., Rosen, A.S., Ceder, G., Persson, K., and Jain, A. (2022). Structured information extraction from complex scientific text with fine-tuned large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2212.05238.

28. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Adv. Neural. Inf. Process Syst. *33*, 1877–1901.

29. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. Adv. Neural. Inf. Process Syst. *35*, 27730–27744.

30. Lopez, S.A., Pyzer-Knapp, E.O., Simm, G.N., Lutzow, T., Li, K., Seress, L.R., Hachmann, J., and Aspuru-Guzik, A. (2016). The Harvard organic photovoltaic dataset. Sci. Data *3*, 160086–160087. https://doi.org/10.1038/sdata.2016.86.

31. Eibeck, A., Nurkowski, D., Menon, A., Bai, J., Wu, J., Zhou, L., Mosbach, S., Akroyd, J., and Kraft, M. (2021). Predicting power conversion efficiency of organic photovoltaics: models and data analysis. ACS Omega *6*, 23764–23775. https://doi.org/10.1021/acsomega.1c02156.

32. Liu, Y., Yan, W., Han, S., Zhu, H., Tu, Y., Guan, L., and Tan, X. (2022). How Machine Learning Predicts and Explains the Performance of Perovskite Solar Cells. Sol. RRL *6*, 1–11. https://doi.org/10.1002/solr.202101100.

33. Jablonka, K.M., Schwaller, P., Ortega-guerrero, A., and Smit, B. (2023). Is GPT-3 all you need for low-data discovery in chemistry. Preprint at ChemRxiv. https://doi.org/10.26434/chemrxiv-2023-fw8n4.

34. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. *28*, 31–36. https://doi.org/10.1021/ci00057a005.

35. Krenn, M., Ai, Q., Barthel, S., Carson, N., Frei, A., Frey, N.C., Friederich, P., Gaudin, T., Gayle, A.A., Jablonka, K.M., et al. (2022). SELFIES and the future of molecular string representations. Patterns *3*, 100588. https://doi.org/10.1016/j.patter.2022.100588.

36. Meftahi, N., Klymenko, M., Christofferson, A.J., Bach, U., Winkler, D.A., and Russo, S.P. (2020). Machine learning property prediction for organic photovoltaic devices. npj Comput. Mater. *6*, 166. https://doi.org/10.1038/s41524-020-00429-w.

37. Xie, T., Wan, Y., Wang, H., Østrøm, I., Wang, S., He, M., Deng, R., Wu, X., Grazian, C., Kit, C., and Hoex, B. (2023). Opinion mining by convolutional neural networks for maximizing discoverability of nanomaterials. J. Chem. Inf. Model. https://doi.org/10.1021/acs.jcim.3c00746.

38. Wan, Y., and Xie, T. (2024). SII MDP LLaMA. https://doi.org/10.6084/m9.figshare.24972699.v1.