Using Large Language Models for Data Extraction from Tables in Materials Literature

Defne Circi and **Ghazal Khalighinejad** and **Bhuwan Dhingra** and **L. Catherine Brinson**
Duke University, USA

Abstract

Advances in materials science require leveraging past findings and data from the vast published literature. Existing materials data repositories typically rely on newly created data in narrow domains because extracting detailed data from the enormous wealth of publications is immensely challenging. The advent of Large Language Models (LLMs) present a new opportunity to rapidly and accurately extract data from the published literature and transform it into structured data formats for easy query and reuse. In this paper, we build on initial strategies for using LLMs for rapid and autonomous data extraction from materials science articles in a format curatable by materials databases. We presented the subdomain of polymer composites as our example use case and demonstrated the success and challenges of LLMs on extracting tabular data. We explored different table representations for use with LLMs, finding that a multimodal model with an image output yielded the most promising results. This model achieved an accuracy score of 0.910 for composition information extraction. We envision that the results and analysis from this study will promote and accelerate further research directions in developing information extraction strategies from materials information sources.

1 Introduction

In this paper, we examine the effect of using different input types for information extraction from tables in polymer composite domain which will help scientists and engineers to easily find data without attempting to search through millions of relevant articles. The current official repository of reliable information on a large variety of materials data are peer reviewed research publications. However, due to their unstructured nature, it is difficult to utilize the vast majority of materials data locked in these articles and reports (Horawalavithana et al., 2022). Moreover, sifting through the articles is tedious, time-consuming, and error prone. Therefore, automation of the data curation process has gained increasing attention to enable rapid growth of a robust repository of prior published data (Yang, 2022; Olivetti et al., 2020; Dunn et al., 2022; Foppiano et al., 2023; Shetty and Ramprasad, 2021; Xie et al., 2023; Gilligan et al., 2023). Leveraging natural language processing (NLP) and large language models (LLMs) can make vital material information such as material identification, composition, properties, or experimental details readily available in a machine-readable format (Choi and Lee, 2023; Polak et al., 2023; Kononova et al., 2019; Wang et al., 2022; Shetty et al., 2023; Venugopal et al., 2021). Of the initial explorations of LLMs for information extraction from the scientific literature, most have focused on extraction from text only. The prior work indicates that while tables can be an excellent form to present condensed information for human readers, automated extraction of information from them remains a challenging task. Toward this end, we complement the structural understanding capabilities of the off-the-shelf LLMs, and their understanding of materials vocabulary, by using unique prompting and input types and evaluation strategies to explore viability of accurate and efficient knowledge extraction from tables in materials science papers. Our study focuses on extracting polymer composite sample information, where each sample is identified by its composition and is associated with property details. We employed LLMs, namely GPT-4 Turbo and GPT-4 Turbo with vision, for named entity recognition and relation extraction tasks in tables. Our study confronted several challenges that underscore the complexity of this task. These challenges include (a) layout challenges, such as merging multiple rows, (b) entity classification challenges, like differentiating between filler names and particle surface treatments (PST), and (c) relationship classification challenges, specifically in associating properties

with their names and metrological parameters. To explore the effectiveness of these models, we investigated how different input formats influence the table extraction process. Our findings contribute to the broader understanding of LLMs' capabilities in information extraction within scientific contexts, demonstrating both their potential and the challenges.

2 Methods

2.1 Article and dataset preparation

The data for this study consists of tables with information about polymer composite samples. Eighteen articles are selected from MaterialsMine (McGuinness et al., 2022). In this study, we focus on the composition and properties. Two graduate students annotated 37 tables to provide the ground truth. Within selected tables, each table has an average of approximately 4.9 samples with a minimum of 2 and a maximum of 15 samples for a total of 182 samples. We used 3 approaches for obtaining inputs of table data. All methods leverage GPT-4, with one using GPT4-Vision, and two approaches using digitization of the table, one in unstructured format using OCR, and the other using a structured tabular format. An example of different input types of a table can be seen in Figure 1.

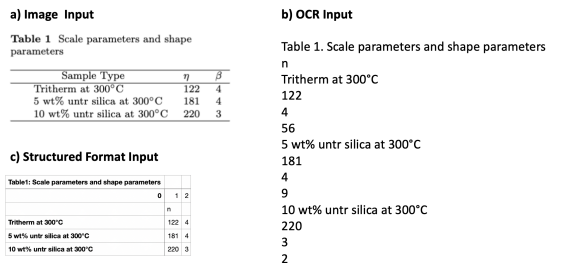

Figure 1: Example of the three different input types: a) GPT-4-V on sample table image b) GPT-4 on unstructured OCR c) GPT-4 on structured extracted table from pdf

2.2 Prompt design

Based on our knowledge of polymer composite materials, the key differentiating fields are matrix, filler, composition and PST. Therefore, we picked this minimal set to define the composition information of the samples. For each sample there are sets of material properties reported in the tables, such as storage modulus and glass transition temperature. For each property, we capture the name of the property, its value, unit and, if reported, conditions at which the property is measured such as temperature or pressure. Each condition has its own value and unit. We utilized the strength of few-shot prompting. The models extract the entities and find the relations simultaneously. The prompt includes a template JSON file to be filled along with a description of the task. Based on the selected option, the type of input table to be incorporated in the prompt is determined.

2.3 Evaluation

Our evaluation focused on comparing the extracted data against the set of annotated ground truth tables. Table 1 shows the accuracy scores of composition information. When a complete list is desired, it is necessary to penalize for missing some samples in the table in the predictions. In this case, the image input performed the best with a score of 0.910 and structured format without captions, structured format with captions and OCR gave accuracy scores of 0.832, 0.816 and 0.790, respectively. We also evaluated the results without penalizing for missing sample and found that structured format with captions gave the highest average accuracy with a score of 0.948.

Input type	Accuracy
Image	0.910 ± 0.037
OCR	0.790 ± 0.107
Structured (w/ captions)	0.816 ± 0.113
Structured (w/o captions)	0.832 ± 0.089

Table 1: Accuracy scores of composition information extraction with their 95% confidence intervals

Conclusion

Our work develops a rigorous method to compare different methodologies for materials science data extraction from tables using GPT-4 offering insights into the effectiveness of various techniques. We introduced an automated evaluation technique, contributing to a nuanced understanding of their performance. We also compiled, annotated and analyzed a dataset of tables in the polymer composite domain, providing a resource for further research. Our results indicate that using GPT4-V with appropriate prompting results in the best performance. This study also highlighted a number of challenges which underscore the complexities involved in information extraction and also pave the way for future research to address these issues.

Limitations

A notable limitation in our current approach is the separate evaluation of each table in an article. A more integrated method that merges information across all tables could offer a holistic view of each sample's properties, leading to a more comprehensive understanding. Additionally, our current methodology does not include the extraction of variations in numerical property values. Moreover, we assume a direct match in the ordering of samples, implying that each sample's position in the model output corresponds to the same position in the human-annotated dataset.

Acknowledgements

We thank NSF for funding support.

References

Jaewoong Choi and Byungju Lee. 2023. Accelerated materials language processing enabled by gpt. *arXiv preprint arXiv:2308.09354*.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.

Luca Foppiano, Pedro Baptista Castro, Pedro Ortiz Suarez, Kensei Terashima, Yoshihiko Takano, and Masashi Ishii. 2023. Automatic extraction of materials and properties from superconductors scientific literature. *Science and Technology of Advanced Materials: Methods*, 3(1):2153633.

Luke PJ Gilligan, Matteo Cobelli, Valentin Taufour, and Stefano Sanvito. 2023. A rule-free workflow for the automated generation of databases from scientific literature. *arXiv preprint arXiv:2301.11689*.

Sameera Horawalavithana, Ellyn Ayton, Shivam Sharma, Scott Howland, Megha Subramanian, Scott Vasquez, Robin Cosbey, Maria Glenski, and Svitlana Volkova. 2022. Foundation models of scientific knowledge for chemistry: Opportunities, challenges and lessons learned. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 160–172.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data*, 6(1):203.

Deborah McGuinness, Cate Brinson, Wei Chen, Chiara Daraio, Cynthia Rudin, Linda Schadler, Rebecca Cowan, Jamie McCusker, Samuel Stouffer, Neha Keshan, et al. 2022. Materialsmine: An open-source, user-friendly materials data resource guided by fair principles.

Elsa A Olivetti, Jacqueline M Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M Hiszpanski. 2020. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4).

Maciej P Polak, Shrey Modi, Anna Latosinska, Jinming Zhang, Ching-Wen Wang, Shanonan Wang, Ayan Deep Hazra, and Dane Morgan. 2023. Flexible, model-agnostic method for materials data extraction from text using general purpose language models. *arXiv preprint arXiv:2302.04914*.

Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kuenneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. 2023. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials*, 9(1):52.

Pranav Shetty and Rampi Ramprasad. 2021. Automated knowledge extraction from polymer literature using natural language processing. *Iscience*, 24(1).

Vineeth Venugopal, Sourav Sahoo, Mohd Zaki, Manish Agarwal, Nitya Nand Gosvami, and NM Anoop Krishnan. 2021. Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, 2(7).

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

Tong Xie, Yuwei Wa, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, and Bram Hoex. 2023. Large language models as master key: Unlocking the secrets of materials science with gpt. *arXiv preprint arXiv:2304.02213*.

Huichen Yang. 2022. Piekm: Ml-based procedural information extraction and knowledge management system for materials science literature. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 57–62.