Integrating Materials and Manufacturing Innovation (2024) 13:445–452
https://doi.org/10.1007/s40192-024-00356-4

THEMATIC SECTION: HARNESSING THE POWER OF MATERIALS DATA

Annotating Materials Science Text: A Semi-automated Approach for Crafting Outputs with Gemini Pro

Hasan M. Sayeed[1] · Trupti Mohanty[1] · Taylor D. Sparks[1]

Received: 26 February 2024 / Accepted: 10 April 2024 / Published online: 13 May 2024

Abstract

Recent advancements in large language models (LLMs) have paved the way for automated information extraction in the materials science domain. However, fine-tuning these models, crucial for effective machine learning pipelines in materials science, is hindered by a lack of pre-annotated data. Manual annotation, a laborious process, exacerbates the challenge. To address this, we introduce a tailored semi-automated annotation process, using Google's Gemini Pro language model. Our approach focuses on two key tasks: extracting information in structured JSON format and generating abstractive summaries from materials science texts. The collaborative process, a symbiotic effort between human annotators and the LLM, driven by structured prompts and user-guided examples, enhances the annotation quality and augments the LLM's capacity to comprehend materials science intricacies. Importantly, it streamlines human annotation efforts by leveraging the LLM's proficient starting point.

Introduction

In the realm of materials science research, where a vast repository of knowledge is encapsulated within peer-reviewed scientific literature, the extraction of structured information and concise summaries plays a pivotal role in advancing the field. The intricate details and discoveries within these texts hold the key to unlocking new insights and driving innovation. Structured information, organized in a machine-readable format, not only facilitates efficient data retrieval but also lays the groundwork for systematic analysis and comparison. Similarly, concise summaries distill the essence of comprehensive research articles, offering researchers, scientists, and industry professionals a quick and insightful overview.

A significant portion of experimental data in materials science remains locked within the confines of literature [1, 2]. Unlocking this wealth of high-quality data presents an opportunity to revolutionize materials informatics, potentially leading to the creation of databases that could, overnight, transform the field. Indeed, such databases have already demonstrated their utility in constructing models capable of identifying regions in materials composition space conducive to superconductivity [3], designing high-entropy alloys [4], predicting the emergence of novel magnets [5], and forecasting the *ZT* thermoelectric figure of merit in inorganic materials [6], among other applications.

However, existing databases in the field often revolve around calculated materials properties [7–10], introducing potential systematic errors and limiting the scope of data to quantities amenable to rapid computations. Additionally, there exists the risk that such calculations may not accurately reflect real-world scenarios. This suboptimal reliance on calculated properties highlights the need for more comprehensive and accessible experimental databases. While some attempts at constructing experimental databases exist, they often remain proprietary due to the associated labor costs [11–13]. Although a few open-access initiatives have initiated database construction efforts [14, 15], there is still ample room for the development of workflows aimed at enhancing accessibility and performance in automated materials database construction from literature—a fertile area for investigation. Our JSON structured data extraction method emerges as a promising avenue in this exploration.

Previous attempts to extract data from literature showcase the potential for automation. Pipelines facilitating automated extraction of compound–property relationships from unstructured battery-specific texts [16] and the extraction of a database of 300,000 polymer property records from

✉ Taylor D. Sparks
sparks@eng.utah.edu

[1] Department of Materials Science & Engineering, University of Utah, Salt Lake City, UT 84112, USA

650,000 abstracts [17] exemplify the strides made. Fine-tuned GPT models have been employed to extract general chemistry information from literature abstracts [18], demonstrating the versatility of such approaches. Domain-specific pre-training and fine-tuning have proven effective in enhancing transformer-based models' performance in domain-specific tasks [19–21]. As the emerging large language models (LLMs) are transformer-based, optimizing LLM pipelines for relevant domains becomes imperative, necessitating annotated data for fine-tuning.

Recent advancements in LLMs have presented exciting prospects for automating information extraction pipelines. Fine-tuning LLMs is crucial for constructing effective pipelines, especially in materials science. However, the scarcity of pre-annotated data poses a formidable challenge to realizing the full potential of LLMs. Manual annotation, a labor-intensive process, emerges as a bottleneck in this context. Relying solely on human annotation proves both expensive and time-consuming, while exclusive dependence on LLMs through few-shot learning risks errors due to the models' tendency to hallucinate. The urgent need for innovative annotation approaches is evident, aiming to facilitate the automation of pipelines that unlock the wealth of information embedded in scientific literature. A few previous semi-automatic approach exists like the semi-automatic annotation methodology for disinformation detection in news articles [22], labeling documents with noisy labels by LM followed by human correction of the labels [23], partially trained LLM in a human-in-the-loop annotation process to gradually decrease annotation time [24].

In this work, we advocate for a collaborative methodology that bridges human expertise with the capabilities of large language models (LLMs). We introduce a semi-automated text annotation process that harnesses the power of Google's Gemini Pro language model through structured prompts, ensuring the efficient handling of this annotation task (Fig. 1). The structured prompt feature of Gemini Pro facilitates incorporating examples directly in the prompt to leverage in-context learning which has already been proven effective by several previous studies [25–28]. Although our initial emphasis lay in annotating data for JSON extraction from plain text, it is crucial to acknowledge the occasional challenge posed by the token size limitations of LLMs, particularly when dealing with lengthy input texts. To address this limitation effectively, we expanded our annotation efforts to encompass an abstractive summarization task. This strategic augmentation provides a flexible solution to mitigate the mentioned constraint, also demonstrating the adaptability and versatility of our collaborative approach.

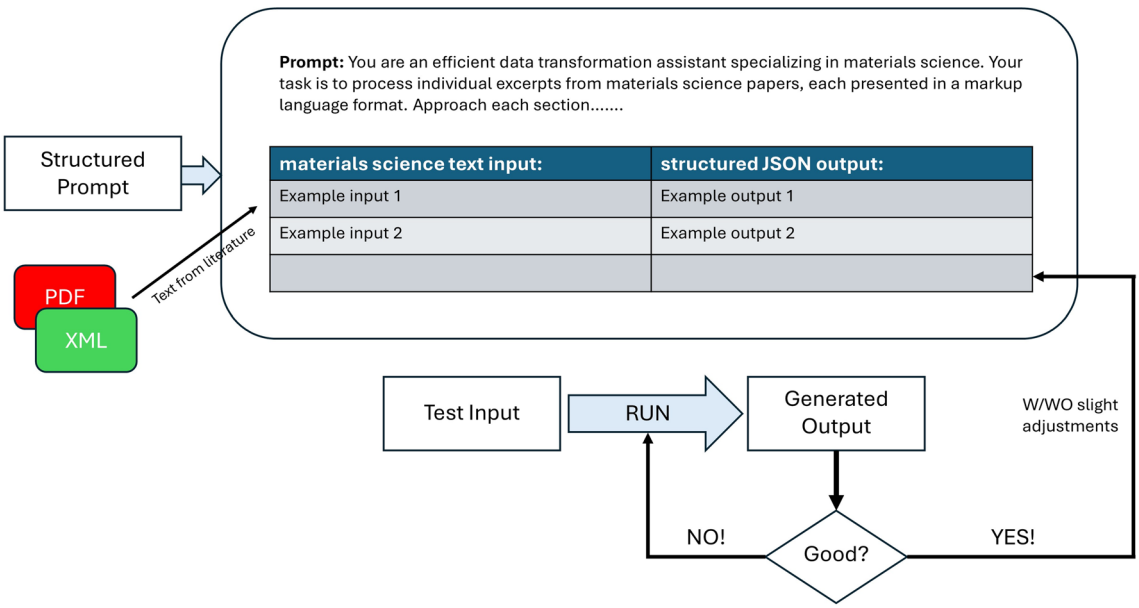

Fig. 1 Illustration of the semi-automated annotation workflow: The figure showcases the structured prompt methodology using Google's Gemini Pro language model for annotating materials science literature. The structured prompt, encompassing detailed instructions for extracting JSON or generating abstractive summaries, includes example input and desired output pairs. The provided examples, originating from literature, are human annotated gold standards. A test input is then processed by the model, and users evaluate the output's satisfaction. If satisfactory, the output is added to the example pool; otherwise, users can rerun the model or manually refine the output before inclusion in the pool. This semi-automated workflow combines human expertise with model guidance for efficient text annotation

Methodology

In this work, our approach revolves around leveraging Google's Gemini Pro language model through Google AI Studio leveraging structured prompts to annotate text data from materials science literature. The structured prompt capabilities of Gemini Pro play a crucial role in guiding the model to generate structured JSON and abstractive summaries.

Creating a structured prompt involves defining the instruction format, which, in our case, aligns with the task of generating structured JSON or abstractive summaries for materials science texts. Following the instructions provided in the documentation, we initiate the structured prompt by defining columns that represent the input and output structure. For the task of JSON extraction, this would entail a clear definition of the input text and the desired output in a structured JSON format. Similarly, for the abstractive summarization task, the columns would reflect the input text and the expected abstractive summary.

The creation of examples within the structured prompt is a critical step. Users provide input–output pairs as examples to guide the model in generating outputs that align with the desired format. In our context, these examples showcase the expected structure of the structured JSON or abstractive summaries for materials science texts. Importantly, these examples serve as the basis for the model's subsequent output when presented with test examples.

User-guided interaction with the model forms an integral part of our semi-automated workflow. This involves running the model with test examples and evaluating the generated output against the provided examples within the structured prompt. Users have the flexibility to determine whether the output is satisfactory. If deemed satisfactory, the output can be added to the example pool for future reference. In cases where the output needs refinement, users have the option to rerun the model for a new output or manually edit the output before adding it to the example pool.

As examples we provided text from materials science literature each section at a time. We opted to segment each paper by its sections to ensure that similar kinds of information remain together, enhancing the clarity and coherence of the annotations. By processing each section as independent examples, we also aimed to mitigate the token size limitations inherent in the language models. This strategic segmentation of texts not only facilitates the management of complex information but also enables more effective utilization of the LLM model's capabilities, ultimately contributing to the overall efficiency of the annotation process.

We illustrated this workflow, in Fig. 2 showcasing the structured prompt creation process. This figure help visualize the steps involved in creating structured prompts and running the model with user-guided examples, providing a comprehensive understanding of our semi-automated annotation approach.

Example Datasets

Our primary task involved annotating materials science text to generate datasets of structured JSONs, aiming to convert unstructured content into a machine-readable format. The structured JSON extraction task was meticulously guided by a detailed prompt, allowing the model to transform individual excerpts into coherent and well-organized JSON files, capturing vital information such as chemical compositions, processing conditions, characterization methods, and performance properties. As our main focus, this task provided a foundation for creating a valuable resource for materials informatics. Additionally, recognizing the token size limitations of LLMs when handling lengthy input texts, we extended our efforts to include an abstractive summarization task. While the summaries were not used for our primary objective, they can serve as a fallback strategy for cases where the token capacity of the LLM is exceeded, showcasing the versatility of our approach. Leveraging the NOUGAT tool [29] for parsing PDF files, we efficiently obtained text data in markup format from literature, facilitating our annotation endeavors.

We selected a diverse set of 10 articles spanning various domains within materials science, encompassing topics such as supercapacitors, high-entropy alloys, batteries, and ceramics. Each paper was meticulously segmented into its respective sections, resulting in a total of 86 individual sections across the 10 papers. By treating each section as an independent example, we aimed to maintain a focused approach, maximizing the model's ability to extract relevant information without being encumbered by details from previous sections. This strategic segmentation not only facilitated efficient processing but also ensured that the model could effectively capture the unique characteristics and nuances present within each section, ultimately contributing to the accuracy and coherence of the annotation process.

Structured JSON

To showcase the effectiveness of our semi-automated annotation workflow for structured JSON extraction from materials science literature, we employed a detailed prompt designed for accurate data transformation. The prompt served as a guiding framework, asking the model to act as an efficient data transformation assistant, specializing in materials science. This assistant was tasked with processing

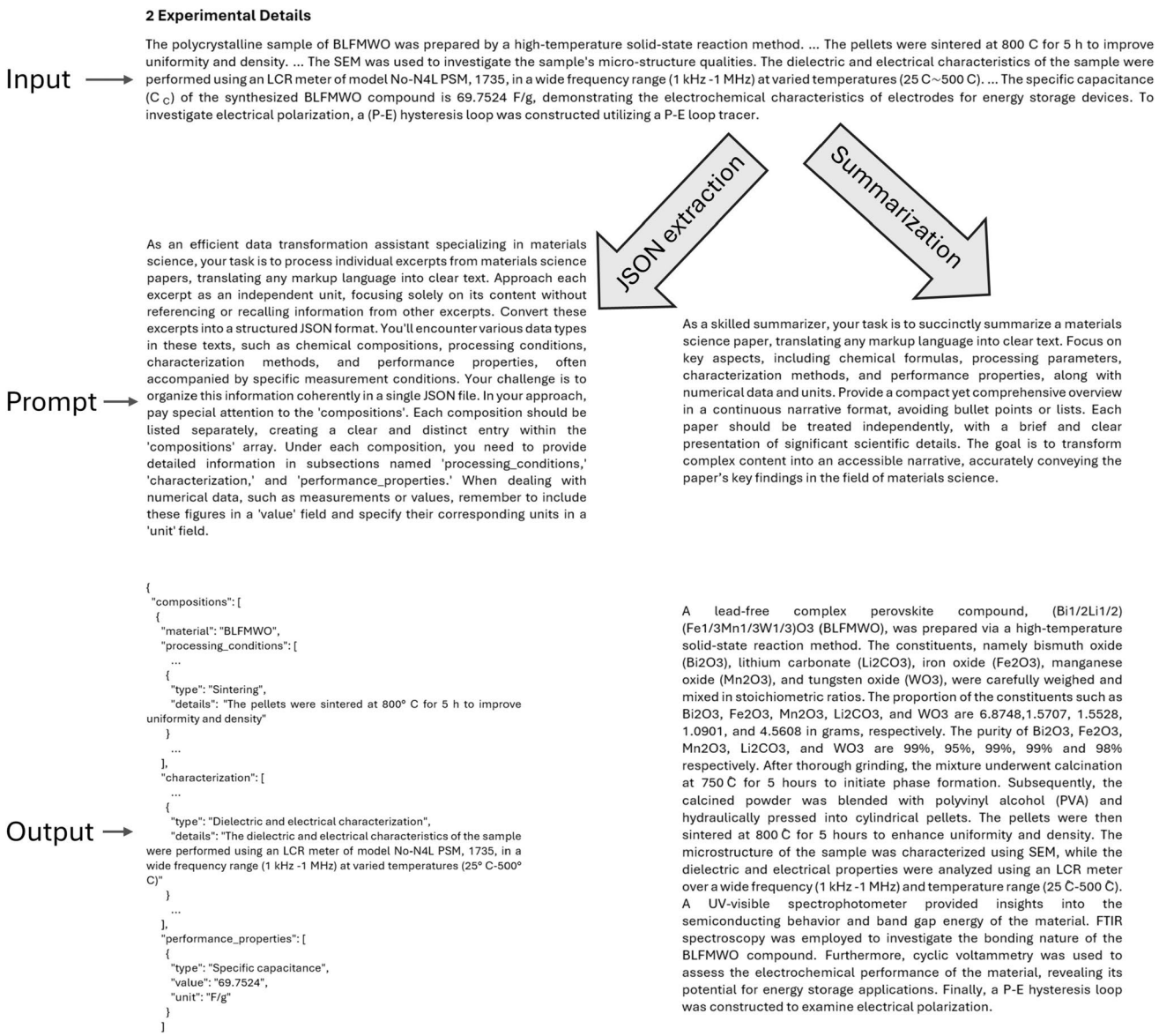

Fig. 2 Example tasks: The image illustrates the application of our semi-automated annotation process to materials science literature. At the top, a sample input text is provided, showcasing the complexity of scientific content. Below, on the left, the structured JSON extraction task is demonstrated, where the input text is transformed into a structured JSON format using our approach. This process provides a solid foundation for human annotators to edit and correct any mistakes, ensuring accuracy and refinement. On the right, the abstractive summary extraction task is showcased, with the input text condensed into a concise summary. We have also included the prompt used for both the task

individual excerpts from materials science papers, presented in a markup language format.

We provided text from materials science literature each section at a time. The comprehensive instructions instructed the assistant to treat each section independently, focusing solely on its content without referencing or recalling information from other sections it has seen before. The goal was to meticulously convert each excerpt into a structured JSON format, treating each section as a distinct entity. The annotated data includes various entities found in materials science texts, such as chemical compositions, processing conditions, characterization methods, and performance properties accompanied by specific measurement conditions.

For clarity and coherence, special emphasis was placed on the ‘compositions’ entity. Each composition was listed separately, creating clear and distinct sub-entries within the ‘compositions’ array. Within each composition, detailed information was organized into subsections named ‘processing_conditions,’ ‘characterization,’ and ‘performance_properties.’ Numerical data, such as measurements or values, was

included with corresponding units in a 'value' field and a 'unit' field, respectively.

To maintain the integrity and distinction of each composition, particularly when multiple compositions were present in the text, careful organization was instructed to avoid overlap and ensure clarity in representation. In instances where texts lacked the specified information, an empty JSON file was returned as an indication of no relevant data found.

To illustrate the structured JSON extraction process, we present one example in Fig. 2, highlighting the transformation of a materials science excerpt into a well-organized JSON file. Additionally, for researchers interested in accessing the annotated data, we have made it publicly available on Figshare, offering a valuable resource for further scientific analysis. This dataset serves as a testament to the precision and clarity embedded in our semi-automated annotation approach, showcasing its potential for efficient text annotation in the materials science domain.

Summaries

To demonstrate the effectiveness of our semi-automated annotation workflow for abstractive summarization of materials science literature, we employed a carefully crafted prompt tailored for skilled summarization. The prompt guided the model to distill the essence of materials science papers originally provided in markup language into well-structured and concise paragraphs.

In this summarization task, the focus was on eloquently capturing key aspects of the paper, including chemical formulas, processing parameters, characterization methods, and performance properties along with their respective measurement conditions. The model was instructed to create summaries that are shorter than the provided section, offering compact yet comprehensive overviews. Emphasis was placed on translating markup language notation into plain, easily understandable text, ensuring a fluid narrative that presents facts and tells the story of the paper.

Separate sections from each paper were treated as a distinct entity, and the model was directed to focus solely on the information provided in the current section without referencing or recalling details from previous papers or sections. The goal was to create summaries that weave in all relevant numerical data and units reported in the paper, maintaining scientific accuracy and integrity.

To illustrate the abstractive summarization process, we present one example in Fig. 2, showcasing the transformation of a materials science excerpt into a concise yet comprehensive summary. Additionally, for researchers interested in accessing the annotated data, we have made it publicly available on Figshare, offering a valuable resource for further scientific analysis. This dataset serves as evidence of the capability of our semi-automated annotation approach in producing informative and succinct summaries in the materials science domain.

Results

Our semi-automated annotation process yielded significant efficiency gains and improvements in annotation speed for both structured JSON extraction and abstractive summaries. Quantitative metrics were employed to evaluate the quality of the annotations, including ROUGE and BERTScore.

To provide insights into the model's performance without human intervention, we generated five additional examples solely using Gemini Pro. We then compared these examples with the gold standard examples prepared by our semi-automated approach, which involved human annotators collaborating with Gemini Pro. The average scores of these five examples were calculated for each metric, considering the stochasticity inherent in language models. These results will help us to assess the model's performance and provided valuable insights into the level of human intervention required to achieve high-quality annotations.

In evaluating structured JSON extraction, we encountered significant complexity stemming from the diverse types of entries found in materials science literature. Each JSON entry encapsulates a unique facet of scientific information, ranging from chemical compositions and processing conditions to characterization methods and performance properties. This diversity poses a significant challenge in devising a singular metric to comprehensively evaluate the fidelity of the extracted data. The heterogeneous nature of structured data complicates the application of traditional evaluation metrics such as F1 score, ROUGE, and BERTScore. Thus, quantifying the accuracy of JSON extraction proves to be a multifaceted task, demanding a more nuanced approach. While our initial attempt to establish a definitive evaluation metric encountered obstacles, we recognize the importance of devising an appropriate assessment framework. In future endeavors, we aim to develop a robust metric tailored to the intricacies of structured JSON extraction tasks like this, facilitating a more comprehensive evaluation of the model's performance in handling structured data. In the interim, we focused our evaluation efforts on the abstractive summarization task, which offered clearer insights into the model's performance and the requisite level of human intervention for data annotation.

Moving to the evaluation of abstractive summaries, we employed both ROUGE and BERTScore metrics to evaluate the quality of the generated summaries. The ROUGE score measured the similarity between the generated summaries and the gold standard summaries provided by human annotators, offering a quantitative assessment of the summarization process's fidelity to the original text. ROUGE-1,

ROUGE-2, and ROUGE-L scores of 0.58, 0.36, and 0.56 respectively were obtained, indicating a substantial overlap between the generated and gold standard summaries. Additionally, the BERTScore provided a more nuanced evaluation by assessing the semantic similarity between the generated and gold standard summaries, offering insights into the coherence and meaningfulness of the generated summaries. A BERTScore of 0.92 was achieved, indicating a high level of semantic similarity between the generated and gold standard summaries.

While evaluating the abstractive summaries offers valuable insights into the model's performance in summarizing the content of materials science literature, it is important to recognize that this assessment alone does not directly confirm the fidelity of the JSON extraction task. However, it is worth noting that abstractive summarization itself is a form of information extraction, albeit on a higher level of abstraction. Therefore, while our evaluation focuses on summarization, it indirectly sheds light on the model's capability to comprehend and extract meaningful information from the input text. This shared aspect of information extraction underscores the interconnectedness of the two tasks and suggests that the proficiency demonstrated in one task can inform our understanding of the model's performance in the other.

Our findings highlight the efficiency and effectiveness of our semi-automated annotation process in maintaining high levels of accuracy and quality. While manual annotation from scratch typically requires an average of 55 min per paper, our approach significantly reduces this labor-intensive process. The collaborative workflow with Gemini Pro streamlines the annotation process, with manual editing and corrections averaging around 15 min per paper for both structured JSON extraction and abstractive summarization tasks. This demonstrates the potential for our approach to minimize manual effort while ensuring high-quality annotations, making it a promising solution for accelerating data extraction from materials science literature.

Discussion

The semi-automated annotation workflow presented in this study represents a promising approach for extracting structured information from materials science literature. However, despite its efficacy, several challenges were encountered during the annotation process, highlighting areas for potential refinement. One significant challenge stems from the fact that the language model utilized, Google's Gemini Pro, is not fine-tuned specifically for materials science. Consequently, the model's performance may vary when applied to domain-specific texts, necessitating a substantial amount of human intervention to ensure annotation accuracy. To address this challenge, future refinements could involve fine-tuning language models specifically tailored to the materials science domain. Notably, our semi-automated approach has the potential to play a pivotal role in this refinement process. By providing high-quality annotated training datasets, our approach facilitates the development of more specialized fine-tuned models capable of accurately extracting information from materials science literature. This iterative loop, wherein our approach enhances its own efficiency and effectiveness, holds promise for continuously improving the semi-automated annotation process and advancing toward the development of fully automated language models tailored for seamless and accurate data extraction from materials science literature.

Another challenge we faced during the annotation process was the evaluation of the accuracy of the JSON extraction task. The absence of a well-defined metric for evaluating structured JSON extraction hindered our ability to comprehensively assess the fidelity of the extracted data. Developing a robust evaluation metric tailored specifically to the complexities of structured data extraction from materials science literature could be a focus for future work, enabling more nuanced assessments of the model's performance. By addressing this need, future refinements could further enhance the effectiveness and reliability of semi-automated annotation workflows, ultimately advancing the development of fully automated language models tailored for materials science literature.

While our study focused on utilizing a single language model, it is worth considering the implications of employing multiple models for benchmarking purposes. Introducing multiple models into the annotation workflow could lead to variations in data extraction preferences, resulting in differences in the extracted data even when all models produce valid outputs. This variability would significantly impact benchmarking practices, particularly concerning the establishment of ground truth annotations. Exploring how different language models interpret and extract information from materials science literature could provide valuable insights into the robustness and generalizability of annotation workflows. Future research endeavors may consider conducting comparative analyses across multiple language models to elucidate the extent of variability in data extraction and its implications for benchmarking fairness and accuracy in materials science literature annotation.

Given the subjective nature of summarization, exploring the ROUGE score across summaries by multiple experts could provide deeper insights into the evaluation metric's comprehensiveness. However, in this work, we did not utilize summaries from multiple experts. Notably, the BERTScore metric, employed in our study for assessing semantic similarity between generated and gold standard summaries, focuses on contextual embeddings, which are

less dependent on subjective judgments. Therefore, the use of BERTScore mitigates the need for multiple expert summaries, as it primarily evaluates the semantic relevance of the generated summaries to the reference summaries, rather than relying on subjective interpretations of summary quality.

In addition to addressing challenges, the practical implications and applications of semi-automated annotation for structured JSON and summaries are profound. One notable advantage is the potential to significantly reduce the financial and labor costs associated with manual annotation efforts. By combining human expertise with the capabilities of language models, our workflow streamlines the annotation process, enabling efficient extraction of structured information from large volumes of text data. Furthermore, the semi-automated nature of the workflow minimizes the risk of errors commonly associated with manual annotation, ensuring the production of high-quality annotations suitable for various downstream applications. Overall, the adoption of semi-automated annotation holds great promise for accelerating research in materials science by facilitating the creation of comprehensive and machine-readable datasets essential for advancing scientific discovery and innovation.

Conclusion

In conclusion, our study introduces a semi-automated annotation methodology tailored for materials science literature, leveraging Google's Gemini Pro language model. This approach offers a promising avenue for structured information extraction, enhancing the efficiency of annotation tasks while ensuring the creation of high-quality datasets essential for training domain-specific language models. By utilizing structured prompts, we have demonstrated the model's capability to accurately extract information and generate concise summaries from complex scientific texts. Our work highlights the potential of semi-automated annotation workflows in bridging the gap between manual annotation and fully automated processes, paving the way for more efficient and accurate data extraction in materials science research.

Moving forward, future research directions include refining language models specifically optimized for materials science literature through fine-tuning and domain-specific training. Additionally, exploring strategies such as active learning, where the model interacts with the user to identify areas requiring further annotation, and integrating domain knowledge bases to provide context-specific information, could enhance the efficiency and accuracy of the annotation process. These advancements aim to reduce the need for extensive human intervention while ensuring the creation of high-quality annotated datasets essential for training and improving language models tailored to the materials science domain. By pursuing these avenues for improvement, we can continue to advance semi-automated annotation techniques, ultimately facilitating more efficient and accurate data extraction in materials science research.

Acknowledgements We acknowledge the assistance provided by ChatGPT, which was used for rephrasing and achieving coherence. However, it is important to note that all core ideas, text, tables, and figures were the original work of the authors. This research was supported by the National Science Foundation (NSF) under grant number DMR 2334411. We extend our appreciation to the NSF for their financial support, which made this study possible.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Olivetti EA, Cole JM, Kim E, Kononova O, Ceder G, Han TY-J, Hiszpanski AM (2020) Data-driven materials research enabled by natural language processing and information extraction. Appl Phys Rev 7:041317
2. Sayeed HM, Smallwood W, Baird SG, Sparks TD (2024) NLP meets materials science: quantifying the presentation of materials data in scientific literature. Mater Sci 7(3):723–727
3. Isayev O, Fourches D, Muratov EN, Oses C, Rasch K, Tropsha A, Curtarolo S (2015) Materials cartography: representing and mining materials space using structural and electronic fingerprints. Chem Mater 27:735–743
4. Lederer Y, Toher C, Vecchio KS, Curtarolo S (2018) The search for high entropy alloys: a high-throughput ab-initio approach. Acta Mater 159:364–383
5. Sanvito S, Oses C, Xue J, Tiwari A, Zic M, Archer T, Tozman P, Venkatesan M, Coey M, Curtarolo S (2017) Accelerated discovery of new magnets in the Heusler alloy family. Sci Adv 3:e1602241
6. Xi L, Pan S, Li X, Xu Y, Ni J, Sun X, Yang J, Luo J, Xi J, Zhu W et al (2018) Discovery of high-performance thermoelectric chalcogenides through reliable high-throughput material screening. J Am Chem Soc 140:10785–10793
7. Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor RH, Nelson LJ, Hart GL, Sanvito S, Buongiorno-Nardelli M et al (2012) AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. Comput Mater Sci 58:227–235
8. Talirz L, Kumbhar S, Passaro E, Yakutovich AV, Granata V, Gargiulo F, Borelli M, Uhrin M, Huber SP, Zoupanos S et al (2020) Materials Cloud, a platform for open computational science. Sci Data 7:299
9. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G et al (2013) Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. APL Mater 1:011002
10. Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, Rühl S, Wolverton C (2015) The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. npj Comput Mater 1:1–15
11. Zagorac D, Müller H, Ruehl S, Zagorac J, Rehme S (2019) Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. J Appl Crystallogr 52:918–925

12. Groom CR, Bruno IJ, Lightfoot MP, Ward SC (2016) The Cambridge structural database. Acta Crystallogr Sect B Struct Sci Cryst Eng Mater 72:171–179
13. Blokhin E, Villars P (2020) The PAULING FILE project and materials platform for data science: from big data toward materials genome. In: Handbook of materials modeling: methods: theory and modeling, pp 1837–1861
14. Vaitkus A, Merkys A, Gražulis S (2021) Validation of the crystallography open database using the crystallographic information framework. J Appl Crystallogr 54:661–672
15. Gallego SV, Perez-Mato JM, Elcoro L, Tasci ES, Hanson RM, Momma K, Aroyo MI, Madariaga G (2016) MAGNDATA: towards a database of magnetic structures. I. The commensurate case. J Appl Crystallogr 49:1750–1776
16. Huang S, Cole JM (2022) BatteryBERT: a pretrained language model for battery database enhancement. J Chem Inf Model 62:6365–6377
17. Shetty P, Rajan AC, Kuenneth C, Gupta S, Panchumarti LP, Holm L, Zhang C, Ramprasad R (2023) A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. npj Comput Mater 9:52
18. Dunn A, Dagdelen J, Walker N, Lee S, Rosen AS, Ceder G, Persson K, Jain A (2022) Structured information extraction from complex scientific text with fine-tuned large language models. arXiv preprint arXiv:2212.05238
19. Trewartha A, Walker N, Huo H, Lee S, Cruse K, Dagdelen J, Dunn A, Persson KA, Ceder G, Jain A (2022) Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns 3:100488
20. Beltagy I, Lo K, Cohan A (2019) SciBERT: a pretrained language model for scientific text. arXiv preprint arXiv:1903.10676
21. Gupta T, Zaki M, Krishnan NA, Mausam (2022) MatSciBERT: a materials domain language model for text mining and information extraction. Npj Comput Mater 8:102
22. Bonet-Jover A, Sepúlveda-Torres R, Saquete E, Martınez-Barco P (2023) A semi-automatic annotation methodology that combines Summarization and Human-In-The-Loop to create disinformation detection resources. Knowl Based Syst 275:110723
23. Jain S, Van Zuylen M, Hajishirzi H, Beltagy I (2020) SciREX: a challenge dataset for document-level information extraction. arXiv preprint arXiv:2005.00512
24. Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, Persson KA, Jain A (2024) Structured information extraction from scientific text with large language models. Nat Commun 15:1418
25. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. In: Advances in neural information processing systems, vol 33, pp 1877–1901
26. Rubin O, Herzig J, Berant J (2021) Learning to retrieve prompts for in-context learning. arXiv preprint arXiv:2112.08633
27. Reynolds L, McDonell K (2021) Prompt programming for large language models: Beyond the few-shot paradigm. In: Extended abstracts of the CHI conference on human factors in computing systems, pp 1–7
28. Zhang H, Zhang X, Huang H, Yu L (2022) Prompt-based meta-learning for few-shot text classification. In: Proceedings of the 2022 conference on empirical methods in natural language processing, pp 1342–1357
29. Blecher L, Cucurull G, Scialom T, Stojnic R (2023) Nougat: neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418