

## Data Extraction from Polymer Literature using Large Language Models

Corresponding Author: Professor Rampi Ramprasad

**This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.**

Version 0:

Decision Letter:

**\*\* Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors \*\***

Dear Professor Ramprasad,

Thank you for submitting your manuscript, "Data Extraction from Polymer Literature using Large Language Models", to Communications Materials. It has now been seen by 3 referees, whose comments are appended below. You will see that while they find your work of interest, some important points are raised about the level of detail and transparency related to the methodology, data validation, assessment and discussion in your work. We are interested in the possibility of publishing your study in Communications Materials, but would like to consider your response to these concerns in the form of a revised manuscript before we make a decision on publication.

We therefore invite you to revise and resubmit your manuscript, taking into account the points raised.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

When submitting your revised manuscript, please include the following:

-A response letter with a point-by-point reply to each of the referee comments and a description of changes made. Please include the complete referee report in the response letter. Please note that the response letter must be separate to the cover letter to the editors.

-A marked-up version of the manuscript with all changes to the text in a different colored font. Please do not include tracked changes or comments. Please select the file type 'Revised Manuscript - Marked Up' when uploading the manuscript file to our online system.

-A clean version of the manuscript. Please select the file type 'Article File'.

-An updated <https://www.nature.com/documents/nr-editorial-policy-checklist.zip> Editorial Policy checklist, uploaded as a 'Related Manuscript File' type. This checklist is to ensure your paper complies with all relevant editorial policies. If needed, please revise your manuscript in response to these points. Please note that this form is a dynamic 'smart pdf' and must therefore be downloaded and completed in Adobe Reader. Clicking this link will download a zip file containing the pdf.

In the event that your manuscript is accepted we will provide detailed guidance on our journal policies and formatting. You may however wish to ensure that the manuscript complies with our house style at this stage. See our style and formatting guide (<https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf>) and checklist (<https://www.nature.com/documents/commsj-phys-style-formatting-checklist-article.pdf>) for reference.

Data availability statements and data citations policy: All Communications Materials manuscripts must include a section titled "Data Availability" at the end of the Methods section or main text (if no Methods). More information on this policy, and a list of examples, is available at <http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf>.

In particular, the Data availability statement should include:

- Accession codes for deposited data
- Other unique identifiers (such as DOIs and hyperlinks for any other datasets)
- At a minimum, a statement confirming that all relevant data are available from the authors
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.

DATA SOURCES: We strongly encourage authors to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided at <http://www.nature.com/sdata/policies/repositories>.

If a community resource is unavailable, data can be submitted to generalist repositories such as <https://figshare.com/> or <http://datadryad.org/> Dryad Digital Repository. Please provide a unique identifier for the data (for example a DOI or a permanent URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publically available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be further cited in the methods reference section.

Please refer to our data policies at <http://www.nature.com/authors/policies/availability.html>.

Please use the following link to submit your documents:

Link Redacted

\*\* This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first \*\*

We hope to receive your revised paper within three months; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we will close your file. In this event, we will still be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Materials or published elsewhere in the meantime.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Milica Todorovic, PhD  
Editorial Board Member  
Communications Materials  
[orcid.org/0000-0003-0028-0105](https://orcid.org/0000-0003-0028-0105)

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

The paper presents a method for extracting polymer property data from materials science literature using a combination of Named Entity Recognition (NER)-based Models (MaterialsBERT) and Large Language Models (LLMs) such as GPT-3.5. While these methods and their use in materials science data extraction are not new, the field is developing very quickly and no single method for information extraction has been established, therefore different approaches or modifications of methods are appreciated, especially if authors can prove that the performance of their approach is superior to previous ones (although I'm not sure if they can in this case). The problem of data extraction is an important one to address in general, and particularly relevant as the scientific community increasingly looks towards big data and AI to drive discoveries in materials science. By making the extracted data publicly available, the study supports further research and application development in the field. However, there are significant issues that would need to be addressed before the paper is published, the key one being a proper assessment of the accuracy of extracted data

I appreciate the authors' effort to validate the data, with validation against possible property names, units and polymer names. However, the key challenge in accurate data extraction is the correct relation, i.e. what material does a given value and property correspond to. Just checking for validity of individual items is insufficient without validating the relations between them, as it may pass tests where items from a whole paragraphs are mixed and therefore incorrect. This may happen often for the simpler NER-based methods.

The few-shot prompting idea is generally good, but it may be tricky - LLMs are very good at pattern recognition, and often in a few-shot prompting approach, when not able to complete the task properly they will simply follow the examples, making up

semantically correct but fake data. This may artificially increase the number of valid datapoints, as they will pass the correct property name, correct polymer name and correct unit tests.

This data validation makes me particularly concerned. The authors state that "We found that the performance of the models is contingent on the property being extracted", and then they report F1 scores only for Tg and band gap. What about the other 22 properties? It would also be useful to report precision and recall - in addition to F1.

In addition, the relatively low overlap between the different methods (Fig 4 c,d) suggests significant problems with accuracy and either low recall of both methods, low precision, or both. The biggest concern is that in Fig. 4 d, for band gap we see a 9.76% overlap between MaterialsBERT and GPT-3.5, and earlier in the text the authors report F1 scores for the two methods of 0.77 and 0.85 or 0.87. This is simply not possible. With a <10% overlap between the two methods, at least one of the methods (or both) must perform badly. Two methods with F1 scores close to 0.8 cannot overlap on only 10% of points.

It also seems like the manual data verification was performed on abstracts only (another thing not really mentioned in the paper and has to be found in the supplementary). This is not necessarily indicative of how the methods would perform on real-life full-texts of research papers, as abstracts are usually written in a very particular way and not the same as the full-text of the paper.

The above points are all to say that it is absolutely crucial to establish estimates of the accuracy of extracted datapoints in their entirety including all the individual fields and their relation. The authors have to be more diligent in performing that assessment and be more transparent on how its done to facilitate reproducibility.

Following that, no comparison of performance with respect to other methods is given. There are established NER-based methods (such as ChemDataExtractor), as well as very successful GPT-based approaches (e.g. <https://pubs.rsc.org/en/content/articlehtml/2023/dd/d3dd00019b>). A comparison to other methods need to be performed and superiority in at least some aspects has to be demonstrated to justify the use of the presented approach.

Since the same data is extracted with various methods, it would be valuable to see parity plots. For those datapoints that are common between two different methods e.g. GPT-3.5 and MaterialsBERT, a parity plot should be a straight x=y line. For each point that does not lay on that line, at least one of the methods is wrong (at least one, because both may be incorrect as well, both if the points match and if they don't). Figure 4 shows both results on one plot, but from such plot we can only infer that the distributions are somewhat similar, which is not sufficient. This will give additional feedback on the quality of the results.

In the introduction the authors write that "an additional NER filter was applied to identify paragraphs containing all necessary named entities such as material name, property name, and property value (...). The texts of the filtered paragraphs were then passed to either MaterialsBERT for NER-based data extraction, or to the OpenAI API for GPT-3.5-based data extraction." This seems unclear - if NER filter is applied for all necessary entities, how does that differ from actual NER-based extraction? Also, the NER is relatively simple, simpler than GPT-3.5. If the NER does the identification incorrectly then it automatically disqualifies that paragraph from further analysis by GPT-3.5, which may have been successful. This also does not really line up with Fig 4 (c,d) where GPT found points that BERT NER has not, how is that possible if they all went through the NER filter first? This is not very clear.

The authors do not discuss any data extraction from tables until "Outlook" at the very end of the paper. And they immediately give some claims why it's not possible without any previous attempts or other details. Tables are one of the simplest sources to extract numerical data from research papers, most certainly simpler than extraction from raw text, as the data is already structured, it essentially only needs to be transferred to the dataframe that the authors use for building their database. The argument about papers being PDFs also seems not to be entirely true - first, most modern papers are available in XML or HTML, second PDFs do not pose a significant challenge, with methods as simple as OCR designed for tables (dozens available through a simple google search), machine learning methods (e.g. <https://github.com/microsoft/table-transformer>), through simply letting GPT-4-vision read the tables, through more advanced (and accurate) workflows combining machine learning and OCR (e.g. <https://iopscience.iop.org/article/10.1088/2632-2153/ad2f52>). The approach to tables has to be revisited, since, as the authors themselves state, there is "substantial amount of extractable and relevant data in tabular format".

The two examples above illustrate one of the other general problems - the authors should try to be more clear about all they do and explain in more detail. Perhaps through dividing the paragraphs into some sub-paragraphs, or some other restructurization. Each thought has to be explained thoroughly, things cannot be mentioned without thorough explanation in the context that it is used in. Similar

The authors decided to structure the data as json in the GPT-3.5 extraction. However, it seems like that structure is prompted for rather than obtained directly using OpenAI capabilities (<https://platform.openai.com/docs/guides/text-generation/json-mode>). Why has this choice been made? It seems like utilizing the capabilities to natively output json rather than prompt for it could be more successful, since this is one less task for the LLM to perform. What is the success rate of GPT-3.5 outputting a proper json output when prompted to do so?

Reproducibility of the results is very important. The methods section severely lacks and is insufficient to even attempt to reproduce the results. Not even a model name for GPT-3.5 was mentioned, and throughout GPT-3.5 existence there have been at least half a dozen, all with different performance. All parameters to all used models: GPT, MaterialsBERT and LLaMA have to be described.

I suggest removing or at least significantly reducing the discussion about monetary costs associated with running proprietary models. These costs change (decrease) rapidly and any discussion will become irrelevant soon, if not even before publishing. As I'm reviewing this, the costs are already outdated. Looking at Fig. 2 (g) and the mentioned numbers of tokens and costs, the current cost would be ~3x lower, and the cost per token will likely keep decreasing.

The paper overall has potential to be good, but only after a substantially major revision. It seems that the authors rushed with its submission. I suggest thoroughly and diligently addressing the concerns expressed above as well as trying to elucidate on things that were not explained.

#### Reviewer #2 (Remarks to the Author):

This manuscript is about materials data extraction from polymer-related literature, using various large language models. The authors compared the performance and cost-effectiveness of GPT-3.5, LLaMa 2, MaterialsBERT, concluding that the performance of GPT-3.5 was the best but MaterialsBERT is useful in terms of cost-effectiveness. They collected more than million dataset consist of 24 properties from 106,000 unique polymers. The extracted data as well as model comparison information will be highly helpful for the researchers in this field. Therefore, the reviewer recommends to the publication of this work to "Communications Materials", after addressing the minor comments below.

- As the authors stated, (for example, non-oriented vs. oriented PMMA layer) the materials property can vary with the synthesis condition, even the name of polymer is same. One can easily imagine that all the materials property will change with the molecular weight. But in this work, only 'material name' is extracted. At least molecular weight or morphology information should be collected.
- The authors collected the data from full text as well. In full text, there can be several property data for multiple materials, for example, the developed material vs. control group. Did authors consider which data belongs to which material? If so, how did it proceed?
- In introduction, these papers [Commun. Mater. 5, 13 (2024)], [J. Mater. Chem. A 11, 17628 (2023)] can be relevant.

#### Reviewer #3 (Remarks to the Author):

Ramprasad et al. presented a well-designed study using different LLMs to extract 24 properties of over 106,000 polymers from a large number of literature sources. The data is analyzed and discussed in detail, and overall, I enjoyed reading this manuscript. Here are my comments:

1. The authors mentioned that the 681,000 documents for extraction were selected by searching the term "poly." It would be helpful for readers if you could provide more information when it is first mentioned (I personally do not see the details until reaching to the end of the article) and the distribution of the sources, as mentioned on page 21. Additionally, please clarify if these documents were automatically downloaded and, if so, which software was used. This information is crucial for reproducibility and understanding the data collection process. Furthermore, I am curious whether the inclusion of papers from other domains containing the term "poly" could influence the extracted data output.
2. It would be beneficial to discuss the selection criteria for the 24 properties and why there is an interest in extracting these specific properties. For example, in ref 24, the extracted data were used for predictive analysis. I believe it would be valuable to inform the reader not only about 'how' to extract data and how well the models perform but also what can be done with the extracted data, showcasing its potential applications.
3. Based on Figure 1, it appears that there is already a well-designed NLP-based workflow, with the last step allowing for a choice between NER-based extraction and LLM-based extraction. Has this workflow (from "full text corpus" to "NER filter") been reported before? In my opinion, one advantage of LLM-based extraction is its ease of use and reduced requirement for coding/NLP knowledge, making it more accessible to the general public. However, mixing traditional NLP workflow and LLM-based extraction may not necessarily highlight this advantage, as a majority of the framework still relies on traditional NLP (and the researchers who are not familiar with these techniques need to invest time on learning). While this is acceptable, I suggest commenting on future perspectives and whether this pipeline could be entirely governed by LLMs alone.
4. How was the 1,000 subset for evaluation selected? Given that this subset consists of only a small portion of the entire dataset, it would be helpful to discuss how the evaluation based on such a small set can be representative of the whole dataset.
5. In the data availability section, it would be beneficial to provide a link to the prompt used and the code/API to run the LLMs like LLaMa 2 and GPT, allowing readers to reproduce the results. Alternatively, the author could include these in the supporting information file or create a dedicated GitHub repository for this purpose.
6. In the supporting information, it is difficult to read the exact false positive numbers in Fig. S1. I recommend labeling the y-

axis values on each panel or providing a separate table listing the exact values to improve readability and interpretation.

**\*\* Visit Nature Research's author and referees' website at <a href="http://www.nature.com/authors">www.nature.com/authors</a> for information about policies, services and author benefits\*\***

Communications Materials is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the <a href="http://platformsupport.nature.com/">Platform Support Helpdesk</a>.

Version 1:

Decision Letter:

**\*\* Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors \*\***

Dear Dr Ramprasad,

Thank you for resubmitting your manuscript, "Data Extraction from Polymer Literature using Large Language Models", to Communications Materials. It has now been seen again by the 3 referees, whose comments are appended below. You will see that while they find your work improved, some important points are still raised, mostly related to the veracity of your performance metrics and the suggested use of your extraction pipeline. We are, of course, still interested in the possibility of publishing your study in Communications Materials, but would like to consider your additional response to these concerns in the form of a revised manuscript (particularly in relation to the remarks raised by Reviewer #1) before we make a decision on publication.

We therefore invite you to revise and resubmit your manuscript, taking into account the points raised.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

When submitting your revised manuscript, please include the following:

-A response letter with a point-by-point reply to each of the referee comments and a description of changes made. Please include the complete referee report in the response letter. Please note that the response letter must be separate to the cover letter to the editors.

-A marked-up version of the manuscript with all changes to the text in a different colored font. Please do not include tracked changes or comments. Please select the file type 'Revised Manuscript - Marked Up' when uploading the manuscript file to our online system.

-A clean version of the manuscript. Please select the file type 'Article File'.

-An updated <a href="https://www.nature.com/documents/nr-editorial-policy-checklist.zip">Editorial Policy</a> checklist, uploaded as a 'Related Manuscript File' type. This checklist is to ensure your paper complies with all relevant editorial policies. If needed, please revise your manuscript in response to these points. Please note that this form is a dynamic 'smart pdf' and must therefore be downloaded and completed in Adobe Reader. Clicking this link will download a zip file containing the pdf.

In the event that your manuscript is accepted we will provide detailed guidance on our journal policies and formatting. You may however wish to ensure that the manuscript complies with our house style at this stage. See our style and formatting guide (<https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf>) and checklist (<https://www.nature.com/documents/commsj-phys-style-formatting-checklist-article.pdf>) for reference.

Data availability statements and data citations policy: All Communications Materials manuscripts must include a section titled "Data Availability" at the end of the Methods section or main text (if no Methods). More information on this policy, and a list of examples, is available at <http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf>.

In particular, the Data availability statement should include:

- Accession codes for deposited data
- Other unique identifiers (such as DOIs and hyperlinks for any other datasets)
- At a minimum, a statement confirming that all relevant data are available from the authors
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.

DATA SOURCES: We strongly encourage authors to deposit all new data associated with the paper in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided at <http://www.nature.com/sdata/policies/repositories>.

If a community resource is unavailable, data can be submitted to generalist repositories such as <https://figshare.com/> or <http://datadryad.org/> Dryad Digital Repository. Please provide a unique identifier for the data (for example a DOI or a permanent URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publically available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be further cited in the methods reference section.

Please refer to our data policies at <http://www.nature.com/authors/policies/availability.html>.

Please use the following link to submit your documents:

Link Redacted

\*\* This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first \*\*

We hope to receive your revised paper within three months; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we will close your file. In this event, we will still be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Materials or published elsewhere in the meantime.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Milica Todorovic, PhD  
Editorial Board Member  
Communications Materials  
[orcid.org/0000-0003-0028-0105](https://orcid.org/0000-0003-0028-0105)

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

I appreciate the author's thorough response to my review. While many of my comments were addressed appropriately, there are still a few points that are problematic and have not been addressed properly - particularly regarding the accuracy assessment in response 4).

I understand that the <10% overlap is low due to differences in the extracted names of the materials. Even if they are technically equivalent, they may not be literally the same. While not unifying them is a deficiency of the presented results (because very straightforward NLP techniques exist for comparing two non-literally identical phrases to determine whether they have the same meaning), it is not that important if it is explained, which the authors did. That being said, the presented values still contradict the previously presented results. Even disregarding the overlap argument, the pure number of data points (13,929 for BERT and 88,128 for GPT) does not mathematically allow for one of the two methods to have a high F1 score. A discrepancy in the number of data points that high must mean that at least one of the methods has either horrible precision or horrible recall, which should be obvious. Therefore, the F1 score - a direct consequence of precision and recall - cannot be high. The lowest F1 score mentioned in the paper seems to be 0.66 for BERT and 0.87 for GPT. A back-of-the-envelope estimate for the maximum F1, assuming perfect recall for GPT, suggests that the maximum F1 for BERT in that

case is around 0.15 - based solely on the significant difference in the amount of extracted data between the two methods, which is a best case scenario. This should be obvious. The authors should revisit all of their accuracy estimates before the publishing process continues. If, however, I am mistaken in my reasoning (which, to my understanding, I am not), that means the text and figure description of the accuracy assessments is unclear or insufficient and should be thoroughly revised.

Reviewer #2 (Remarks to the Author):

The authors fully addressed the reviewer's concerns. Now the reviewer recommends to the publication of this work to "Communications Materials".

Reviewer #3 (Remarks to the Author):

I have recently reviewed this manuscript again and have read through the author's response. I believe they have done a good job addressing most of the concerns I raised in the last round. Therefore, I think this paper can be accepted.

One minor point that I feel is not adequately addressed is my comment #3. This comment is more of an advice rather than a strict requirement. Specifically, I mentioned that the author should better explain how their work aims to train downstream ML models. While the author has provided some examples from their previous work, a more systematic approach would be beneficial. I suggest including a table that outlines specific properties and examples, along with relevant work. This table can be placed in the supplementary information for readers.

\*\* Visit Nature Research's author and referees' website at <http://www.nature.com/authors> for information about policies, services and author benefits\*\*

Communications Materials is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the [Platform Support Helpdesk](http://platformsupport.nature.com/).

Version 2:

Decision Letter:

\*\* Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors \*\*

Dear Professor Ramprasad,

Your manuscript titled "Data Extraction from Polymer Literature using Large Language Models" has now been seen again by our referees, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Materials.

We therefore invite you to edit your manuscript to comply with our journal policies and formatting style in order to maximise the accessibility and therefore the impact of your work.

#### EDITORIAL REQUESTS

\* Your manuscript should comply with our policies and format requirements, detailed in our style and formatting guide (<https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf>).

\* Please edit your manuscript according to the editorial requests in the attached table, and outline revisions made in the right hand column. If you have any questions or concerns about any of our requests, please do not hesitate to contact me. It is

important that each request be addressed in order to avoid delays in accepting your manuscript. Please upload the completed table with your manuscript files as a Related Manuscript file.

\* The editorial requests table also includes a full list of the files that must be provided upon resubmission. Please upload your files according to this table.

\* An updated editorial policy checklist that verifies compliance with all required editorial policies must be completed and uploaded with the revised manuscript. All points on the policy checklist must be addressed; if needed, please revise your manuscript in response to these points. Please note that this form is a dynamic 'smart pdf' and must therefore be downloaded and completed in Adobe Reader. Clicking this link will download a zip file containing the pdf.

Editorial policy checklist: <https://www.nature.com/documents/nr-editorial-policy-checklist.zip>

#### OPEN ACCESS

Communications Materials is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit <https://www.nature.com/commsmat/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit your revised files:

Link Redacted

\*\* This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first \*\*

We hope to hear from you within two weeks; please let us know if the process may take longer.

Best regards,

Milica Todorovic, PhD  
Editorial Board Member  
Communications Materials  
[orcid.org/0000-0003-0028-0105](https://orcid.org/0000-0003-0028-0105)

#### REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

The authors addressed most of my concerns. I have no further comments.

\*\* Visit Nature Research's author and referees' website at <http://www.nature.com/authors> for information about policies, services and author benefits\*\*

**Open Access** This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

**Manuscript COMMSMAT-24-0322-T**  
**Response to Reviewers**

Dear Editor and Reviewers,

Thank you for your feedback on our manuscript titled “*Data Extraction from Polymer Literature using Large Language Models*”. We have carefully studied the valuable comments and questions raised by the reviewers and tried to address them in this revision. Our detailed reply to each of the reviewers’ comments, along with the corresponding modifications made to the manuscript, are outlined below.

**Reviewer #1 (Remarks to the Author):**

The paper presents a method for extracting polymer property data from materials science literature using a combination of Named Entity Recognition (NER)-based Models (MaterialsBERT) and Large Language Models (LLMs) such as GPT-3.5. While these methods and their use in materials science data extraction are not new, the field is developing very quickly and no single method for information extraction has been established, therefore different approaches or modifications of methods are appreciated, especially if authors can prove that the performance of their approach is superior to previous ones (although I'm not sure if they can in this case). The problem of data extraction is an important one to address in general, and particularly relevant as the scientific community increasingly looks towards big data and AI to drive discoveries in materials science. By making the extracted data publicly available, the study supports further research and application development in the field.

However, there are significant issues that would need to be addressed before the paper is published, the key one being a proper assessment of the accuracy of extracted data

1. I appreciate the authors' effort to validate the data, with validation against possible property names, units and polymer names. However, the key challenge in accurate data extraction is the correct relation, i.e. what material does a given value and property correspond to. Just checking for validity of individual items is insufficient without validating the relations between them, as it may pass tests where items from a whole paragraphs are mixed and therefore incorrect. This may happen often for the simpler NER-based methods.

**Response:**

We thank the reviewer for this comment, and we agree with the sentiment. We note and emphasize though that the primary purpose of the NER and LLM-based pipelines, or any NLP-based data extraction pipelines in materials science, per se, is to extract the entities of the data so that accurate relationship is clearly identified. If such relations don't need to be recognized, complicated methods or LLMs are not necessary, as a simple script can extract the individual items from a given text. Thus, our usage of language models implicitly comes with recognizing, preserving and extracting appropriate material-property relationships (as further elaborated with validation below).

To emphasize this aspect, we have added the following text to the ‘Overview of the Data Extraction Pipelines’ section of the manuscript.

“During the extraction process, the relationship extraction module of MaterialsBERT processes the identified entities to determine and establish correct relationships using heuristic rules. GPT-3.5, on the other hand, automatically identifies relationships between the entities by itself.”

The F1 score validation metrics implemented in our work checks if the relationships are correctly identified, it does not simply check if the individual items are extracted. We have added new texts to the ‘Performance Benchmarking for a Labeled Subset of the Full Corpus’ section of the manuscript to clarify this.

“Our  $F_1$  score computation methodology necessitates that the extracted data be completely present within the provided text. Thus, to ensure precise data extraction, it is imperative to identify all entities, specifically ‘material’, ‘property’, ‘value’, and ‘unit’ within the specified paragraph for each extracted data point.”

2. The few-shot prompting idea is generally good, but it may be tricky - LLMs are very good at pattern recognition, and often in a few-shot prompting approach, when not able to complete the task properly they will simply follow the examples, making up semantically correct but fake data. This may artificially increase the number of valid datapoints, as they will pass the correct property name, correct polymer name and correct unit tests.

**Response:**

We have considered the probability of hallucination during the validation process. We do this by not only checking if all the entities are extracted correctly but also checking if they are from the correct source paragraph. As a result, any semantically correct fake data points generated by LLMs from the given few shots are also caught by the validation process because those fake data points do not exist in the curated ground truth data extracted from the given text.

We have added the following text to the ‘Performance Benchmarking’ section of the manuscript to explain this.

“Furthermore, to verify the correct relationships among these entities, they must correspond accurately with their respective entities in the curated dataset to account for hallucination by the LLMs. The  $F_1$  score calculation method not only verifies the accuracy of the completely extracted data but also confirms their origin from the appropriate source paragraph, because any semantically accurate but fabricated data points produced by LLMs would be absent in the curated ground truth data extracted from the provided text. Additional details about the calculation of the  $F_1$  score are discussed in the **supplementary information.**”

3. This data validation makes me particularly concerned. The authors state that "We found that the performance of the models is contingent on the property being extracted", and then they report F1 scores only for Tg and band gap. What about the other 22 properties? It would also be useful to report precision and recall - in addition to F1.

**Response:**

While we generally agree with this comment and the desire to provide validation for the other 22 properties, we would like to emphasize that creating labeled data for validation purposes is extremely time- and resource-intensive. For instance, the labeled Tg and bandgap dataset used for validation purposes were curated by two of our very talented graduate and undergraduate students over the course of two semesters to collect, cross-check and curate. Needless to say, validating for all 24 properties will require considerable effort.

In addition to providing validation studies for Tg and band gap (which we think would be representative of other properties, albeit with variations between properties which are to be expected), we have performed multiple stages of filtering and validation (see the steps mentioned in the beginning of the ‘Data Extraction from Full Texts’ section) and in Fig. 4f-k to confirm the validity of the extracted data. Specifically, to provide indirect validation, we checked for correct trends and correlations between the other 22 properties as well as their range based on extensive literature review and domain knowledge. Finally, we are making all the extracted data available to the community at <https://polymerscholar.org> so that anyone can directly check the data and its source and report erroneous datapoints. Thus, while no one expects any such framework to be 100% perfect, this approach significantly simplifies the task of data access and data availability.

We have updated the text in the ‘Data Extraction from Full Texts’ section of the manuscript to clarify this.

“Given that neither pipeline can achieve perfect accuracy, and manually curating data sets for all 24 properties requires a significant effort, we conducted additional validation of the extracted data in a post-processing step, including examining general trends and correlations between properties.”

We have discussed the true positive, false positives and false negative values for Tg and bandgap in the SI in detail.

4. In addition, the relatively low overlap between the different methods (Fig 4 c,d) suggests significant problems with accuracy and either low recall of both methods, low precision, or both. The biggest concern is that in Fig. 4 d, for band gap we see a 9.76% overlap between MaterialsBERT and GPT-3.5, and earlier in the text the authors report F1 scores for the two methods of 0.77 and 0.85 or 0.87. This is simply not possible. With a <10% overlap between the two methods, at least one of the methods (or both) must perform badly. Two methods with F1 scores close to 0.8 cannot overlap on only 10% of points.

**Response:**

This is a good point and something we have tried to address in the manuscript.

To clarify, the low overlap in Fig. 4c is a consequence of the results shown in Fig. 3c and 3d. Because GPT-3.5 can understand more complex relationships and extract more data, there is considerably a large number of paragraphs where MaterialsBERT fails and GPT-3.5 is successful. This is also captured by Fig. 3g, where MaterialsBERT obtained an F-1 score of 0.66 vs. the F-1 score of 0.85 obtained by GPT-3.5. Please note that, the F-1 score of MaterialsBERT is 0.66, and the F-1 score of 0.77 is for LLaMa-2. Therefore, the results shown in Fig. 4c agree with the F-1 scores reported earlier and do not contradict them.

Further clarifications are added to the discussion of Fig. 4c to explain it better.

“This supports the previously reported  $F_1$  scores, demonstrating that GPT-3.5 is capable of understanding more intricate relationships and extracting more data. Notably, there is a significant number of paragraphs where MaterialsBERT fails ( $F_1 = 0.66$ ), in contrast to the success of GPT-3.5 ( $F_1 = 0.85$ ).”

The reviewer’s concern of just 10% overlap with lower individual F1 scores than he/she points out, is still valid. We clarify why this could occur below. One factor is that MaterialsBERT faces issues while extracting materials names correctly (please see the SI section on “Limitations of MaterialsBERT”); the low overlap between the material names in Fig. 4d is a consequence of how material names are extracted differently by the pipelines. We have investigated this phenomenon with examples in Table S3, and in the conclusions that GPT-3.5 tends to extract detailed names of the materials, while MaterialsBERT fails to correctly extract the names of blends and composites. To summarize, even if two pipelines extract a data point correctly, there can be mismatch between material names extracted by the pipelines. For instance, using the example discussed in the manuscript, while calculating F-1 scores, we treated instances such as ‘non-oriented PMMA layer’ and ‘oriented PMMA layer’ extracted by GPT-3.5 as valid material names. At the same time, we also treated ‘PMMA’, as extracted by MaterialsBERT, as a correct material name. Obviously, none of the name variations are wrong, and materials, especially polymers can have multiple co-referents. Fig. 4d demonstrates this exact phenomenon to highlight the difference between the two pipelines.

We have added the following text to the discussion of Fig. 4d to explain it better.

“It is important to mention that variations in naming and co-referents of polymers were considered correct during the calculation of the  $F_1$  score. Consequently, although there is a low overlap of the names exactly extracted by the methods as depicted in Fig. 4d, a high  $F_1$  score is still maintained.”

5. It also seems like the manual data verification was performed on abstracts only (another thing not really mentioned in the paper and has to be found in the supplementary). This is not necessarily indicative of how the methods would perform on real-life full-texts of research papers, as abstracts are usually written in a very particular way and not the same as the full-text of the paper.

**Response:**

In the submitted version of the manuscript, we have mentioned the source of the manually curated dataset multiple times, e.g., in the LLM Data Extraction section, in the Performance Benchmarking section, and also in the Methods section of the manuscript that the data were curated from 630 abstracts and used to calculate the F1 scores.

Even if there are semantic differences in how the abstracts are written (e.g., subject verb agreement, tone, voice, tense, brevity) compared to how a paragraph of the full text of a paper is written, the pipelines are robust enough to handle such cases, because, the semantic style of the materials related entities, i.e., ‘material’, ‘property’, ‘value’ and ‘unit’ usually remain the same no matter where they are discussed in the document. Based on our decade worth of experience in performing informatics on polymer properties, and we believe the abstracts are representative of the full texts for this purpose.

One of the strengths of our method is that the pipelines work on small chunks of texts, e.g. (a paragraph or abstract). The pipelines extract data that is present in the given text in its entirety, such that, for a data point to be correctly extracted, all the items (i.e., ‘material’, ‘property’, ‘value’ and ‘unit’) need to be present in the given paragraph. This is a characteristic (or limitation) we have discussed in the Outlook section of the manuscript. Despite that, the pipelines could extract more than a million complete data points.

To clarify, following sentence is added to the Outlook section.

“Our pipelines extract data that is described completely in each paragraph by looking for all the required named entities (i.e., ‘material’, ‘property’, ‘value’ and ‘unit’) to establish correct relationships.”

6. The above points are all to say that it is absolutely crucial to establish estimates of the accuracy of extracted datapoints in their entirety including all the individual fields and their relation. The authors have to be more diligent in performing that assessment and be more transparent on how its done to facilitate reproducibility.

**Response:**

We thank the reviewer for emphasizing this important point. We believe we have addressed this in comment (1) earlier and as part of our other comment responses. To reiterate, we have added further discussion on F1 score calculation, and discussed how we check for the validity of all items to mark it as a true positive or false positive, etc. in detail in the SI.

7. Following that, no comparison of performance with respect to other methods is given. There are established NER-based methods (such as ChemDataExtractor), as well as very successful GPT-based approaches (e.g. <https://pubs.rsc.org/en/content/articlehtml/2023/dd/d3dd00019b>). A comparison to other

methods need to be performed and superiority in at least some aspects has to be demonstrated to justify the use of the presented approach.

**Response:**

We thank the reviewer for this suggestion. We have added texts mentioning the presence of such comparative study and metrics in the manuscript.

Our present work builds on top of our previous work on the MaterialsBERT pipeline, (as discussed in the Introduction). We have compared different NER-based methods in that work. For example, we have calculated the performance of MaterialsBERT against PubMedBERT, MatBERT, BioBERT, ChemBERT on ChemDNER, ChemRxnExtractor datasets (See Table 2, 3 of Ref. [1]). Overall, the MaterialsBERT pipeline, tuned to extract material properties from the abstracts of journal articles, demonstrated superior performance over the other models for property data extraction tasks. Therefore, the metrics reported here extends the previously reported ones, and the LLM clearly shows improvements over the existing methods.

We have added the following text in the Performance Benchmarking section of the manuscript.

“We previously demonstrated the superior performance of MaterialsBERT compared to other existing NER models [1]. The higher  $F_1$  scores compared to the MaterialsBERT obtained in this work thus clearly underscores the advantage of the LLMs over the NER models.”

As for a fair comparison with other GPT-based approaches, we are not aware of any similar general-purpose pipelines that can extract property data from the polymer literature. The reference mentioned in the comment, for example, is tuned to perform data extraction tasks for Au NPs. As materials science domain experts, we appreciate the considerable difference in the polymer and metal nanoparticle literature, properties and applications.

8. Since the same data is extracted with various methods, it would be valuable to see parity plots. For those datapoints that are common between two different methods e.g. GPT-3.5 and MaterialsBERT, a parity plot should be a straight  $x=y$  line. For each point that does not lay on that line, at least one of the methods is wrong (at least one, because both may be incorrect as well, both if the points match and if they don't). Figure 4 shows both results on one plot, but from such plot we can only infer that the distributions are somewhat similar, which is not sufficient. This will give additional feedback on the quality of the results.

**Response:**

We thank the reviewer for this suggestion and agree that this is a great idea. We have included such parity plots in the SI (see Fig. S2).

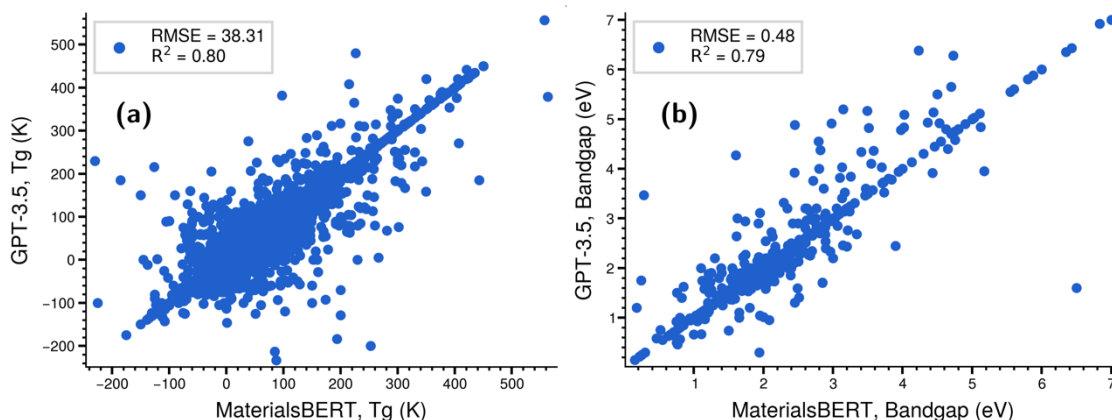


FIG. S2. Parity plot for (a) Tg and (b) bandgap data extracted by the GPT-3.5 and the MaterialsBERT pipelines from the full texts of all polymer related articles. The plots show data where both pipelines extracted the same polymer names and are from the same papers.

We updated the following sentence to refer to this figure while discussing Fig. 4d.

“Notably, only 11, 042 material names exactly matched between the pipelines, as illustrated in Fig. 4d, and in the parity plots showing overlap between the pipelines in Fig. S2.”

9. In the introduction the authors write that "an additional NER filter was applied to identify paragraphs containing all necessary named entities such as material name, property name, and property value (...) The texts of the filtered paragraphs were then passed to either MaterialsBERT for NER-based data extraction, or to the OpenAI API for GPT-3.5-based data extraction." This seems unclear - if NER filter is applied for all necessary entities, how does that differ from actual NER-based extraction?

**Response:**

There are two stages of a typical NER-based data extraction pipeline. NER-filtering (called “NER-filter” in Figure 1) is the first step that identifies the named entities mentioned in a given text. In the second step (called “NER-based extraction” in Figure 1), the NER pipeline attempts to establish correct relationships between the identified named entities and extracts the data. This is a common approach adapted by many past works which we briefly discussed in the introduction (see Ref. [1] for MaterialsBERT, Ref. [2], [3], [4] for other similar methods for additional details).

10. Also, the NER is relatively simple, simpler than GPT-3.5. If the NER does the identification incorrectly then it automatically disqualifies that paragraph from further analysis by GPT-3.5, which may have been successful. This also does not really line up wit Fig 4 (c,d) where GPT found points that BERT NER has not, how is that possible if they all went through the NER filter first? This is not very clear.

**Response:**

To clarify this, we have added this text to the Overview of the Pipelines section of the manuscript.

“Regardless of the final data extraction model, the NER filter is utilized to verify the presence of ‘material’, ‘property’, ‘value’, and ‘unit’ entities in the given paragraph. The absence of any of these entities would preclude the extraction of a complete data point. This filter also assists LLMs in accurately identifying relationships without which they may generate placeholder values such as ‘not mentioned’, ‘n/a’, ‘-’, or even hallucinate false data if an entity is not present in the text.”

As for Fig. 4 (c,d), the apparent discrepancy arises because the NER-filter and the NER-based extraction modules are separate from each other. Even if the NER-filter could find the presence of the target named entities in a text, the extraction module of MaterialsBERT can fail to establish a correct relationship. Fig. 4c indicates that GPT-3.5 is significantly better at such task.

11. The authors do not discuss any data extraction from tables until "Outlook" at the very end of the paper. And they immediately give some claims why it's not possible without any previous attempts or other details. Tables are one of the simplest sources to extract numerical data from research papers, most certainly simpler than extraction from raw text, as the data is already structured, it essentially only needs to be transferred to the dataframe that the authors use for building their database.

**Response:**

Effective tabular data extraction is by no means trivial. Our main point is that it is very difficult to create a general pipeline to extract tabular data because of the arbitrary nature of how the tables are formatted in scientific texts, how the column headers are named and how they are discussed in the full texts. These challenges are already known in the domain, see Ref. [5] for an excellent discussion on this matter.

The primary challenge of the NLP-methods is not simply extracting the entities mentioned in a table, but rather establishing correct relationships between the extracted entities. Extracting the contents of a table in a data frame is obviously a trivial process and a simple Python script can be used to do that. However, creating a general pipeline that can correctly establish relationships between the various types of property data mentioned in the various cells of the table, and with the discussion in the texts (e.g., for the 24 properties we have considered here), is an open challenge.

Existing methods, as a result, attempt to simplify the problem by focusing on a handful of properties, or only a specific class of materials for which tables are usually constructed following a consistent format (see Ref. [6], [7]). These relationship extraction algorithms are completely rule-based systems that suffer from generalizability or are expert-defined and cannot be translated algorithmically at scale to 610,000 articles and the 24 properties that we have successfully processed in this work.

We thank the reviewer for this comment, and we have tried to clarify this further by adding relevant citations to the ‘Outlook’ section.

“However, tables are frequently arbitrarily structured, necessitating meticulous filtering, classification, and pre-processing for correct relationship establishment [5], [8].”

12. The argument about papers being PDFs also seems not to be entirely true - first, most modern papers are available in XML or HTML, second PDFs do not pose a significant challenge, with methods as simple as OCR designed for tables (dozens available through a simple google search), machine learning methods (e.g. <https://github.com/microsoft/table-transformer>), through simply letting GPT-4-vision read the tables, through more advanced (and accurate) workflows combining machine learning and OCR (e.g. <https://iopscience.iop.org/article/10.1088/2632-2153/ad2f52>). The approach to tables has to be revisited, since, as the authors themselves state, there is "substantial amount of extractable and relevant data in tabular format".

**Response:**

We agree with the statement that most modern papers are available in XML or HTML format. This is why it was not necessary for our current pipelines to rely on PDF files.

The reason PDFs pose a significant challenge is because the PDF format does not have any set standard, and different applications generate PDF files differently. In the case of XML format or its derivatives (e.g. HTML), the parts of a document are clearly formatted with unique tags or identifiers (e.g. H1 for heading, P or SPAN for texts, IMG for the figures, TABLE for the tables, CSS classes for the document title, author list, affiliations and so on). In contrast, PDF is a format that simply tells a document reader where glyphs, lines, and boxes go on a page and in what font, size, and color.

Currently available PDF file parsers (whether it is OCR-based or not), therefore, tries to scan the characters one by one to extract chunks of texts from a file. They are not able to detect what section that text is from or how it connects to the document in any string-like way. For example, the texts from the footnote, headers or footers of a document very easily get mixed with the paragraph of a main body, which significantly ‘confuses’ the downstream pipelines employing NER or LLM models.

Existing works that attempt to extract data from PDF files need to first identify the coordinates or bounding box of each section of the document, which depends on the publisher and the formatting of the journal article. This requires a lot of careful processing using heuristic rules or similar techniques, which is out of the scope of this article. See Ref. [9] for an example.

OCR methods are useful to extract text if the texts in the PDF file is formatted as images (usually when the PDF is generated by scanning physical documents). That is why, OCR-based approach is not needed in most cases since most modern scientific articles are published with text-based/digitally generated PDF formats. Multiple software and packages are currently available to read the texts form digitally generated PDF files (e.g. PyPDF, MuPDF). However, the challenge of clearly identifying the section of text chunks without mixing with each other remains.

Having worked in the polymer informatics domain for over a dozen years, we have great appreciation for the challenges involved in collecting data from various sources (including tables and PDF files). We are thus reluctant to trivialize the difficulties inherent in effectively extracting data in a format that can be truly useful.

13. The two examples above illustrate one of the other general problems - the authors should try to be more clear about all they do and explain in more detail. Perhaps through dividing the paragraphs into some sub- paragraphs, or some other restructurization. Each thought has to be explained thoroughly, things cannot be mentioned without thorough explanation in the context that it is used in. Similar

**Response:**

As stated in the Abstract and in the Introduction, the purpose of the manuscript is to conduct a comparative study of how the newer LLM-models perform against the already existing NER-based pipelines for property data extraction tasks. During the writing of this manuscript, we have assumed that readers are already familiar with the NER-related methods, and there are excellent review articles already available in the literature. We have briefly introduced the readers to these concepts in the Introduction, and extensively cited the relevant articles for additional details when available. The concepts we use to discuss our workflow and the issues mentioned in the Outlook section of the manuscript are, therefore, not brand new. The SI also contains adequate information on many of the details.

Efforts of this type straddle two very different areas of enquiry, namely materials science and language models, with widely different needed backgrounds to attain 100% comprehension of both topics, which we are not sure is achievable through one manuscript. In writing this manuscript, we have been mindful of how long the manuscript needs to be, while also capturing the essential details; as mentioned above, we have been diligent in pointing the reader towards other resources/references they may wish to consult.

14. The authors decided to structure the data as json in the GPT-3.5 extraction. However, it seems like that structure is prompted for rather than obtained directly using OpenAI capabilities (<https://platform.openai.com/docs/guides/text-generation/json-mode>). Why has this choice been made? It seems like utilizing the capabilities to natively output json rather than prompt for it could be more successful, since this is one less task for the LLM to perform. What is the success rate of GPT-3.5 outputting a proper json output when prompted to do so?

**Response:**

This is a great point. While on the surface, enforcing an exact output format (JSON in this case) seems like a good idea, existing implementations of the LLM format enforcing algorithms can steer the LLMs in an unwanted direction. When a specific output format is enforced, the language model is not allowed to freely control the text generation, for example, white spacing, token optionality and ordering, etc. It means that the language model may not be generating the texts that it ‘wants to’ generate and could put its internal states in a suboptimal level, reducing the quality of the output in later timesteps. That is

why, the OpenAI API (please see the mentioned URL) requires specifying the word ‘JSON’ in the prompt. Omitting such a word causes the model to generate “an unending stream of whitespace and the request may run continually until it reaches the token limit”. The JSON-mode only guarantees a valid JSON formatted output that can be parsed without any errors.

In contrast, parsing JSON response using Python is trivial; incorrect ones can be easily flagged and avoids forcing the LLM to generate valid JSON response. In our work, GPT-3.5, being a large LLM processing a small chunk of texts, could easily produce valid JSON strings in more than 99.8% of the cases when prompted to do so. So, it was not necessary to enforce the JSON-mode.

15. Reproducibility of the results is very important. The methods section severely lacks and is insufficient to even attempt to reproduce the results. Not even a model name for GPT-3.5 was mentioned, and throughout GPT-3.5 existence there have been at least half a dozen, all with different performance. All parameters to all used models: GPT, MaterialsBERT and LLaMA have to be described.

**Response:**

As reported in the Methods section, we have made all our codes and model publicly available in Github and HuggingFace Hub, so interested readers can run the pipelines and reproduce our results.

Please see <https://github.com/Ramprasad-Group/PromptDataExtraction> for the codes and <https://huggingface.co/pranav-s/MaterialsBERT> for the NER model. The extracted datasets are publicly available free of charge via Polymer Scholar (<https://polymerscholar.org/>). Researchers can validate each individual data point by clicking on it which links to the source article in the web interface to check the accuracy themselves.

We have updated the Method section to include the version number and other parameters of GPT-3.5 used in this work.

“The GPT-3.5-turbo-0613 model hosted by OpenAI was used to extract data from text. The OpenAI Python package was used to access the OpenAI API. The temperature parameter was set to 0.001 for text generation by the model, with all other parameters remaining at their default values.”

16. I suggest removing or at least significantly reducing the discussion about monetary costs associated with running proprietary models. These costs change (decrease) rapidly and any discussion will become irrelevant soon, if not even before publishing. As I'm reviewing this, the costs are already outdated. Looking at Fig. 2 (g) and the mentioned numbers of tokens and costs, the current cost would be ~3x lower, and the cost per token will likely keep decreasing.

**Response:**

We thank the reviewer for the suggestion. Though the API cost has decreased and is likely to decrease even more in the future, the cost of API usage will likely be more than zero, which is not the case for running MaterialsBERT and the open-source models locally, since there are no associated API costs. Our discussions highlight this seemingly small but important difference, especially when the pipelines need to be run at scale. We have reported the price-per-token during the evaluations of this study in the Methods section, so in the future, astute readers can quickly estimate the adjusted costs.

Moreover, such pricing information would be an extremely important piece of historical record, that will provide the community with a sense of how things change over time. This, we think is invaluable (think Moore's Law, price/performance ratio of processors, etc., as a function of time).

17. The paper overall has potential to be good, but only after a substantially major revision. It seems that the authors rushed with its submission. I suggest thoroughly and diligently addressing the concerns expressed above as well as trying to elucidate on things that were not explained.

**Response:**

We thank the reviewer for the diligent review and raising important points, which have helped us significantly improve the quality and impact of the manuscript.

**Reviewer #2 (Remarks to the Author):**

This manuscript is about materials data extraction from polymer-related literature, using various large language models. The authors compared the performance and cost-effectiveness of GPT-3.5, LLaMa 2, MaterialsBERT, concluding that the performance of GPT-3.5 was the best but MaterialsBERT is useful in terms of cost-effectiveness. They collected more than million dataset consist of 24 properties from 106,000 unique polymers. The extracted data as well as model comparison information will be highly helpful for the researchers in this field. Therefore, the reviewer recommends to the publication of this work to "Communications Materials", after addressing the minor comments below.

1. As the authors stated, (for example, non-oriented vs. oriented PMMA layer) the materials property can vary with the synthesis condition, even the name of polymer is same. One can easily imagine that all the materials property will change with the molecular weight. But in this work, only 'material name' is extracted. At least molecular weight or morphology information should be collected.

**Response:**

This is an excellent point. Properties that depend only on the molecular structure of the polymers are mostly suitable for the current method. Many other properties such as the mechanical properties are heavily dependent on molecular weight, temperature, synthesis

and processing conditions, etc. It is crucial to extract the additional relevant information for downstream usage that pertain to these properties. Our pipelines currently do not extract such additional information. This is because such additional information is usually mentioned in a different paragraph (e.g., the experimental methods section) than where the properties are discussed in the text (e.g., the results section). As such, the pipeline that can correctly extract such additional information needs to be capable of understanding the context when a material or property is discussed in the article. This requires keeping track of each time a material is mentioned in different sections of the document, construct a knowledge graph like description based on such occurrences and prompt the language model to perform reasoning and deductive analysis to extract the correct data maintaining valid relationships.

We consider the inability to extract such additional information as a limitation of our current method. We briefly discussed this limitation in the second bullet point of the Outlook section. We have added the following text for further clarification.

“Our current pipelines extract data that is described completely in a specific paragraph by looking for all the required named entities (i.e., ‘material’, ‘property’, ‘value’ and ‘unit’) to establish correct relationships. However, the properties of polymers often rely on further information, such as molecular weights, temperature, synthesis and processing conditions, and morphology. This additional data also needs to be extracted from multiple paragraphs, while ensuring the preservation of valid relationships.”

Since the correct relationship extraction across multiple paragraphs is an important capability of a data extraction pipeline, we plan to address this in our upcoming works.

2. The authors collected the data from full text as well. In full text, there can be several property data for multiple materials, for example, the developed material vs. control group. Did authors consider which data belongs to which material? If so, how did it proceed?

**Response:**

Our pipelines create new data points if multiple properties are mentioned for the same materials. MaterialsBERT creates a new entry for the material name entity when multiple properties of the same material are found in the text. GPT-3.5 also generates multiple rows for each of the property values in the JSON output.

As for the cases of developed material vs. control group, MaterialsBERT cannot distinguish between such cases, which is one of its main limitations we discussed in the text. In contrast, GPT-3.5 performs satisfactorily when the control group is mentioned along with the material’s name. This works because materials are usually assigned different identifiers or prefixes when such differences are discussed in the full texts.

In the cases where the additional information is discussed in a different paragraph, or the same name is used for different materials groups, our pipelines cannot distinguish between the two groups for the reasons mentioned in reply to comment (1).

3. In introduction, these papers [Commun. Mater. 5, 13 (2024)], [J. Mater. Chem. A 11, 17628 (2023)] can be relevant.

**Response:**

We thank the reviewer for this suggestion to include these relevant papers. We have added discussion citing these papers in the Introduction section.

“In recent years, the transformer-based BERT architecture, with its ability to capture contextual and semantic relationships within scientific texts [10], has especially demonstrated superior performance compared to traditional neural network models [11].”

“These models have shown remarkable performance in handling various NLP tasks, showcasing their robustness and versatility [12], especially in high-performance text classification, named entity recognition, and extractive question answering with limited datasets [13].”

**Reviewer #3 (Remarks to the Author):**

Ramprasad et al. presented a well-designed study using different LLMs to extract 24 properties of over 106,000 polymers from a large number of literature sources. The data is analyzed and discussed in detail, and overall, I enjoyed reading this manuscript. Here are my comments:

1. The authors mentioned that the 681,000 documents for extraction were selected by searching the term "poly." It would be helpful for readers if you could provide more information when it is first mentioned (I personally do not see the details until reaching to the end of the article) and the distribution of the sources, as mentioned on page 21. Additionally, please clarify if these documents were automatically downloaded and, if so, which software was used. This information is crucial for reproducibility and understanding the data collection process.

**Response:**

We thank the reviewer for this comment and suggestion. The term “poly” was used as a simple keyword filter to quickly identify the articles that may contain results or discussions about polymers and eliminate the articles studying other classes of materials.

To compile the corpus, we first collected an index of materials science related journal articles from the Crossref database, then obtained permissions from the corresponding publishing houses via a licensing agreement of the Georgia Tech Library. Finally, we used in-house Python scripts to automatically download the corresponding HTML files from the journal websites or via API requests using instructions provided by the publishers. Additional details are available in one of our previous publications [Ref 14].

We have added texts discussing the source of the documents to the ‘Overview of the Data Extraction’ section of the manuscript.

“The articles were initially indexed through the Crossref database, followed by authorized downloads from 11 publishers, including Elsevier, Wiley, Springer Nature, American Chemical Society, and the Royal Society of Chemistry. Further details regarding the articles can be found in the Methods section and in Ref [14]. Specifically focusing on polymer-related content, we identified 681, 000 documents by searching for the term ‘poly’ in the title and abstract of the articles to disregard the articles reporting any other classes of materials.”

2. Furthermore, I am curious whether the inclusion of papers from other domains containing the term "poly" could influence the extracted data output.

**Response:**

That is correct; some articles contain the term 'poly' but are not primarily related to polymers. As a result, the pipelines, particularly the LLMs, also extract common properties like bandgap and hardness from other materials such as nanocrystals, inorganic materials, and nanoparticles. We excluded these materials during the post-processing stage and only kept those identified as polymers (or their known variants) for analysis.

The pipelines collect property values specific to a certain material, so the simple ‘poly’ filter doesn't adversely affect the extraction of data for that material. However, they do gather a lot of extraneous data for materials that are not polymers.

This is a good point; we have updated text in the ‘Data Extraction from Full Texts’ section to clarify this.

“In addition to the polymers, the pipelines also extracted property data for other classes of materials due to the use of the simple ‘poly’ keyword filter. To further identify the polymers, we checked if the extracted material is a valid polymer name by cross-referencing it with a comprehensive, albeit non-exhaustive, list of polymer names manually collected from the literature.”

3. It would be beneficial to discuss the selection criteria for the 24 properties and why there is an interest in extracting these specific properties. For example, in ref 24, the extracted data were used for predictive analysis. I believe it would be valuable to inform the reader not only about 'how' to extract data and how well the models perform but also what can be done with the extracted data, showcasing its potential applications.

**Response:**

We aimed to cover a broad range of property classes including thermal, mechanical, electronic, transport, and optical properties with the goal of training downstream ML models. These classes of properties are relevant for a wide variety of applications, ranging from structural to electronic to energy.

We have added text to the ‘Overview’ section discussing this point.

“Commonly reported thermal and optical properties were selected for their efficacy in training multi-task machine learning (ML) models, using highly correlated properties as substitutes for less prevalent ones. Additionally, we included properties that are beneficial for various polymer application areas. For instance, the bandgap and refractive index are vital for electrostatic energy storage and electronics, gas permeability properties are crucial for energy-efficient filtering and distillation applications, and mechanical properties are significant for numerous structural applications.”

We have previously implemented similar pipelines to optimize the power conversion efficiency of polymer solar cells and predict synthesis pathways of common polymers using data extracted from literature. We are currently using the recently extracted datasets in our ongoing downstream ML efforts. To showcase such potential applications, we have added the following sentences.

“The datasets we extracted from literature containing the 24 properties are integral for training downstream machine learning (ML) models. In a previous study, we employed similar datasets to optimize the material system and develop robust predictive models to optimize power conversion efficiency of polymer solar cells and demonstrated a significant reduction in new polymer discovery time [15]. In another study, we assembled a comparable dataset to predict the retrosynthesis pathways for a target polymer [16].”

4. Based on Figure 1, it appears that there is already a well-designed NLP-based workflow, with the last step allowing for a choice between NER-based extraction and LLM-based extraction. Has this workflow (from "full text corpus" to "NER filter") been reported before? In my opinion, one advantage of LLM-based extraction is its ease of use and reduced requirement for coding/NLP knowledge, making it more accessible to the general public. However, mixing traditional NLP workflow and LLM-based extraction may not necessarily highlight this advantage, as a majority of the framework still relies on traditional NLP (and the researchers who are not familiar with these techniques need to invest time on learning). While this is acceptable, I suggest commenting on future perspectives and whether this pipeline could be entirely governed by LLMs alone.

**Response:**

The reviewer makes a very insightful and forward-looking point. Indeed, the primary advantages of the LLMs are their versatility and ease of use. Based on our experience with various LLMs, we have found that bigger models are more effective at natural language reasoning tasks. This is crucial for understanding the complex relationships between different entities in the scientific texts. However, existing pre-trained or instruction-tuned LLMs, are general-purpose and cannot accurately extract scientific data on their own. Often, prompting for data extraction results in increased hallucinations and irrelevant or inconsistent results. The computational or API costs for running the large models are other major issues, for which smaller models are usually better. We used multi-stage filtering and few-shot techniques to address these issues of the LLMs. The “full text corpus” to “NER filter” parts of our workflow, or variations of it, have been used in NER-based extraction

pipelines before (e.g., Ref. [2], [3], [4], [15] employ similar workflow). But to our knowledge, they were not used with LLMs before.

We have added the following text to the “Outlook” section of the manuscript discussing our future perspective on this.

“Future works may involve development of models that can concurrently handle searching, filtering, NER and data extraction tasks. As advancements in smaller open-source models persist, we anticipate the potential to substitute our entire workflow with a cost-efficient, single, fine-tuned, open-source LLM. Such a system could democratize the process of extracting data at-scale from the literature. However, this transition would require comprehensive analysis and validation of all parts of the protocol by the materials science community.”

5. How was the 1,000 subset for evaluation selected? Given that this subset consists of only a small portion of the entire dataset, it would be helpful to discuss how the evaluation based on such a small set can be representative of the whole dataset.

**Response:**

We believe that the small subset is representative of the full corpus, because the subset contains papers randomly selected from 8 out of the 11 publishers. We have added the following figure in the SI to show the distribution of the papers in the subset compared to the total number of papers in the entire corpus (see Fig. S1).

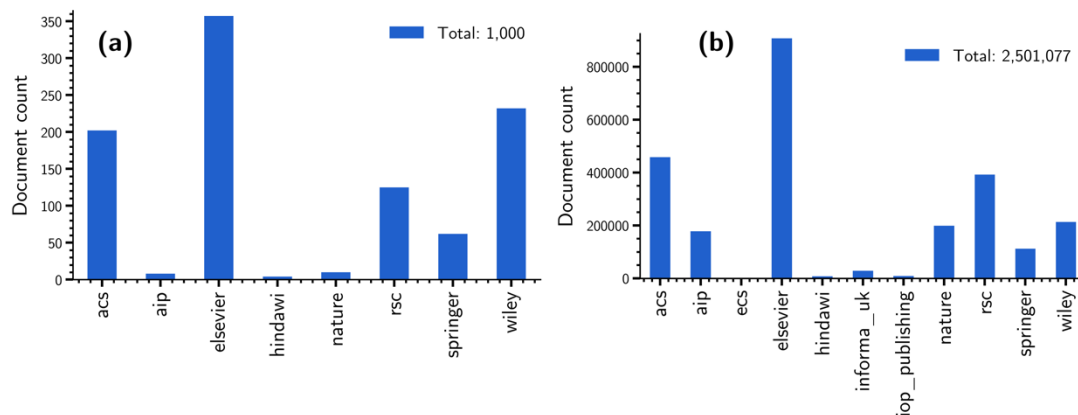


FIG. S1. (a) Number of documents randomly selected from different publishers for performance benchmarking using a subset of the full corpus. (b) Total number of documents available in the corpus from different publishers. The subset represents journal articles from the eight most common publishing houses.

Additionally, to address this question, we have added the following text to the manuscript.

“Consequently, the evaluation involves a subset of 1,000 articles from the larger pool of the 681,000 polymer-related papers. In this subset, we have manually curated data from the abstracts of 630 articles, which reported one or more T<sub>g</sub> and bandgap values in their

abstracts and were selected randomly. The rest of the 370 articles were randomly chosen from the polymer papers. Bar charts containing the distribution of the selected articles compared to the full corpus are shown in Fig. S1.”

6. In the data availability section, it would be beneficial to provide a link to the prompt used and the code/ API to run the LLMs like LLaMa 2 and GPT, allowing readers to reproduce the results. Alternatively, the author could include these in the supporting information file or create a dedicated GitHub repository for this purpose.

**Response:**

We made the codes and model publicly available in GitHub and HuggingFace Hub, so interested readers can run the pipelines and reproduce our results. The prompt we mentioned in Fig. 2(a-b) is used for all the properties.

Please see <https://github.com/Ramprasad-Group/PromptDataExtraction> for the codes and <https://huggingface.co/pranav-s/MaterialsBERT> for the NER model.

We have included this information to the ‘Data Availability’ and the ‘Code Availability’ sections of the manuscript.

7. In the supporting information, it is difficult to read the exact false positive numbers in Fig. S1. I recommend labeling the y-axis values on each panel or providing a separate table listing the exact values to improve readability and interpretation.

**Response:**

Thank you for this suggestion. We have included Table S1 in the SI to separately report the related metrics.

TABLE S1. Detailed metrics for Tg and bandgap data shown in Fig. S3.

| Entity class   | Property | True positives | False positives | False negatives |
|----------------|----------|----------------|-----------------|-----------------|
| Material Name  | Tg       | 546            | 211             | 137             |
| Property Value | Tg       | 543            | 214             | 69              |
| Name and Value | Tg       | 441            | 316             | 157             |
| Material Name  | Bandgap  | 302            | 45              | 64              |
| Property Value | Bandgap  | 311            | 36              | 47              |
| Name and Value | Bandgap  | 285            | 62              | 71              |

## References:

- [1] P. Shetty *et al.*, “A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing,” *Npj Comput. Mater.*, vol. 9, no. 1, Art. no. 1, Apr. 2023, doi: 10.1038/s41524-023-01003-w.
- [2] L. Foppiano, P. B. Castro, P. Ortiz Suarez, K. Terashima, Y. Takano, and M. Ishii, “Automatic extraction of materials and properties from superconductors scientific literature,” *Sci. Technol. Adv. Mater. Methods*, vol. 3, no. 1, p. 2153633, Dec. 2023, doi: 10.1080/27660400.2022.2153633.
- [3] L. P. J. Gilligan, M. Cobelli, V. Taufour, and S. Sanvito, “A rule-free workflow for the automated generation of databases from scientific literature,” *Npj Comput. Mater.*, vol. 9, no. 1, Art. no. 1, Dec. 2023, doi: 10.1038/s41524-023-01171-9.
- [4] J. Dagdelen *et al.*, “Structured information extraction from scientific text with large language models,” *Nat. Commun.*, vol. 15, no. 1, Art. no. 1, Feb. 2024, doi: 10.1038/s41467-024-45563-x.
- [5] K. Hira, M. Zaki, D. Sheth, Mausam, and N. M. Anoop Krishnan, “Reconstructing the materials tetrahedron: challenges in materials information extraction,” *Digit. Discov.*, vol. 3, no. 5, pp. 1021–1037, 2024, doi: 10.1039/D4DD00032C.
- [6] G. H. Yi *et al.*, “MaTableGPT: GPT-based Table Data Extractor from Materials Science Literature.” 2024.
- [7] H. Oka, A. Yoshizawa, H. Shindo, Y. Matsumoto, and M. Ishii, “Machine extraction of polymer data from tables using XML versions of scientific articles,” *Sci. Technol. Adv. Mater. Methods*, vol. 1, no. 1, pp. 12–23, Jan. 2021, doi: 10.1080/27660400.2021.1899456.
- [8] T. Gupta, M. Zaki, D. Khatsuriya, K. Hira, N. M. A. Krishnan, and Mausam, “DiSCoMaT: Distantly Supervised Composition Extraction from Tables in Materials Science Articles,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 13465–13483. doi: 10.18653/v1/2023.acl-long.753.
- [9] M. Zhu and J. M. Cole, “PDFDataExtractor: A Tool for Reading Scientific Text and Interpreting Metadata from the Typeset Literature in the Portable Document Format,” *J. Chem. Inf. Model.*, vol. 62, no. 7, pp. 1633–1643, Apr. 2022, doi: 10.1021/acs.jcim.1c01198.
- [10] J. Choi *et al.*, “Deep learning of electrochemical CO<sub>2</sub> conversion literature reveals research trends and directions,” *J. Mater. Chem. A*, vol. 11, no. 33, pp. 17628–17643, Aug. 2023, doi: 10.1039/D3TA02780E.
- [11] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A Pretrained Language Model for Scientific Text.” arXiv, Sep. 10, 2019. doi: 10.48550/arXiv.1903.10676.
- [12] S. Lappin, “Assessing the Strengths and Weaknesses of Large Language Models,” *J. Log. Lang. Inf.*, Nov. 2023, doi: 10.1007/s10849-023-09409-x.
- [13] J. Choi and B. Lee, “Accelerating materials language processing with large language models,” *Commun. Mater.*, vol. 5, no. 1, pp. 1–11, Feb. 2024, doi: 10.1038/s43246-024-00449-9.
- [14] P. Shetty and R. Ramprasad, “Automated knowledge extraction from polymer literature using natural language processing,” *iScience*, vol. 24, no. 1, p. 101922, Jan. 2021, doi: 10.1016/j.isci.2020.101922.

- [15] P. Shetty, A. Adeboye, S. Gupta, C. Zhang, and R. Ramprasad, “Accelerating materials discovery for polymer solar cells: Data-driven insights enabled by natural language processing.” arXiv, Feb. 29, 2024. doi: 10.48550/arXiv.2402.19462.
- [16] L. Chen, J. Kern, J. P. Lightstone, and R. Ramprasad, “Data-assisted polymer retrosynthesis planning,” *Appl. Phys. Rev.*, vol. 8, no. 3, p. 031405, Jul. 2021, doi: 10.1063/5.0052962.

**Manuscript COMMSMAT-24-0322-T**  
**Response to Reviewers**

Dear Editor and Reviewers,

Thank you for your comments and feedback to the responses about our manuscript titled “*Data Extraction from Polymer Literature using Large Language Models*”. We have added additional details regarding the remaining questions asked by the reviewers. Our detailed reply to each of the reviewers’ comments, along with the corresponding modifications made to the manuscript, are outlined below.

**Reviewer #1 (Remarks to the Author):**

I appreciate the author's thorough response to my review. While many of my comments were addressed appropriately, there are still a few points that are problematic and have not been addressed properly - particularly regarding the accuracy assessment in response 4).

1. I understand that the  $<10\%$  overlap is low due to differences in the extracted names of the materials. Even if they are technically equivalent, they may not be literally the same. While not unifying them is a deficiency of the presented results (because very straightforward NLP techniques exist for comparing two non-literally identical phrases to determine whether they have the same meaning), it is not that important if it is explained, which the authors did. That being said, the presented values still contradict the previously presented results. Even disregarding the overlap argument, the pure number of data points (13,929 for BERT and 88,128 for GPT) does not mathematically allow for one of the two methods to have a high F1 score. A discrepancy in the number of data points that high must mean that at least one of the methods has either horrible precision or horrible recall, which should be obvious. Therefore, the F1 score - a direct consequence of precision and recall - cannot be high. The lowest F1 score mentioned in the paper seems to be 0.66 for BERT and 0.87 for GPT. A back-of-the-envelope estimate for the maximum F1, assuming perfect recall for GPT, suggests that the maximum F1 for BERT in that case is around 0.15 - based solely on the significant difference in the amount of extracted data between the two methods, which is a best case scenario. This should be obvious. The authors should revisit all of their accuracy estimates before the publishing process continues. If, however, I am mistaken in my reasoning (which, to my understanding, I am not), that means the text and figure description of the accuracy assessments is unclear or insufficient and should be thoroughly revised.

**Response:**

We have attempted to address this very pertinent point in the revised manuscript. It is likely that there may have been some misunderstandings regarding the F<sub>1</sub> score calculation method and we admit that we may have not been clear enough on the definitions that have gone into the F<sub>1</sub> score calculations. Below, we attempt to clarify and address these potential sources of confusion.

First, we note that, at present, no standard metric exists for measuring the performance of complex data extraction tasks undertaken here, similar to standard metrics such as F<sub>1</sub>

score, RMSE and  $R^2$  appropriate for more straightforward or common classification and regression tasks. The present work on data extraction poses unique challenges to specifying performance metrics, as the extracted information contains many parts, and success can be measured in many ways. In this work, while we have used the  $F_1$  score metric to measure our data extraction performance, we have made several assumptions to better fit this performance metric to our specific requirements. Therefore, some results not expected with traditional application of the  $F_1$  metric for standard classification tasks should be expected here.

To describe the performance metric used here in a more step by step detail, we have added the following additional explanation to the “ $F_1$  score calculation method” section of the SI. We hope that this clarification is helpful, and will spawn other more appropriate approaches to quantifying performance for complex data extraction such as undertaken here.

“To further illustrate the  $F_1$  score calculation method, we consider an example scenario, where we extract bandgap data from the abstract of Ref. [1] for which we have manually curated data available. After processing through the GPT-3.5 and MaterialsBERT pipelines, we obtain the following set of data points, shown in **Table S1**. The decision on data quality determined using the ground truth values is given in the ‘Result’ column.

**Table. S1** Datapoints extracted from Ref. [1] for bandgap by GPT-3.5 and MaterialsBERT. The star (\*) symbol indicates a material not extracted by the pipeline.

| No. | Pipeline      | Material | Property        | Value   | Result |
|-----|---------------|----------|-----------------|---------|--------|
| 1   | GPT-3.5       | P-diF2T  | bandgap         | 1.83 eV | TP     |
| 2   | GPT-3.5       | P-TT     | bandgap         | 1.85 eV | TP     |
| 3   | GPT-3.5       | P-BDT    | bandgap         | 1.90 eV | TP     |
| 4   | GPT-3.5       | P-NDT    | bandgap         | 2.10 eV | FP     |
| 5   | MaterialsBERT | P-diF2 T | optical bandgap | 1.83 eV | TP     |
| 6   | MaterialsBERT | TT       | optical bandgap | 1.85 eV | TP     |
| 7   | MaterialsBERT | P-diF2 T | optical bandgap | 1.90 eV | FP     |
| 8   | MaterialsBERT | P-diF2 T | optical bandgap | 2.04 eV | FP     |
| 9   | MaterialsBERT | * P-BDT  | -               | -       | FN     |
| 10  | MaterialsBERT | * P-NDT  | -               | -       | FN     |

As discussed in the manuscript, there are differences in the extracted material names and the property names by the two pipelines (e.g., note the presence of a space in ‘P-diF2 T’ in Rows 5, 7 and 8). We normalize such names to identify the true positive cases. Data for ‘P-BDT’ and ‘P-NDT’ was not extracted by MaterialsBERT and were counted as false negatives (FN). Additionally, we treat datapoint 4 as false positive (FP) because the material name was correctly extracted but the value was hallucinated by GPT-3.5.

Proceeding to the  $F_1$  score calculation, we obtain the following numbers shown in **Table S2**.

**Table. S2** Metrics obtained for the datapoints extracted in Table S1.

| Pipeline      | TP | FP | FN | precision | recall | $F_1$ score | Unique Materials |
|---------------|----|----|----|-----------|--------|-------------|------------------|
| GPT-3.5       | 3  | 1  | 0  | 0.75      | 1      | 0.86        | 4                |
| MaterialsBERT | 2  | 2  | 2  | 0.5       | 0.5    | 0.50        | 2                |

The Venn diagrams shown in Fig. 4, which highlight the differences in the material names extracted by the pipelines, were drawn using the number of unique paragraphs and materials obtained from the Tg and bandgap data points extracted from all papers (and not just the papers and paragraphs used in determining the  $F_1$  scores).”

If we construct Venn diagrams using the material names specific to each pipeline from the example data shown in Table S1, we obtain the following diagrams.

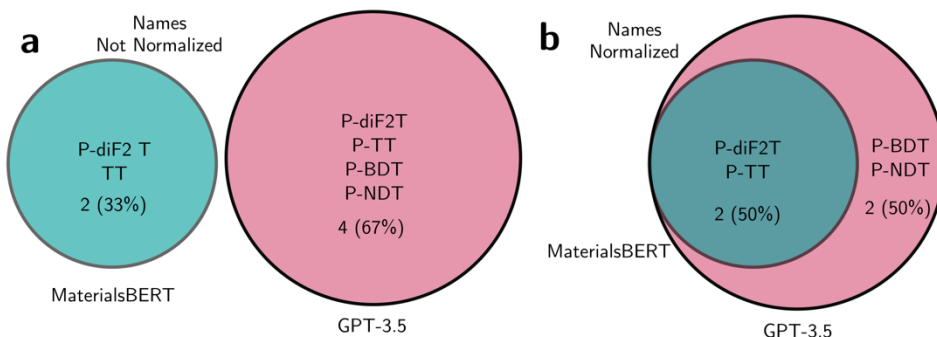


**Figure R1.** Venn diagram for the example data extraction shown in Table S1 and Table S2 (a) without normalizing the material names and (b) after normalizing the material names.

We find that, even if there is no overlap between the material names, the  $F_1$  score obtained for MaterialsBERT and GPT-3.5 pipelines are 0.5 and 0.86, respectively. These values are very close to what is shown in Fig. 4d of the manuscript. We also note that, in Table S2, the number of unique materials is already two times higher for the GPT pipeline and it is independent of the number of false negatives. Data extraction for composites and blends from the full texts further increase the number of materials names extracted by the GPT pipeline (discussed in the manuscript).

Thus, this example scenario shows that the pipelines can have high  $F_1$  scores even if there are no overlap between the material names in the Venn diagram.

The diagram shown in Fig. 4d is calculated for all texts, while we calculated  $F_1$  scores on the abstracts of 630 papers. We added the Venn diagrams for Tg and bandgap for the 630

abstracts in the SI to provide a better picture. Additionally, we have added detailed precision, recall and  $F_1$  score for Tg and bandgap to Table S3.

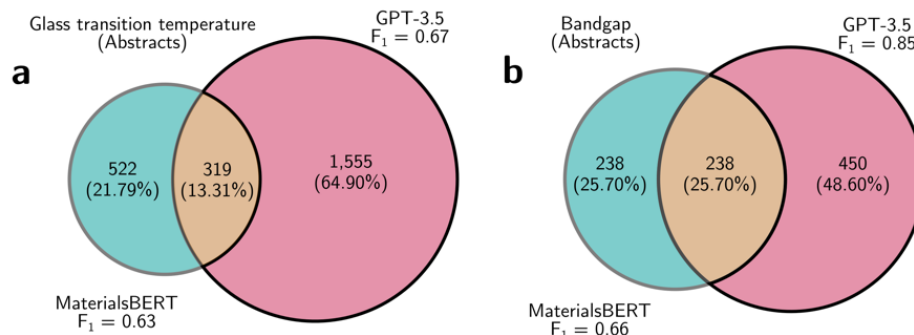


Figure S2. Number of unique material names extracted for (a) Tg and (b) bandgap using the Materials and GPT-3.5 pipelines from the abstracts of 630 selected papers.

Following reference to this figure has been added to the main text while discussing Fig. 4d.

“For comparison, similar Venn diagrams for Tg and bandgap data extracted from the 630 abstracts are shown in **Fig. S1**.”

To summarize, the apparent discrepancies between the  $F_1$  scores reported in the manuscript and the Venn diagram shown in Fig. 4d are due to the following reasons:

1. During the  $F_1$  score calculation, we have normalized the material names using the available curated ground truth dataset, while in the Venn diagram shown in Fig. 4d, we have kept the original (unnormalized) material names. This is because the GPT-3.5 pipeline extracts additional modifiers (e.g. ‘vertical’, ‘horizontal’, ‘blend’ etc) along with the material names (discussed in the manuscript) and normalization of these names are not straightforward.
2. Venn diagrams are constructed of sets, which are mathematically defined as collections of well-defined and distinct objects. Thus, the number of material names shown in the diagram in Fig. 4d should not be used to infer about the  $F_1$  score, as Venn diagrams show the relationship between objects, while  $F_1$  score calculation requires precise values of the objects. The main purpose of Fig. 4d was to highlight the difference between the pipelines.

### Reviewer #3 (Remarks to the Author):

I have recently reviewed this manuscript again and have read through the author's response. I believe they have done a good job addressing most of the concerns I raised in the last round. Therefore, I think this paper can be accepted.

1. One minor point that I feel is not adequately addressed is my comment #3. This comment is more of an advice rather than a strict requirement. Specifically, I mentioned that the author should better explain how their work aims to train downstream ML models. While the author has provided some examples from their previous work, a more systematic approach would be beneficial. I suggest including a table that outlines specific properties and examples, along with relevant work. This table can be placed in the supplementary information for readers.

**Response:**

Thank you for reviewing our manuscript again and for your positive feedback. We appreciate your suggestions for a clearer explanation of how our work can lead to training downstream ML models.

In our previous response, we briefly mentioned the use of extracted datasets for training ML models, citing examples from our work on polymer solar cells and retrosynthesis pathways. The polymer solar cell work (Ref. [2] cited below) has very recently been published, clearly demonstrating the possibilities for downstream ML model training, which can accelerate materials discovery. This work is now cited in our modified manuscript as Ref. 39.

Several other works related to property predictions are underway that uses our automated extraction pipelines. Since, these activities are still in progress, and associated challenges for making the extracted data truly informatics-ready are ongoing, we hope to share such advancements with the community in a future contribution.

**References:**

- [1] M. He *et al.*, “Wide bandgap donor-acceptor conjugated polymers with alkylthiophene as side chains for high-performance non-fullerene polymer solar cells,” *Org. Electron.*, vol. 65, pp. 31–38, Feb. 2019, doi: 10.1016/j.orgel.2018.10.034.
- [2] P. Shetty, A. Adeboye, S. Gupta, C. Zhang, and R. Ramprasad, “Accelerating Materials Discovery for Polymer Solar Cells: Data-Driven Insights Enabled by Natural Language Processing,” *Chem. Mater.*, vol. 36, no. 16, pp. 7676–7689, Aug. 2024, doi: 10.1021/acs.chemmater.4c00709.