

Toward Reliable Ad-hoc Scientific Information Extraction: A Case Study on Two Materials Datasets

Satanu Ghosh¹ Neal R. Brodnik² Carolina Frey² Collin Holgate²
Tresa M. Pollock² Samantha Daly² Samuel Carton¹

¹University of New Hampshire

²University of California, Santa Barbara

{satanu.ghosh, samuel.carton}@unh.edu

{nrbodnik, cfrey, holgate, tresap, samdaly}@ucsb.edu

Abstract

We explore the ability of GPT-4 to perform ad-hoc schema based information extraction from scientific literature. We assess specifically whether it can, with a basic prompting approach, replicate two existing material science datasets, given the manuscripts from which they were originally manually extracted. We employ materials scientists to perform a detailed manual error analysis to assess where the model struggles to faithfully extract the desired information, and draw on their insights to suggest research directions to address this broadly important task.

1 Introduction

A key use case for large language models in science is **ad-hoc schema-based information extraction**. In this scenario, a scientist has a specific information need, such as the compressive yield strength of known multi-principal element alloys (MPEAs), and wants to (1) identify papers containing relevant information instances, and (2) comprehensively extract these instances according to a desired schema, which might include secondary information such as testing conditions and synthesis details (Fig 1).

Such datasets are a key element of scientific informatics, transforming information reported inconsistently throughout the literature into a structured format appropriate for use in machine learning. In materials science, such datasets are used in a wide range of applications, including model-driven materials discovery, where they can be used to train or warm-start surrogate models of material properties which are then incorporated into various kinds of discovery algorithms (Pilania, 2021).

This is a challenging extraction task. The materials science literature contains a plethora of experimental results, often spanning multiple decades, in a multitude of different formats with varying physical units. Prior generations of models needed large amounts of data to fine-tune, and were brittle and

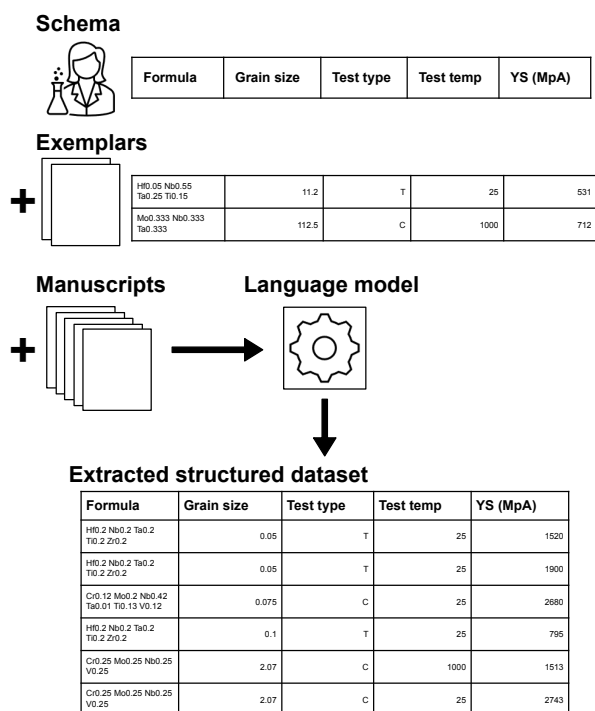


Figure 1: Extracting large-scale structured data from scientific literature should be as simple as specifying a schema, a corpus of manuscripts, and a few exemplars, and letting the LLM perform the extraction.

non-transferable, limiting their practicality for this type of ad-hoc, one-off extraction task. To date, manual annotation is still the standard approach for extraction of this type of unstructured data (Xu et al., 2023b).

Contemporary large language models such as GPT-4 (OpenAI, 2023) and Gemini (Team et al., 2023) have the potential to conquer this barrier. With the emergent few-shot capabilities of these models, a scientist can potentially present a LLM with a (1) collection of manuscripts, (2) a schema defining what types of data they want to extract, and (3) a few exemplars of performing this extraction, and receive back a tabular dataset of information extracted from those manuscripts, ready to use for

their chosen purpose. This ad-hoc paradigm has been referred to as “on-demand” information extraction (ODIE) (Jiao et al., 2023) and it promises to speed up the search for new materials with unique properties.

We argue that this very ambitious extraction setting should be a key target for LLM-driven scientific information extraction, as it can potentially allow scientists to extract a maximum of clinically relevant data with a minimum of effort, both in materials science and other domains such as biomedicine. Two significant questions, then, are **(1) to what extent can contemporary LLMs perform this type of ad-hoc extraction on scientific text, and (2) what are the major barriers impeding their effectiveness?**

In this paper, we explore these questions using two manually-extracted material property datasets, one focused on multi-principal element alloys (MPEAs), sometimes known as high-entropy alloys (HEAs) (Borg et al., 2020), and one focused on elemental diffusion in silicate melts (Zhang et al., 2010). We use these existing datasets as sources of schema structure and exemplars, and evaluate whether a representative contemporary LLM (GPT-4) is capable of replicating them given the manuscripts from which they were originally extracted. We perform a detailed error analysis, including annotation by domain experts, to understand where and why the model falls short, to guide further development of LLMs to advance extraction paradigms for materials data.

Generally, we find that the model shows great potential for extracting information described narratively or in the form of a conventionally-formatted table. The majority of errors are attributable to figures, which our current pipeline does not address, and PDF parsing issues. Other sources of error include non-standard table formats, the need for additional postprocessing of extracted values, as well as true reading comprehension errors which could be addressed via improved prompt engineering. All relevant code and data can be found in our github repository ¹.

2 Related work

Information extraction (IE) is a core NLP task with a long history, and one of many which contemporary large language models (LLMs) such as GPT-4

(OpenAI, 2023) and Gemini (Team et al., 2023) are able to perform in a zero- or few-shot setting (Xu et al., 2023a). This few-shot extraction ability has enabled a new extraction paradigm, involving extracting a complex schema with little or no training data, termed “on-demand information extraction” (ODIE) by (Jiao et al., 2023). Our task setting falls into this paradigm.

Zero- and few-shot scientific information extraction forms a major aspect of the excitement over LLMs’ potential for accelerating scientific discovery, because of its potential to create structured information from diffuse unstructured data sources such as scientific papers and clinical notes (Hope et al., 2022; Morris, 2023). Ad-hoc extraction from scientific literature is particularly challenging because the amount of manually-annotated data needed to fine-tune a model to perform a particular extraction task would constitute a significant fraction of all the data in existence, obviating the benefit of training such a model. Scientific literature is also highly multimodal, with key information presented in narrative, tabular, and visual formats. SciDaSynth (Wang et al., 2024) approaches the problem with an interactive QA-based interface, evaluating based on a 12-scientist user study.

LLM-powered IE has attracted attention in materials science specifically (Olivetti et al., 2020; Xie et al., 2023). (Polak et al., 2024) explore the use of GPT-4 to extract bulk modulus information from sentences extracted from scientific papers, while (Yang et al., 2023) uses zero-shot ChatGPT with a verification step to extract band gap information from sentences collected in a dataset by (Dong and Cole, 2022). (Walker et al., 2023) fine-tunes GPT-3.5 to extract growth procedures and outcomes for gold nanorods from paper text. (Montanelli et al., 2024) uses the Cohere model to extract phase-property relationships from a corpus of full-text materials papers. Finally, the A-Lab (Szymanski et al., 2023) incorporates extracted synthesis recipes into an AI-guided experimentation protocol. Our work differs from these recent approaches in seeking to extract a rich, complex schema from non-curated data in one pass with no fine-tuning, instead of narrowly-defined tasks over carefully-curated corpora. This is a highly ambitious and difficult extraction task, but one that places a minimum of burden on the scientist, and we see it as a key milestone for scientific IE.

¹https://github.com/SatanuG/ad_hoc_information_extraction

| | | MPEA | Diffusion |
|------------------|----------|-----------|-----------|
| Total | # Papers | 164 | 71 |
| | # Rows | 1211 | 2838 |
| Retrieved | # Papers | 128 | 55 |
| | # Rows | 971 | 2359 |
| Publication year | | 2004-2022 | 1964-2009 |

Table 1: Basic statistics for the two datasets, publication year range. For both datasets, a subset of total papers were retrieved, parsed and analyzed in this work.

3 Datasets

We experiment with two material properties datasets, one pertaining to multi-principal element alloys (MPEAs) (Borg et al., 2020), and one to elemental diffusion in silicate melts (Zhang et al., 2010). Table 1 reports basic statistics of each dataset, including the number of unique papers, number of unique records (since a single paper can yield multiple records), and publication year range. Table 4 describes the individual columns in each dataset.

MPEA Mechanical Properties Dataset Multi-principal element alloys (MPEAs) are a new class of chemically complex metallic alloys of interest for applications in energy, defense, and transportation, distinguished from conventional alloys by having relatively even proportions of many elements (Miracle and Senkov, 2017; Miracle et al., 2020). The MPEA dataset used for this analysis specifically aims to record key properties reported for MPEAs derived primarily from the refractory metals (Cr, Hf, Mo, Nb, Ru, Ta, Ti, V, W, Zr). These properties include yield strength, which represents the force necessary to make a metal permanently deform, and elongation, which represents the degree to which a material can deform prior to failure. Hardness, which represents the ability of the material resist localized deformation, can be used to estimate the yield strength of a material. Other useful details recorded in the MPEA dataset where available include the concentration of interstitial elements present (i.e. oxygen, nitrogen, and carbon concentrations), the method by which the material was processed, the test temperature, and the method of testing.

The MPEA dataset used in this study is an expansion of the dataset published in (Borg et al., 2020), which has been continually updated. Records

were taken from experimental literature published in 2004-2022 and are identified by a Digital Object Identifier (doi) that was assigned by the publisher.

Diffusion in Silicate Melts Dataset Our second dataset (Zhang et al., 2010) covers the diffusion (transport of elements) inside silicate glasses and melts (e.g., magma), which affects many processes in geology and materials science. The key value in this dataset is *diffusion coefficient*, which is often symbolized as D and has units of m^2/s . The diffusion coefficient is a proportionality constant that describes how rapidly one element is transported within a substance. As one may expect, the value of D depends on what the diffusing element is, the medium (magma) it is moving in (i.e., its composition), and also environmental conditions like temperature, pressure, and how much water is present (which greatly affects magmatic structure). The diversity of magmas found in the earth and the elements being transported within them motivates building and maintaining databases of diffusion coefficients to help elucidate underlying truths.

Outside of its fundamental value, we chose this dataset for our study because (1) diffusion as a process is heavily sensitive to experimental conditions, meaning value extraction is only useful with the associated large quantity of metadata; (2) it is a large dataset, affording sufficient measurements and papers to draw meaningful conclusions; (3) the multi-decade span of publication dates offers exposure to a range of scientific conventions and pdf quality levels from the perspective of typesetting, character recognition, and table format; and (4) it represents a field where updates occur on a decadal basis, thus valuing a tool for automatic extraction of new measurements.

4 Method

Our extraction pipeline consists of the following steps: (1) retrieving and parsing source PDFs; (2) prompting and response postprocessing; and (3) extracted row alignment.

4.1 Retrieving and parsing source PDFs

The vast majority of PDFs available in both datasets are not open-access, and are referred to only by their DOI within each dataset. As a first stage, we manually download as many PDFs as we can easily access, resulting in 128 PDFs of the original 164 for the MPEA dataset, and 55 out of 71 for the diffusion dataset. Our github repository includes

the DOIs of the PDFs we downloaded but not, for copyright reasons, the PDFs of these papers.

As of June 2024, none of the major LLM APIs allow programmatic uploading of PDF files. Therefore, to process PDFs at scale, they must be converted to an intermediate readable format. We use GROBID² to convert paper PDFs to XMLs. When parsed correctly, the XML contains all textual information from the paper, including tables and their respective captions. However, GROBID cannot process figures, leaving any figure-bound information inaccessible by definition. The error analysis presented in Section 6.2 assesses the magnitude of this problem.

4.2 Prompting

We experiment with three basic prompting techniques: zero- and one-shot prompting, and LangChain. For each technique, we try both a “complex” schema including all columns from the original dataset, and a reduced “simple” schema including only a subset of column.

Role Setting:

```
I am a helpful assistant capable of extracting information from text. I will not generate any new tokens... return the information in format of the schema.
```

Exemplar:

```
{{Prompt instructions}}
{{Exemplar XML}}
{{Exemplar output JSON}}
```

Prompt:

```
Extract all information relating to every High Entropy Alloys (HEA) from the following text:
{{Paper XML}} using the following schema:
{{JSON schema}}. Return a list of schemas in JSON format for every unique compound.
```

```
Additional context: HEA composition sometimes have variable ratio... not available then return 'No information'.
```

```
Restriction: Do not extract any tokens if a HEA... not available then return 'No information'. Do not violate the schema.
```

Figure 2: Abbreviated one-shot prompt. The prompt begins with role-setting, includes a single exemplar with prompt instructions, then repeated prompt instructions with additional context and clarifications.

Zero- and one-shot prompting Working with the gpt-4-1106-preview GPT-4 Turbo model variant via the OpenAI API, we use a very basic

²<https://github.com/kermitt2/grobid>

and general prompting approach, illustrated in Fig. 2. Our goal is to extract all columns from a single full paper XML at once, requesting them in the form of a JSON. In the one-shot condition, the prompt includes a single exemplar, selected randomly to be (Long et al., 2019) and (Rüssel and Wiedenroth, 2004) for the MPEA and diffusion dataset respectively.

We additionally make use both role-setting and task instructions in our prompt. Role-setting has been found to be generally helpful in prompt engineering (Shanahan et al., 2023; Kong et al., 2023), so we set various desired aspects of the LLM’s behavior via this mechanism, i.e., extracting information from the given text without generating new tokens and following the provided schema (if any). The task instructions are specific and detailed, customized to each dataset based on expert input and include a schema that should be followed. We add additional context with background knowledge and special instructions. A more complete version of the prompt is available in the appendix (Section A.2).

LangChain LangChain³ is a popular open-source framework for developing solutions that combine LLMs with other tools. Information extraction is one of the many use cases that can be performed with LangChain. We follow the guidelines on their website to perform structured information extraction for long text. In short, the LLM is provided with a schema of various entities and a short description of each entity. In our use case, these entities are the column headers in the two datasets. An important distinction of this method from basic prompting methods was *chunking the papers into smaller blocks of 2000 tokens*, with each block having a 1% overlap with the previous one.

Postprocessing Minimal postprocessing is needed for the model’s JSON output, which varies slightly in format from run to run. This includes wrapping JSON-formatted output in ‘‘ ‘json’ tags. The model also sometimes produces Javascript-style ‘//’ comments explaining its extraction decisions; therefore, we use the `jstyleson` Python library⁴ to parse this JSON output rather than the default Python `json` library.

³<https://python.langchain.com/v0.2/docs/introduction/>

⁴<https://github.com/linjackson78/jstyleson>

| Dataset | Method | Matched | Recall | Hallucinated | Precision |
|-----------|-------------------|---------|--------|--------------|-----------|
| MPEA | 0-shot-simple | 56 | 0.062 | 166 | 0.245 |
| | 0-shot-complex | 56 | 0.066 | 235 | 0.369 |
| | 1-shot-simple | 243 | 0.280 | 259 | 0.502 |
| | 1-shot-complex | 174 | 0.202 | 207 | 0.475 |
| | langchain-simple | 188 | 0.253 | 746 | 0.219 |
| | langchain-complex | 129 | 0.182 | 564 | 0.213 |
| Diffusion | 0-shot-simple | 0 | 0 | 93 | 0 |
| | 0-shot-complex | 0 | 0 | 126 | 0 |
| | 1-shot-simple | 2 | 0.001 | 52 | 0.037 |
| | 1-shot-complex | 23 | 0.013 | 193 | 0.106 |
| | langchain-simple | 70 | 0.032 | 1606 | 0.060 |
| | langchain-complex | 47 | 0.022 | 1423 | 0.048 |

Table 2: Extraction performance based on match, missed, and hallucinated values on different variations of dataset and methods.

4.3 Extracted row alignment

In order to evaluate our success in replicating the original datasets, including which rows were matched, which original rows were missed, and which extracted rows were hallucinated, we first need to determine which extracted row corresponds with which original dataset row. This is nontrivial, because the columns that define a unique record will vary from paper to paper. One paper might test multiple different compositions, while another might test the same composition under different testing conditions, while a third might perform multiple trials of the same experiment. Given that the model will miss or hallucinate many individual cell values, this becomes a nuanced matching task.

Our approach is to combine domain-specific minimum criteria for a possible match with a greedy matching process within each paper. An extracted row is a potential match for an original row if it shares values for hand-chosen required columns, which we chose to be **formula and yield strength** for the MPEA dataset and **diffusing element and diffusion coefficient** for the diffusion dataset. Beyond this minimum, match strength is determined by proportion of shared column values between the two rows, after which rows are greedily paired with each other until no possible pairs remain for a given paper. Paired rows are classified as a ‘match’, unpaired rows from the original dataset are classified as a ‘miss’, and unpaired extracted rows are classified as a ‘hallucination’.

5 Automated evaluation

After aligning extracted rows with dataset rows as described above, we report the number of matches and hallucinations in Table 2 along with the recall and precision. For both the datasets, two variations of schema and three types of prompting.

The only difference between simple and complex schema is the number of properties mentioned in each schema.

The raw performance is underwhelming. On MPEA, the model hallucinates roughly as many rows as it matches, and misses twice as many. We are able to match significantly more entries using the simple schema than the complex schema. While using Langchain we extracted more relevant entries, but we can see a drop of precision (more hallucination). On the Diffusion dataset, we are able to match no rows at all for 0-shot and, though the model hallucinates at a proportionally lower rate than for the MPEA dataset. Using Langchain clearly improved the performance on this dataset but the hallucinations also increased significantly.

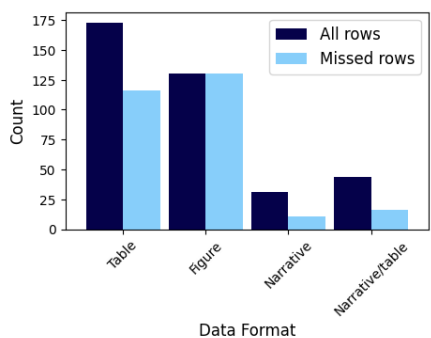
Table 3 gives a detailed breakdown of extraction performance for related properties (all properties other than the ones used to match the entries), properties of high variance (experimental properties that were subjected to change in a paper) and properties of low variance (experimental constants). The two different schemas present the two different notions of information extraction: maximal and minimal. Tasks that require less exploration of contextually related information extraction can follow the minimal approach. From the results, we can see that the extraction of fewer properties aids the model, as it allows more useful extractions with higher recall (on the primary property). The precision value gives us an idea about the hallucinated extractions, and the simpler schema resulted in lesser hallucination compared to the complex schema. The 0-shot prompting did not perform well for low-variance properties, but performed much better at extracting high-variance properties. Langchain performs better than 0-shot in most cases but outperformed by the 1-shot, considering that we cannot draw any conclusion from the Diffusion dataset because the number of extractions are so low.

6 Manual error analysis

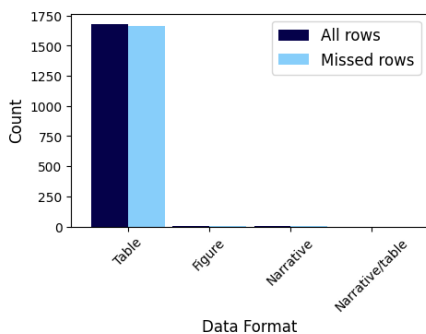
To elucidate what formats key information occurs in, and under what circumstances the model struggles to extract that information, two materials science experts hand-annotated a sample of papers and attempted LLM extractions from each dataset, using the 1-shot prompting variant. 60 papers’ worth of attempted extractions were annotated from the MPEA dataset, constituting a total of 428 rows, and 40 papers from the Diffusion dataset, constituting

| | | 0-shot | | | | 1-shot | | | | Langchain | | | |
|------------------|--------------------------|----------|-----------|----------|-----------|--------|-----------|---------|-----------|-----------|-----------|---------|-----------|
| | | simple | | complex | | simple | | complex | | simple | | complex | |
| | | Recall | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall | Precision |
| MPEA | related properties | 0.360 | 0.245 | 0.311 | 0.163 | 0.323 | 0.248 | 0.269 | 0.180 | 0.449 | 0.254 | 0.301 | 0.155 |
| | low-variance properties | 0.071 | 0.082 | 0.146 | 0.146 | 0.457 | 0.555 | 0.289 | 0.301 | 0.287 | 0.287 | 0.076 | 0.076 |
| | high-variance properties | 0.408 | 0.273 | 0.394 | 0.171 | 0.301 | 0.197 | 0.259 | 0.119 | 0.472 | 0.249 | 0.391 | 0.187 |
| Diffusion | related properties | no match | no match | no match | no match | 0.500 | 0.500 | 0.147 | 0.138 | 0.333 | 0.333 | 0.125 | 0.333 |
| | low-variance properties | no match | no match | no match | no match | 0.500 | 0.500 | 0.250 | 0.250 | 0.500 | 0.500 | 0.250 | 0.250 |
| | high-variance properties | no match | no match | no match | no match | 0.500 | 0.500 | 0.112 | 0.104 | 0.250 | 0.250 | 0.083 | 0.500 |

Table 3: Extraction results for both the datasets can be found in this table. Recall and precision are calculated based on the matched entries found in Table. 2 (if no match is found in required properties, then we cannot report any other results).



(a) MPEA dataset



(b) Diffusion dataset

Figure 3: Counts of rows with key information in different formats.

1714 rows. Both sets were sampled randomly from the set of retrieved and parsed papers for which extractions were attempted. Each row was labeled for the format in which the key information occurred (table, figure, narrative text, etc). Each failure, either missing or hallucinated rows, was annotated for its failure reason, to the best of the annotator’s ability to diagnose the model’s behavior by looking at the PDF, parsed XML, and extracted JSON.

In any given paper, different values will be reported in different forms, such as yield strengths in table and testing temperature in narrative text. Our annotators were instructed to note the format of the majority of the “key information” for each

paper, and based their failure evaluations based on that information.

6.1 Data format

Figure 3 summarizes the data format of key information in both datasets, including both original rows in the dataset and rows the model failed to extract. Possible data formats include table, figures, narrative text, calculated, meaning manually calculated by the creator of the dataset from other information within the paper, and other, a miscellaneous category.

Figure 3A shows that, while a plurality of relevant MPEA data is located in tables, a significant percentage of the data is located in figures, which is inaccessible to any text-only extraction method. A significant percentage of relevant information is also located in written narrative form. Looking at the distribution of data format for rows the LLM failed to extract, we can see that it trivially misses 100% of figure-formatted data, 2/3 of table-formatted data, and a relatively smaller 1/3 of narratively-described data. Figure 3B shows the same numbers for the diffusion dataset, where the vast majority of relevant data was reported in tables, which the LLM in almost every case failed to extract.

6.2 Error analysis

Figure 4 summarizes annotated causes of error between the two datasets. It includes a separate plot for misses and matches (covering all the original rows in the dataset covered by the annotator), and one for hallucinations, indicating extracted rows that could not be matched to an original dataset row. Reported error types include **XML parsing errors**, where GROBID failed to extract key information from the PDF; **figure errors** where key information was stored in figures invisible to the model; **dataset errors** where the original dataset had missing or

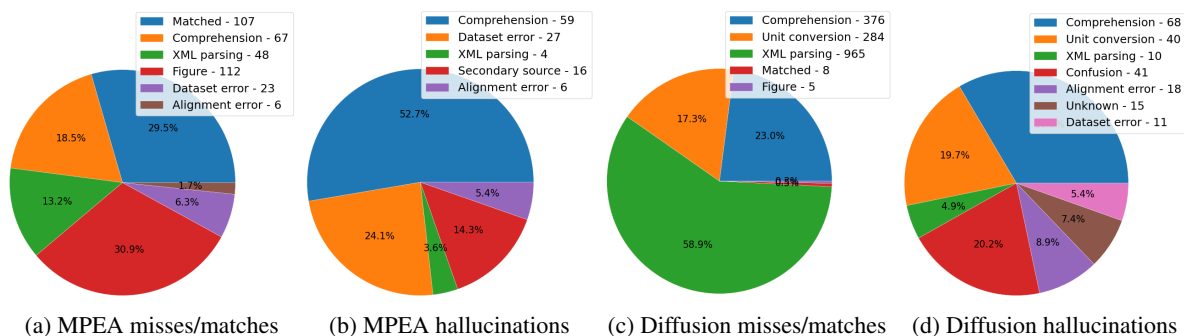


Figure 4: Proportions of differing error reasons, divided by error type (missed row vs. hallucination) and dataset.

incorrect values which the model successfully extracted; **generalized comprehension errors** where the LM overlooked key information; **unit conversion errors** where values were reported in a different unit from that recorded in the dataset; **confusion errors** where the model confused one value for another; **secondary source**, where the model picked up values from other sources referred to in passing; and **alignment errors** where the alignment algorithm described above failed to match extracted to original rows.

MPEA dataset. Figure 4A summarizes error causes reasons for MPEA original dataset rows (both matched and missed). For the 368 ground-truth entries evaluated from the MPEA dataset, a total of 107 were matched (29.5%). Of the missed 70.5%, major sources of error were model comprehension problems (18.5%), key information being presented in figures (30.9%), XML parsing problems (13.2%), and errors in the original dataset (6%).

Thus for this dataset the model’s performance on the extraction task was hindered more by the form of data presentation than by limitations in comprehension ability. The majority of annotated entries in the MPEA set that were missed by the model (160 of 256, or 62.5%) were entries that were located in either a figure or a table that did not parse properly into the XML input. For both of these cases, it is unreasonable to expect the extraction of information that was never presented to the model in the first place, so even though these database values could not be recreated, the failure does not provide much insight from a text comprehension perspective. For the 67 missed MPEA entries that were comprehension-related (annotators manually determined that relevant information was present in text or a table that properly parsed into XML), 2/3 of them were values that appeared

only in tables. Cases of missed values in a table often occurred with tables that were relatively complex in structure, such as in [Senkov et al. \(2020\)](#), where information on yield strength is merged in with information on flow stress and categorized by temperature and relative contribution from material constituent structure—making it difficult to directly extract the yield strength as demanded by the dataset schema.

For values missed in the text itself, some of these errors can be attributed to cases where the value was mentioned in discussion around specific figure or table, making the text itself more difficult to interpret. For example, in [Hu et al. \(2019\)](#), the discussion of MPEA yield strength in the text occurs in the excerpt “[t]he maximum compressive strength is... for the NbZrTi and NbHfZrTi alloys, respectively,” which is stated as part of a discussion around a related figure showing that this “maximum compressive strength” corresponds to the yield strength. However, without the associated figure, the text alone is difficult to properly interpret. Still, there were apparent cases of comprehension failure, such as in [\(Tseng et al., 2018\)](#), where, for one of their target test temperatures, the authors plainly state “the compressive yield strength was 1005 MPa,” a point which the model failed to detect.

Figure 4B summarizes reasons for hallucinated rows in the MPEA dataset. The most common form of hallucination on the MPEA dataset were generalized miscomprehensions (59 cases, 52.7%), often cases where the model identified a valid composition but none of the other relevant properties. The next most common form of reported hallucination (27, 25.6%) were not actually hallucinations, but rather the model finding accurate compositions and property values that were not collected in the construction of the original dataset. The third most

common type of “hallucination” were cases where composition/yield strength combinations reported in prior work were mentioned in the paper at hand, and then extracted by the model. XML parsing was not a major source of error for the MPEA dataset, where most of the included papers are relatively recent.

Diffusion dataset Figure 4C summarizes the percentage of expert-annotated error reasons for diffusion dataset rows (both matched and missed).

The overall performance of the model was distinctly different for this dataset. Out of a total of 1714 dataset entries, only 23 entries were able to be automatically aligned with original records, and 15 of these were missing so many other values as to be effectively useless. Thus, extraction almost completely failed for this model. By far the largest source of error was XML parsing issues, which were responsible for 965 misses. Errors in XML parsing were often related to the age of the publication. Many of the older diffusion publications in the set have outdated typesetting practices and non-standard table structures with multiply-merged cells, both of which created numerous issues in the model input information.

After XML issues, the next largest error source was generalized model comprehension, which was responsible for 387 (24.2%) of misses. The model had significant difficulty correctly reporting textual values like melt descriptions, as these often required domain knowledge connections that were difficult to make in the context of the prompt. For example, Alletti et al. (2007) studied diffusion in a Hawaiitic melt, but during database compilation, this ‘Hawaiitic’ description was instead reported as ‘basalt’ as part of a standardizing effort by Zhang et al. (2010) to reduce the number of unique mineral terms in the database. When the model retrieved summarizing values from the abstract of Alletti et al. (2007), the melt was labeled as “Hawaiitic melt from Mt. Etna”, which is factually correct, but does not align with the database description of basalt. Other key forms of miscomprehension included misreading complex table formats like in the MPEA dataset, and erroneously preferring narratively-presented information to tabular information. In processing (MacKenzie and Canil, 2008), for instance, the model seems to have extracted values presented in the abstract rather than in the centerpiece tables.

Finally, unit mismatches were a significant

source of error, which were related to 17.1% of all rows. When the database was compiled by Zhang et al. (2010), the standard units of m^2/s were chosen for the diffusion coefficient, but it is not uncommon for this coefficient to also be reported in cm^2/s , $\mu m^2/s$, or even as a logarithm ($\log(D)$), depending on the context of the investigation. Unit-based comprehension issues often involved failure to report diffusion coefficient values in the correct units, and failure to correctly convert logarithm-form diffusion coefficients. Without access to an external conversion tool, it is unreasonable to expect the LLM to perform these unit conversions.

Figure 4D summarizes hallucinated rows for the diffusion dataset. Of the 203 reported hallucinations reported on the annotated portion of the Diffusion dataset, 109 (53.6%) were annotated as comprehension issues, including “confusion” instances where the model confused the desired value for something else. An example is (Roselieb and Jambon, 2002), where it incorrectly took Do values from the abstract. Another reported type hallucination was the retrieval of simplified or summarizing values. The remaining hallucinations were a combination of XML issues, assignment comprehension issues, and entries not collected during dataset construction.

7 Discussion

The manual error analysis identifies several major barriers to effective ad-hoc schema-based information extraction from scientific articles.

PDF parsing. PDF parsing is a problem for older PDFs which, for some areas such as geology and certain subfields of materials science, can still contain valuable information worth extracting. This suggests a need for native LM support for PDFs, which, while present in major models such as Claude and ChatGPT via their interfaces, is not yet available in programmatic API form at the time of writing.

Figure comprehension. The MPEA dataset demonstrates that a considerable amount of valuable quantitative scientific information is presented visually rather than in tabular form, suggesting a need for vision-language models capable of reading values from plots and figures. For tasks like determining yield strength from a load curve, a vision model would need to be able to recognize the curve itself and accurately read a singular specific point

off of that curve, which may require specialized tuning. Recent work like Shi et al. (2023) provides a starting point for this.

Unit conversion. A characteristic problem in the diffusion dataset is the presentation of results in units like $\log(m^2/s)$ or cm^2/s when the schema requires m^2/s . A LM capable of normalizing across these different units would need access to external tools capable of doing so, like ToolFormer (Schick et al., 2023).

Table comprehension. By far the most common format for experimental results is tables, and table miscomprehensions, especially those associated with complex or non-standard table formats, formed the bulk of true comprehension errors made by the model on both datasets. Thus, a key research direction needs to be improved table comprehension specifically.

Narrative context for tabular information. While tabular data is typically (though not always) the centerpiece of a given paper, narrative information provides important context for understanding it. Constant values such as the temperature at which experiments were run is often presented narratively (which we want to extract), and selected values from a given table are often summarized narratively (which we don't want to extract). Secondly-reported values from prior work also falls into this category, as it is often presented as context for the primary set of results. Thus, another important research direction toward ad-hoc scientific information extraction is to correctly parse the relationship between tabular and narratively-presented data, perhaps by explicitly treating tabular data as higher priority than narrative data when performing holistic extraction.

Deeper schemas. There are also significant challenges associated with the use of scientific terminology across different fields. This is particularly apparent for the diffusion dataset, which represents a significant effort by an experienced geologist to compile silicate diffusion information and present it in a form that translates across disciplines. The result of this effort shows through particularly in the melt and experiment descriptions in the compiled dataset, both of which have been standardized as much as possible. For an IE model to accurately recreate these types of standardized descriptions, it would need to understand the full schema asso-

ciated with the standardization, and also be sufficiently flexible to generate the unique tokens associated with these schema labels, even when those tokens do not occur within the text itself. Creating a role and prompt for the model that permits this flexibility while also mitigating hallucination is a nontrivial task.

8 Conclusion

We perform an analysis of whether a representative contemporary large language model, GPT-4, is capable of performing complex ad-hoc scientific information extraction sufficient for replicating two materials property datasets. A detailed manual expert error analysis shows that, while the model shows promise for narrative and tabular data, there are significant barriers in both modality and comprehension ability that prevent the model from being adequate to the task using a baseline one-shot prompting approach. Nevertheless, our initial results on this only-recently-recognized information extraction paradigm and our careful analysis will help guide further work in this area.

Limitations

Our goal in this paper is to study the limitations of a representative contemporary large language model (GPT-4) in performing ad-hoc schema-based information extraction. Our analysis approach has its own limitations, however, including the assumption of representativeness in the one model we evaluate and the two materials datasets we evaluate it on. GPT-4's behavior may not generalize to other models, nor do other manually-curated scientific datasets necessarily conform to the structure of the two that we study. There will need to be extensive further experimentation with a wider variety of model and datasets to fully characterize the research frontier for this topic.

Acknowledgments

This work was supported by the OpenAI Researcher Access Program.

References

- Marina Alletti, Don R Baker, and Carmela Freda. 2007. Halogen diffusion in a basaltic melt. *Geochimica et Cosmochimica Acta*, 71(14):3570–3580.
- Don R Baker, Aida Conte, Carmela Freda, and Luisa Ottolini. 2002. The effect of halogens on zircon dif-

- fusion and zircon dissolution in hydrous metaluminous granitic melts. *Contributions to Mineralogy and Petrology*, 142(6):666–678.
- Hélène Balcone-Boissard, DR Baker, Benoît Villemant, and Georges Boudon. 2009. F and cl diffusion in phonolitic melts: Influence of the na/k ratio. *Chemical Geology*, 263(1-4):89–98.
- Christopher K. H. Borg, Carolina Frey, Jasper Moh, Tresa M. Pollock, Stéphane Gorsse, Daniel B. Miracle, Oleg N. Senkov, Bryce Meredig, and James E. Saal. 2020. [Expanded dataset of mechanical properties and observed phases of multi-principal element alloys](#). *Scientific Data*, 7(1):430. Number: 1 Publisher: Nature Publishing Group.
- H Chen, A Kauffmann, S Laube, I-C Choi, R Schwaiger, Y Huang, K Lichtenberg, F Müller, B Gorr, H-J Christ, et al. 2018. Contribution of lattice distortion to solid solution strengthening in a series of refractory high entropy alloys. *Metallurgical and Materials Transactions A*, 49:772–781.
- J-Ph Couzinié, L Lilensten, Y Champion, G Dirras, L Perrière, and I Guillot. 2015. On the room temperature deformation mechanisms of a tizrhfnbta refractory high-entropy alloy. *Materials Science and Engineering: A*, 645:255–263.
- Qingyang Dong and Jacqueline M. Cole. 2022. [Auto-generated database of semiconductor band gaps using ChemDataExtractor](#). *Scientific Data*, 9(1):193. Number: 1 Publisher: Nature Publishing Group.
- Shaofan Ge, Huameng Fu, Long Zhang, Huahai Mao, Hong Li, Aimin Wang, Weirong Li, and Haifeng Zhang. 2020. Effects of al addition on the microstructures and properties of monbtativ refractory high entropy alloy. *Materials Science and Engineering: A*, 784:139275.
- Tom Hope, Doug Downey, Oren Etzioni, Daniel S. Weld, and Eric Horvitz. 2022. [A Computational Inflection for Scientific Discovery](#). Technical Report arXiv:2205.02007, arXiv. 1 citations (Semantic Scholar/arXiv) [2022-08-26] arXiv:2205.02007 [cs] type: article.
- Yu-Min Hu, Xiao-Dong Liu, Na-Na Guo, Liang Wang, Yan-Qing Su, and Jing-Jie Guo. 2019. Microstructure and mechanical properties of nbzrti and nbhfzrti alloys. *Rare Metals*, 38:840–847.
- Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. [Instruct and Extract: Instruction Tuning for On-Demand Information Extraction](#). ArXiv:2310.16040 [cs].
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. [Better Zero-Shot Reasoning with Role-Play Prompting](#). ArXiv:2308.07702 [cs].
- L Lilensten, J-P Couzinié, L Perrière, A Hocini, Clément Keller, G Dirras, and I Guillot. 2018. Study of a bcc multi-principal element alloy: Tensile and simple shear properties and underlying deformation mechanisms. *Acta Materialia*, 142:131–141.
- Yan Long, Xiaobiao Liang, Kai Su, Haiyan Peng, and Xiaozhen Li. 2019. A fine-grained nbmotawvcr refractory high-entropy alloy with ultra-high strength: Microstructural evolution and mechanical properties. *Journal of Alloys and Compounds*, 780:607–617.
- Jason M. MacKenzie and Dante Canil. 2008. [Volatile heavy metal mobility in silicate liquids: Implications for volcanic degassing and eruption prediction](#). *Earth and Planetary Science Letters*, 269(3):488–496.
- D. B. Miracle and O. N. Senkov. 2017. [A critical review of high entropy alloys and related concepts](#). *Acta Materialia*, 122:448–511.
- Daniel B. Miracle, Ming-Hung Tsai, Oleg N. Senkov, Vishal Soni, and Rajarshi Banerjee. 2020. [Refractory high entropy superalloys \(RSAs\)](#). *Scripta Materialia*, 187:445–452.
- Luca Montanelli, Vineeth Venugopal, Elsa A. Olivetti, and Marat I. Latypov. 2024. [High-Throughput Extraction of Phase–Property Relationships from Literature Using Natural Language Processing and Large Language Models](#). *Integrating Materials and Manufacturing Innovation*.
- Meredith Ringel Morris. 2023. [Scientists’ Perspectives on the Potential for Generative AI in their Fields](#). ArXiv:2304.01420 [cs].
- Elsa A. Olivetti, Jacqueline M. Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M. Hiszpanski. 2020. [Data-driven materials research enabled by natural language processing and information extraction](#). *Applied Physics Reviews*, 7(4):041317.
- OpenAI. 2023. [GPT-4 Technical Report](#). ArXiv:2303.08774 [cs].
- Ghanshyam P. P. Pilania. 2021. [Machine learning in materials science: From explainable predictions to autonomous design](#). *Computational Materials Science*, 193:110360.
- Maciej P. Polak, Shrey Modi, Anna Latosinska, Jiming Zhang, Ching-Wen Wang, Shanonan Wang, Ayan Deep Hazra, and Dane Morgan. 2024. [Flexible, Model-Agnostic Method for Materials Data Extraction from Text Using General Purpose Language Models](#). ArXiv:2302.04914 [cond-mat].
- Knut Roselieb and Albert Jambon. 2002. [Tracer diffusion of Mg, Ca, Sr, and Ba in Na-aluminosilicate melts](#). *Geochimica et Cosmochimica Acta*, 66:109–123. ADS Bibcode: 2002GeCoA..66..109R.

- Christian Rüssel and Achim Wiedenroth. 2004. The effect of glass composition on the thermodynamics of the $\text{Fe}^{2+}/\text{Fe}^{3+}$ equilibrium and the iron diffusivity in $\text{Na}_2\text{O}/\text{MgO}/\text{CaO}/\text{Al}_2\text{O}_3/\text{SiO}_2$ melts. *Chemical Geology*, 213(1-3):125–135.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language Models Can Teach Themselves to Use Tools](#). ArXiv:2302.04761 [cs].
- ON Senkov, J-P Couzinie, SI Rao, V Soni, and R Banerjee. 2020. Temperature dependent deformation behavior and strengthening mechanisms in a low density refractory high entropy alloy $\text{Al}_{10}\text{Nb}_{15}\text{Ta}_{5}\text{Ti}_{30}\text{Zr}_{40}$. *Materialia*, 9:100627.
- ON Senkov, AL Pilchak, and SL Semiatin. 2018. Effect of cold deformation and annealing on the microstructure and tensile properties of a hfnbtatizr refractory high entropy alloy. *Metallurgical and Materials Transactions A*, 49:2876–2892.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. [Role play with large language models](#). *Nature*, 623(7987):493–498. Number: 7987 Publisher: Nature Publishing Group.
- Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. [Exploring OCR Capabilities of GPT-4V\(ision\) : A Quantitative and In-depth Evaluation](#). ArXiv:2310.16809 [cs].
- Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. 2023. [An autonomous laboratory for the accelerated synthesis of novel materials](#). *Nature*, 624(7990):86–91. Publisher: Nature Publishing Group.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Ko-Kai Tseng, Chien-Chang Juan, Shuen Tso, Hsuan-Chu Chen, Che-Wei Tsai, and Jien-Wei Yeh. 2018. Effects of mo, nb, ta, ti, and zr on mechanical properties of equiatomic hf-mo-nb-ta-ti-zr alloys. *Entropy*, 21(1):15.
- Nicholas Walker, Sanghoon Lee, John Dagdelen, Kevin Cruse, Samuel Gleason, Alexander Dunn, Gerbrand Ceder, A. Paul Alivisatos, Kristin A. Persson, and Anubhav Jain. 2023. [Extracting structured seed-mediated gold nanorod growth procedures from scientific text with LLMs](#). *Digital Discovery*, 2(6):1768–1782. Publisher: Royal Society of Chemistry.
- Xingbo Wang, Samantha L. Huey, Rui Sheng, Saurabh Mehta, and Fei Wang. 2024. [SciDaSynth: Interactive Structured Knowledge Extraction and Synthesis from Scientific Literature with Large Language Model](#). ArXiv:2404.13765 [cs].
- Tong Xie, Yuwei Wan, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, and Bram Hoex. 2023. [Large Language Models as Master Key: Unlocking the Secrets of Materials Science with GPT](#). ArXiv:2304.02213 [cs].
- Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023a. [Large Language Models for Generative Information Extraction: A Survey](#). ArXiv:2312.17617 [cs].
- Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. 2023b. [Small data machine learning in materials science](#). *npj Computational Materials*, 9(1):1–15. Number: 1 Publisher: Nature Publishing Group.
- Samuel J. Yang, Shutong Li, Subhashini Venugopalan, Vahe Tshitoyan, Muratahan Aykol, Amil Merchant, Ekin Dogus Cubuk, and Gowoon Cheon. 2023. [Accurate Prediction of Experimental Band Gaps from Large Language Model-Based Data Extraction](#).
- Youxue Zhang, Huaiwei Ni, and Yang Chen. 2010. [Diffusion Data in Silicate Melts](#). *Reviews in Mineralogy and Geochemistry*, 72(1):311–408.
- Youxue Zhang, David Walker, and Charles E Leshner. 1989. Diffusive crystal dissolution. *Contributions to Mineralogy and Petrology*, 102:492–513.

A Appendix

A.1 Dataset column descriptions

Table 4 describes each major column in each dataset.

A.2 Full prompt

The prompts used for both extraction tasks are given below.

A.2.1 Role setting for both tasks

I am a helpful assistant capable of extracting information from text. I will not generate any new tokens, only extract tokens containing information from the text you provide. If provided with a schema, then I will follow it and return the information in format of the schema.

A.2.2 Prompt for MPEA extraction

Extract all information relating to every High Entropy Alloys (HEA) from the following text: {exemplar text / other text to extract from} using the following schema: schema. Return a list of

| MPEA dataset | | Diffusion dataset | |
|---|--|--------------------|--|
| Column | Definition | Column | Definition |
| FORMULA | Composition of the alloy | melt | Common name of substance being diffused into (e.g. Andesite, Basalt11) |
| PROPERTY: Microstructure | Crystal structure of alloy | diffusing species | Element diffusing into melt |
| PROPERTY: Processing method | Synthesis method (e.g. CAST, ANNEAL, POWDER, WROUGHT, OTHER) | type of diffusion | How atoms are diffusing into melt (e.g. Tracer, SEBD, FEBD, self) |
| PROPERTY: BCC/FCC/other | Categorical subset of crystal structure | type of experiment | Physical experiment setup |
| PROPERTY: grain size (μm) | Grain size within alloy | T(K) | Experiment temperature |
| PROPERTY: Exp. Density (g/cm^3) | Experimental density of alloy | P(MPa) | Experiment pressure |
| PROPERTY: Computed Density (g/cm^3) | Computed density of alloy | D(m2/s) | Diffusivity; speed of diffusion |
| PROPERTY: HV | Vicker's hardness of alloy | 1s error | SD error in diffusivity when reported. |
| PROPERTY: Type of test | Type of test (C=compression, T=tensile) | SiO2 | |
| PROPERTY: Test temperature ($^{\circ}\text{C}$) | Testing temperature | TiO2 | |
| PROPERTY: YS (MPa) | Yield strength | Al2O3 | |
| PROPERTY: UTS (MPa) | Ultimate tensile strength | FeOt | |
| PROPERTY: Elongation (%) | Elastic elongation at break | MnO | |
| PROPERTY: Elongation plastic (%) | Plastic elongation at break | MgO | Percentage of each compound in the composition of the melt |
| PROPERTY: Exp. Young modulus (GPa) | Experimental Young's Modulus | CaO | |
| PROPERTY: Computed Young modulus (GPa) | Computed Young's modulus | Na2O | |
| PROPERTY: O content (wppm) | Oxygen content | K2O | |
| PROPERTY: N content (wppm) | Nitrogen content | P2O5 | |
| PROPERTY: C content (wppm) | Carbon content | H2Ot | |
| | | Composition unit | Unit of compound composition |
| | | dry total | Non-H2O fraction of melt |

Table 4: Descriptions of individual columns in MPEA and diffusion datasets.

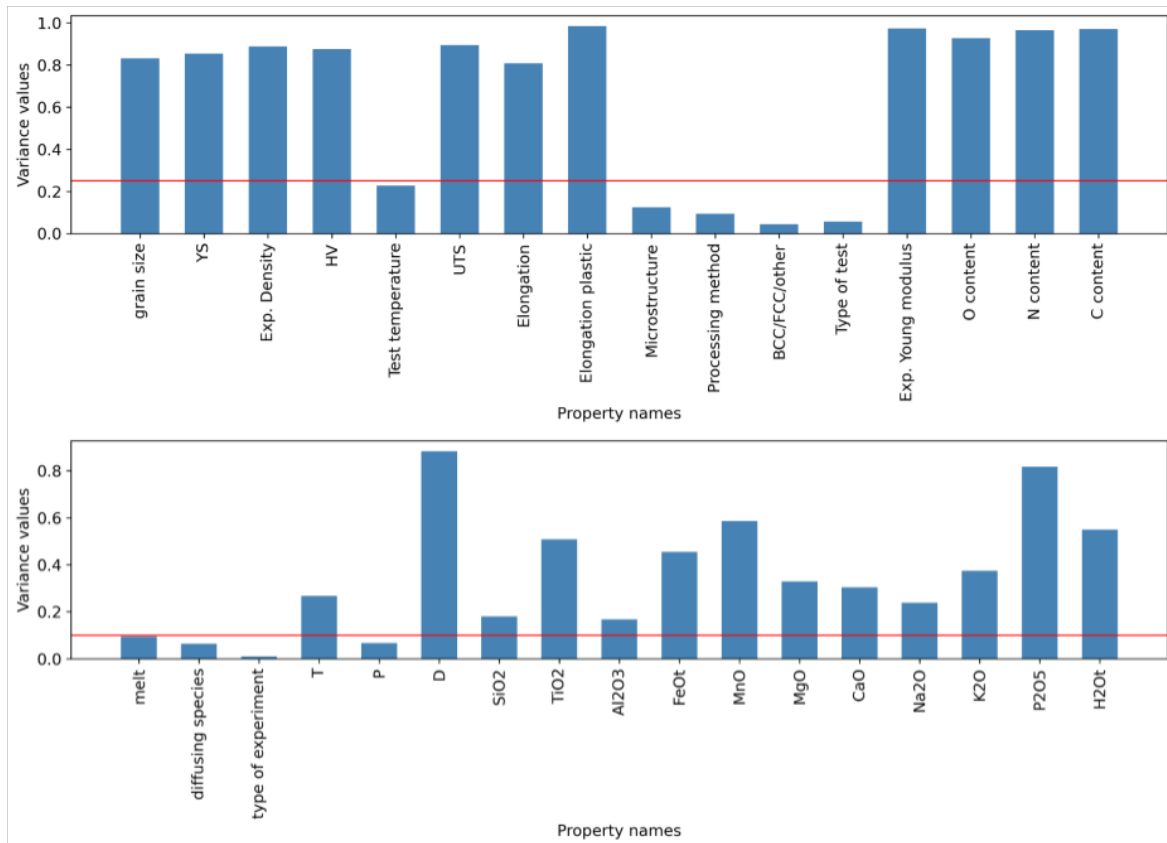


Figure 5: Plot to visualize the variance of properties found in two datasets. High variance means properties that are altered in experiments and low variance show experimental constants.

schemas in JSON format for every unique compound.

Additional context: HEA composition sometimes has variable ratio denoted by 'x', if so, then replace x with a float or integer as applicable by the information in the paper.

Restriction: Do not extract any tokens if a HEA material does not have a property mentioned in the schema. If a property is not available then return 'No information'. Do not violate the schema.

A.2.3 Schema for MPEA extraction

```
{
  "high entropy alloy formula": {"type": "string"},
  "microstructure": {"type": "string"},
  "processing method": {"type": "string"},
  "BCC/FCC/other": {"type": "string"},
  "grain size": {"type": "float"},
  "experimental density": {"type": "float"},
  "hardness": {"type": "float"},
  "type of test": {"type": "string"},
  "test temperature": {"type": "float"},
  "yield strength": {"type": "float"},
  "ultimate tensile strength": {"type": "float"},
  "elongation": {"type": "float"},
  "elongation plastic": {"type": "float"},
  "experimental young modulus": {"type": "float"},
  "oxygen content": {"type": "float"},
  "nitrogen content": {"type": "float"},
  "carbon content": {"type": "float"}
}
```

A.2.4 Exemplar output for MPEA extraction

One entry from the one-shot output exemplar that we used is given below. The actual output contains 5 of these entries in a JString list.

```
{
  "high entropy alloy formula": 'NbMoTaWVCr',
  "microstructure": 'BCC+Laves+Sec.',
  "processing method": 'POWDER',
  "BCC/FCC/other": 'other',
  "grain size": 0.54,
  "experimental density": 'No information',
  "hardness": 1072.0,
  "type of test": 'C',
  "test temperature": 25.0,
  "yield strength": 'No information',
  "ultimate tensile strength": 'No information',
  "elongation": 'No information',
  "elongation plastic": 'No information',
  "experimental young modulus": 'No information',
  "oxygen content": 7946.0,
  "nitrogen content": 'No information',
  "carbon content": 'No information'
}
```

A.2.5 Prompt for Diffusion extraction

Extract all information relating to diffusion of elements into a silicate melt from the following text: {one shot text / other text to extract from} using

the following schema: {schema}. Return a list of schemas in JSON format for every unique compound.

Additional context: there will generally be one or multiple liquids, and one or more elements that move through that liquid at a certain speed. That speed is called the "diffusivity" or the "diffusion coefficient" $D(\text{m}^2/\text{s})$. Each element will have a unique diffusivity in each melt and temperature and pressure, so it is important to keep track of every combination.

Restriction: Do not extract any tokens if no relevant property is present as mentioned in the schema. If a property is not available then return 'No information'. Do not violate the schema.

A.2.6 Schema for Diffusion extraction

```
{
  'melt': {'type': 'string'},
  'diffusing species': {'type': 'string'},
  'type of experiment': {'type': 'string'},
  'test temperature': {'type': 'float'},
  'pressure': {'type': 'float'},
  'diffusivity': {'type': 'float'},
  'SiO2': {'type': 'float'},
  'TiO2': {'type': 'float'},
  'Al2O3': {'type': 'float'},
  'FeO': {'type': 'float'},
  'MnO': {'type': 'float'},
  'MgO': {'type': 'float'},
  'CaO': {'type': 'float'},
  'Na2O': {'type': 'float'},
  'K2O': {'type': 'float'},
  'P2O5': {'type': 'float'},
  'H2O': {'type': 'float'}
}
```

A.2.7 Exemplar output for Diffusion extraction

The following is one of 21 entries in the one-shot exemplar output. The actual output is a list of such entries given to the model as a JString.

```
{
  "melt": "NCMAS6",
  "diffusing species": "Fe",
  "type of experiment": "electrochemistry",
  "test temperature": 1573.15,
  "pressure": "No information",
  "diffusivity": 1.35e-07,
  "SiO2": 80.6793201360426,
  "TiO2": "No information",
  "Al2O3": 0.0,
  "FeO": "No information",
  "MnO": "No information",
  "MgO": 0.0,
  "CaO": 14.11921335907197,
  "Na2O": 5.201466504885413,
  "K2O": "No information",
  "P2O5": "No information",
  "H2O": "No information"
}
```

A.2.8 Exemplar text for both tasks

The text is too long to be presented here but can be viewed in (Long et al., 2019) (MPEA) and (Rüssel and Wiedenroth, 2004) (Diffusion). Input text comprised of the abstract along with other sections and tables (including their captions) sequentially as present in the paper.

A.3 Illustrative examples of errors

As mentioned in the discussion section we found errors occurring primarily due to some reasons like table comprehension, text comprehension, ground-truth information present in plots/images, unit conversion problems, poor XML parsing, simplification of information, and partial information extraction from secondary sources. We wanted to show what those problems look like so in the subsections below we will give an example for each of the errors. We also show an example of an instance when the model was actually correct, in-spite not being matched with the human annotator.

A.3.1 Table comprehension

In Figure 6 taken from Senkov et al. (2018) we found that the model was able to extract the yield strength of three out of eight yield strength value even when the caption clearly mentioned which column corresponds to the true yield strength.

Table IV. Tensile Properties (Young's Modulus, E , True Yield Strength, $\sigma_{0.2}$, True Peak Stress, σ_p , Fracture Stress, σ_f , and True Fracture Strain, ϵ_f) of As-Rolled and Rolled-and-Annealed Sheet Samples of HfNbTaTiZr

| Condition | E (GPa) | $\sigma_{0.2}$ (MPa) | σ_p (MPa) | σ_f (MPa) | ϵ_f (pt) |
|----------------|-----------|----------------------|------------------|------------------|-------------------|
| As-rolled | 91.5 | 1438 | 1495 | 1218 | 5.0 |
| 800 °C, 4 h | 99.2 | 1159 | 1174 | 1127 | 4.7 |
| 800 °C, 16 h | 109.1 | 1134 | 1165 | 1134 | 7.6 |
| 800 °C, 64 h | 105.3 | 1094 | 1110 | 1066 | 6.0 |
| 800 °C, 256 h | 91.6 | 1071 | 1074 | 964 | 4.6 |
| 1100 °C, 0.5 h | 92.6 | 1107 | 1159 | 977 | 13.9 |
| 1100 °C, 1 h | 92.6 | 1115 | 1195 | 830 | 18.5 |
| 1100 °C, 2 h | 101.3 | 1114 | 1168 | 935 | 12.2 |

Figure 6: This is an example table that the model failed to comprehend even when the XML parsing was perfect.

A.3.2 Text comprehension

The text in Figure 7 is a paragraph of text taken from Couzinié et al. (2015), where the model could not understand that the text is mentioning information about yield strength.

A.3.3 Information in plots/images

A lot of missed extraction were due to the information being present in plots. As the XML parsing cannot process the image this information is skipped. In Figure 8, we can see such a plot where the information about the yield strength was present in the plot (Chen et al., 2018).

The monotonic compressive behavior of the equimolar HEA is illustrated by the representation of the true stress with plastic strain $\sigma_T=f(\epsilon_p)$ in Fig. 4. The micro yield stress is estimated to 820 ± 20 MPa, the yield stress $\sigma_{0.2}$ to 905 ± 10 MPa, close to the 929 MPa obtained in the pioneering study of this alloy [23]. The evolution of the work hardening $\Theta=d\sigma_T/d\epsilon_p$ with plastic strain reveals the presence of 3 different stages. The stage 1 corresponds to a sharp decrease of Θ for low plastic strain, in the range of $\epsilon_p=0$ up to $\epsilon_p \approx 3\%$. In stage 2, the work hardening is close to 1300 ± 50 MPa, corresponding to $\Theta/\mu \approx 0.033$ if one consider $\mu=40$ GPa following Senkov and co-authors. It should be

Figure 7: This image shows the paragraph which the model failed to comprehend and did not extract the yield strength value (highlighted text).

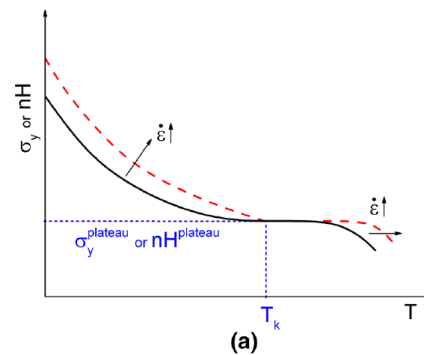


Figure 8: Figure of a plot that contains information about the relationship between yield strength and temperature.

A.3.4 Unit conversion

In multiple situation of the diffusion dataset we saw that the model extracted the correct information as present in the paper but as the units are not converted as the gold-extractions therefore we could not match them. An example this error can be found in Figure 9 from Zhang et al. (1989).

A.3.5 XML parsing

XML parsing of some older papers from the diffusion dataset was not of good quality and this contributed to some failed extraction. One such example can be found in Figure 10 from Baker et al. (2002). The XML created from the PDF file was not parsed properly and the actual values in the cells were just missed during the parsing process. The actual table is provided in Figure 11.

A.4 Simplification

In Figure 12, we can see the abstract from Balcone-Boissard et al. (2009), that mentions different values of diffusion rates at different temperatures. The model extracted all the diffusion rates correctly but missed the varying temperatures associated with them and associated them with one temperature

Table 6. EBDC

| Run # | D_{MgO} | $D_{Al_2O_3}$ | D_{SiO_2} | D_{CaO} |
|----------------------|-----------|---------------|-------------|-----------|
| 212(OL) ^a | 5.3 | 3.3 | 2.0 | — |
| 216(OL) | 4.3 | 1.4 | 2.5 | — |
| 219(OL) | 3.0 | 1.6 | 1.0 | — |
| 222(OL) | 6.6 | 4.2 | 2.3 | — |
| 223(DI) | 4.4 | 2.5 | — | 8.6 |
| 225(DI) | 5.1 | 3.8 | — | 9.5 |
| 226(FO) | 6.0 | 2.6 | 1.7 | — |
| 227(FO) | 7.1 | 3.0 | 2.6 | — |
| 228(OL) | 8.5 | 4.0 | 2.2 | — |
| 228(FO) | 9.4 | 4.7 | 2.6 | — |
| 229(OL) | 1.8 | 1.1 | 0.57 | — |
| 231(OL) | 11.4 | 8.5 | 3.6 | — |
| 231(SP) | 8.5 | 3.8 | 2.5 | — |
| 231(DI) | 10 | 7.3 | — | 22 |
| 234(QT) | 0.2 | 0.42 | 0.28 | 0.21 |
| 235(OL) | 12 | 7.0 | 5.0 | — |
| 236(OL) | 24 | 16 | 12 | — |
| 236(RT) ^b | — | — | 5.5 | — |
| 239(OL) | 20 | 12 | 9.4 | — |
| 242(DI) | 18 | 10 | — | 26 |

Note: The EBDC's are for oxide components in μ^2/s . Run #: Included in parenthesis is the crystal for which the dissolution distance is measured and near which the profile is considered. Relative errors (at 1σ level) from the curve fitting are small, usually less than 10%. However, a more realistic measure of errors comes from inter-run comparisons and can be estimated from Fig. 17
— the EBDC cannot be calculated because of uphill diffusion or because the concentration is almost constant; ^a Cr diffusion coefficient is $3.6 \mu^2/s$; ^b TiO₂ diffusion coefficient is $8.2 \mu^2/s$

Figure 9: Example of a table with failed extraction due to unit conversion error. The table presented rate of diffusion in μ/s^2 whereas the gold extraction where in m/s^2 .

```
<figure xmlns="http://www.tei-c.org/ns/1.0" type="table" xml:id="tab_2">
<head>Table 2</head>
<tbl_struct>
<tbl_header>
<tbl_info cols="5">
<tbl_r cells="5" ix="1" maxcspan="1" maxrspan="1" usedcols="5">
<tbl_r cells="5" ix="2" maxcspan="1" maxrspan="1" usedcols="5">
<tbl_r cells="5" ix="3" maxcspan="1" maxrspan="1" usedcols="5">
<tbl_r cells="5" ix="4" maxcspan="1" maxrspan="1" usedcols="5">
<tbl_r cells="5" ix="5" maxcspan="1" maxrspan="1" usedcols="5">

```

Figure 10: The given XML text was the parsed text from the PDF file.

value, thereby simplifying the text.

A.5 Secondary source

In Liliensten et al. (2018), we can find values of yield strength reported from other papers that were reviewed by the author. The authors reported these values as a summary of the entire paper without going into the details and that can be seen in Figure 13. We considered these extraction as errors as they are

Table 2 Experimental conditions and results. Numbers in parentheses are the standard deviations (1σ) based upon multiple analyses. Counting statistics for Zr at a concentration of ~2,000 ppm produce an uncertainty of ~170 ppm. FeO* concentrations have not been corrected for low totals because of the presence of water in the glasses. Experiments starting from AA29 were performed in a pyrophyllite-bearing assembly (see text). Zircon crystals used in experiments up to AA37 were from the Mud Tank Carbonatite (Australia); later experiments used zircon crystals from Coloma (Sri Lanka)

| Sample | Temp. (°C) | Time (s) | FeO* (wt%) | H ₂ O* (wt%) | n^b | c_o (ppm Zr) | D_o ($m^2 s^{-1}$) |
|--------|------------|----------|-------------|-------------------------|-------|----------------|-------------------------------|
| LCO | | | | | | | |
| AA6 | 1,050 | 154,800 | 0.72 (0.03) | 3.92 | 3 | 875 (20) | 5.81 (1.05) $\times 10^{-14}$ |
| AA32 | 1,050 | 154,800 | 0.71 (0.03) | 4.28 | 3 | 767 (52) | 5.13 (2.66) $\times 10^{-14}$ |
| AA3 | 1,200 | 72,000 | 0.64 (0.07) | 3.05 | 3 | 2,100 (71) | 2.14 (0.38) $\times 10^{-13}$ |
| AA29 | 1,200 | 72,000 | 0.20 (0.02) | 4.83 | 2 | 3,088 (619) | 6.06 (6.28) $\times 10^{-13}$ |
| AA37 | 1,200 | 28,800 | 0.67 (0.03) | 4.52 | 3 | 2,738 (18) | 1.19 (0.59) $\times 10^{-12}$ |
| AA39 | 1,200 | 28,800 | 0.20 (0.02) | 4.48 | 3 | 2,252 (554) | 8.96 (4.47) $\times 10^{-12}$ |
| AA38 | 1,200 | 28,800 | 0.73 (0.03) | 4.68 | 3 | 3,125 (35) | 1.69 (0.38) $\times 10^{-12}$ |
| AA9 | 1,400 | 7,200 | 0.08 (0.01) | 3.54 | 3 | 11,492 (418) | 1.18 (0.47) $\times 10^{-12}$ |
| AA50 | 1,400 | 7,200 | 0.20 (0.01) | 3.91 | 3 | 11,538 (336) | 1.02 (0.06) $\times 10^{-12}$ |
| AA33 | 1,200 | 21,600 | 0.70 (0.04) | 0.41 | 1 | 925 | 6.39 $\times 10^{-13}$ |
| AA12 | 1,200 | 72,000 | 0.66 (0.02) | 0.56 | 2 | 1,500 (50) | 3.60 (0.91) $\times 10^{-13}$ |
| AA40 | 1,200 | 28,800 | 0.60 (0.04) | 1.78 | 3 | 2,742 (144) | 5.84 (1.41) $\times 10^{-14}$ |
| AA41 | 1,200 | 21,600 | 0.75 (0.03) | 1.93 | 3 | 2,550 (863) | 4.98 (1.49) $\times 10^{-14}$ |
| AA16 | 1,300 | 18,000 | 0.44 (0.05) | 0.63 | 3 | 5,100 (163) | 6.77 (1.77) $\times 10^{-14}$ |
| AA20 | 1,400 | 12,240 | 0.29 (0.13) | 1.43 | 5 | 8,330 (581) | 2.35 (0.44) $\times 10^{-13}$ |
| AA53 | 1,400 | 12,240 | 0.54 (0.03) | 1.59 | 3 | 5,200 (1,713) | 1.33 (0.36) $\times 10^{-13}$ |
| LCl | | | | | | | |
| AA4 | 1,050 | 154,800 | 0.64 (0.05) | 3.69 | 3 | 905 (19) | 4.46 (1.86) $\times 10^{-14}$ |
| AA58 | 1,050 | 154,800 | 0.70 (0.04) | 4.52 | 3 | 1,088 (194) | 4.60 (1.74) $\times 10^{-14}$ |
| AA1 | 1,200 | 72,000 | 0.76 (0.03) | 2.54 | 3 | 3,087 (209) | 1.14 (0.14) $\times 10^{-13}$ |
| AA31 | 1,200 | 72,000 | 0.69 (0.02) | 4.97 | 2 | 3,113 (407) | 1.48 (0.16) $\times 10^{-13}$ |
| AA24 | 1,400 | 7,200 | 0.02 (0.01) | 3.05 | 2 | 10,613 (213) | 7.00 (1.63) $\times 10^{-13}$ |
| AA51 | 1,400 | 7,200 | 0.16 (0.02) | 3.93 | 2 | 9,750 (468) | 3.12 (0.60) $\times 10^{-13}$ |
| AA10 | 1,200 | 72,000 | 0.76 (0.02) | 0.51 | 2 | 1,538 (176) | 1.06 (0.30) $\times 10^{-14}$ |
| AA3 | 1,200 | 21,600 | 0.65 (0.02) | 0.91 | 3 | 1,617 (278) | 9.97 (0.76) $\times 10^{-15}$ |
| AA42 | 1,200 | 21,600 | 0.53 (0.05) | 2.20 | 1 | 1,975 | 4.80 $\times 10^{-14}$ |
| AA28 | 1,400 | 3,780 | 0.74 (0.03) | 0.66 | 4 | 7,500 (966) | 1.19 (0.33) $\times 10^{-13}$ |
| AA18 | 1,400 | 12,240 | 0.63 (0.02) | 1.20 | 3 | 6,533 (292) | 1.09 (0.15) $\times 10^{-13}$ |
| AA54 | 1,400 | 12,600 | 0.70 (0.10) | 0.99 | 3 | 6,830 (141) | 8.67 (0.91) $\times 10^{-14}$ |
| MRP | | | | | | | |
| AA5 | 1,050 | 154,800 | 1.69 (0.02) | 4.90 | 4 | 916 (7) | 2.85 (0.53) $\times 10^{-14}$ |
| AA33 | 1,050 | 154,800 | 1.76 (0.05) | 5.08 | 2 | 913 (159) | 5.23 (1.03) $\times 10^{-14}$ |
| AA2 | 1,200 | 72,000 | 1.85 (0.04) | 3.34 | 2 | 1,983 (24) | 2.44 (0.16) $\times 10^{-13}$ |
| AA59 | 1,200 | 169,200 | 1.69 (0.04) | 4.96 | 2 | 3,200 (121) | 2.14 (0.28) $\times 10^{-13}$ |
| AA27 | 1,400 | 3,780 | 1.65 (0.03) | 4.46 | 3 | 11,267 (47) | 2.63 (0.05) $\times 10^{-12}$ |
| AA52 | 1,400 | 7,200 | 0.55 (0.02) | 3.48 | 1 | 9,400 | 8.51 $\times 10^{-14}$ |
| AA2 | 1,200 | 21,600 | 1.80 (0.03) | 1.34 | 3 | 928 (33) | 6.94 (0.08) $\times 10^{-14}$ |
| AA11 | 1,200 | 72,000 | 1.78 (0.04) | 1.99 | 3 | 2,333 (24) | 8.38 (0.69) $\times 10^{-14}$ |
| AA60 | 1,200 | 28,800 | 1.69 (0.14) | 2.57 | 3 | 2,000 (566) | 4.50 (1.47) $\times 10^{-14}$ |
| AA15 | 1,300 | 18,000 | 1.59 (0.04) | 1.17 | 3 | 3,350 (74) | 5.16 (0.39) $\times 10^{-14}$ |
| AA19 | 1,400 | 12,240 | 1.63 (0.07) | 2.01 | 2 | 10,400 (495) | 3.85 (0.14) $\times 10^{-13}$ |
| AA55 | 1,400 | 12,600 | 0.74 (0.03) | 2.42 | 3 | 7,700 (495) | 3.34 (1.99) $\times 10^{-13}$ |

^aH₂O measured by ion microprobe
^bNumber of traverses

^cExperiments performed with graphite and pyrophyllite powder around the capsules (see text)

Figure 11: Table that was parsed incorrectly leading to information loss.

simultaneous diffusion of a halogen mixture (F, Cl, Br) in order to evaluate the interactions between the halogens during diffusion. Diffusion coefficients for F range between $2 \times 10^{-11} m^2/s$ at 1250 °C and $7 \times 10^{-11} m^2/s$ at 1450 °C for the Na-rich melt and between $1 \times 10^{-11} m^2/s$ at 1250 °C and $8 \times 10^{-11} m^2/s$ at 1450 °C for the K-rich melt at anhydrous conditions. Diffusion coefficients for Cl range between $2 \times 10^{-11} m^2/s$ at 1250 °C and $1 \times 10^{-11} m^2/s$ at 1450 °C for the Na-rich melt and between $7 \times 10^{-12} m^2/s$ at 1250 °C and $2 \times 10^{-11} m^2/s$ at 1450 °C for the K-rich

Figure 12: Example of text simplification performed by the model.

from secondary sources. In the given example the model extracted the value 880 MPa along with the generic formula of TiZrNbHf as mentioned in the text without any other information.

Possible control of the microstructure was afterwards demonstrated, using thermo-mechanical treatments [10,11]. Some uncommon features were already reported, as for example the case of the recrystallized quaternary bcc TiZrNbHf displaying a yield strength of 880 MPa in tension, at room temperature [12].

Figure 13: An example of secondary information that was collected by the model without a detailed context around it.

A.6 Actually correct

In this example taken from Ge et al. (2020) the model extracted correct yield strength value as reported. The gold extractions reported by the expert had rounded off yield strength values and the formula was reported as ratio of elements, leading to matching failure.

| x | ρ (g/cm ³) | Lattice parameters | Bv | c_{11} (MPa) | c_{12} (MPa) | c_{44} (MPa) | c_{12}/c_{44} | Δc_{11} | Δc_{12} | Δc_{44} |
|-----|-----------------------------|--------------------|-------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|
| 0 | 9.33 ± 0.01 | 3.2155 ± 0.0026 | 421.7 ± 7.9 | 1227.9 | 2066 | 1791 | 30.00 | -13.773 | 0.055 | 0.35 |
| 0.2 | 9.12 ± 0.03 | 3.2054 ± 0.0016 | 473.6 ± 9.2 | 1329.0 | 1791 | 1600 | 16.00 | -16.974 | 0.031 | 0.11 |
| 0.4 | 8.98 ± 0.02 | 3.2035 ± 0.0021 | 510.0 ± 8.5 | 1332.7 | 1776 | 1300 | 13.00 | -8.379 | 0.017 | 0.17 |
| 0.6 | 8.79 ± 0.02 | 3.2014 ± 0.0019 | 530.0 ± 7.7 | 1352.5 | 1668 | 1100 | 11.00 | -3.227 | 0.020 | 0.26 |
| 1 | 8.15 ± 0.01 | 3.1998 ± 0.0045 | 574.8 ± 9.6 | 1391.2 | 1496 | 900 | 9.00 | -4.283 | 0.033 | 0.33 |

Figure 14: Model extracted values from this table as found in the paper.