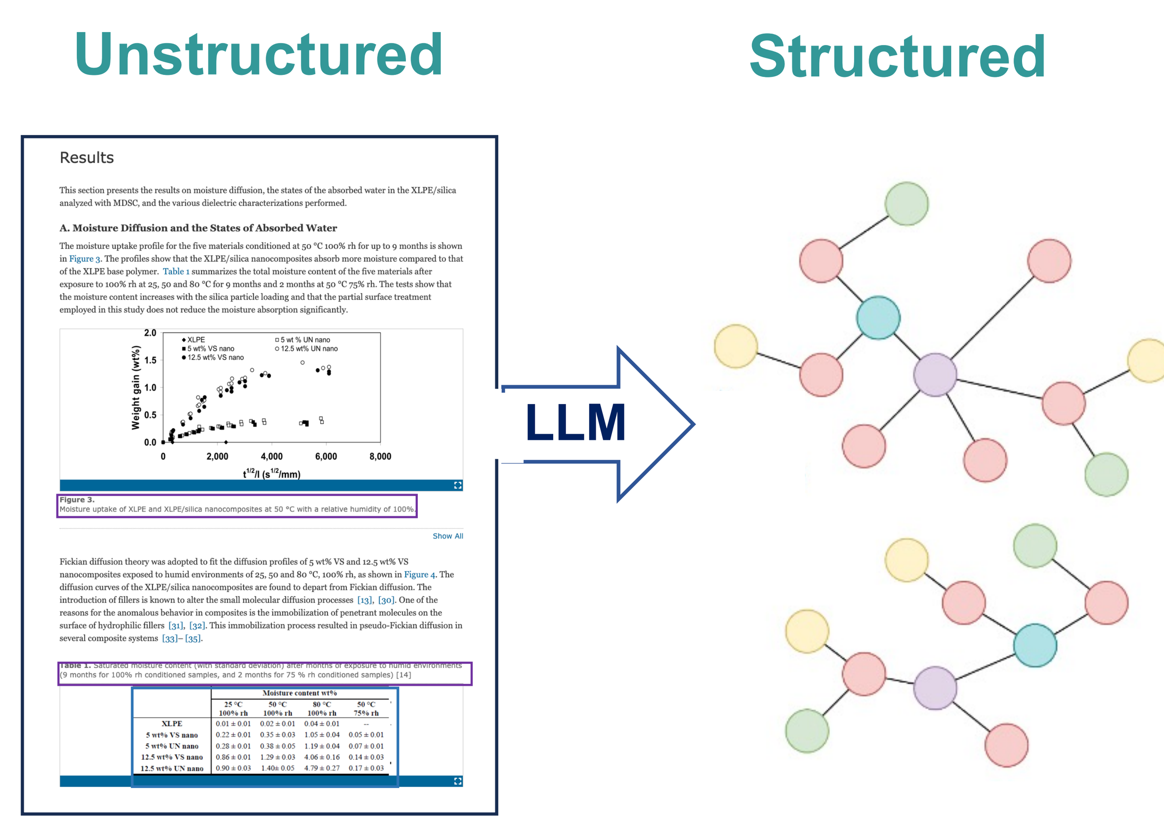
Using Large Language Models for Data Extraction from Tables in Materials Literature

Defne Circi, Ghazal Khalighinejad, Anlan Chen
Bhuwan Dhingra, L. Catherine Brinson

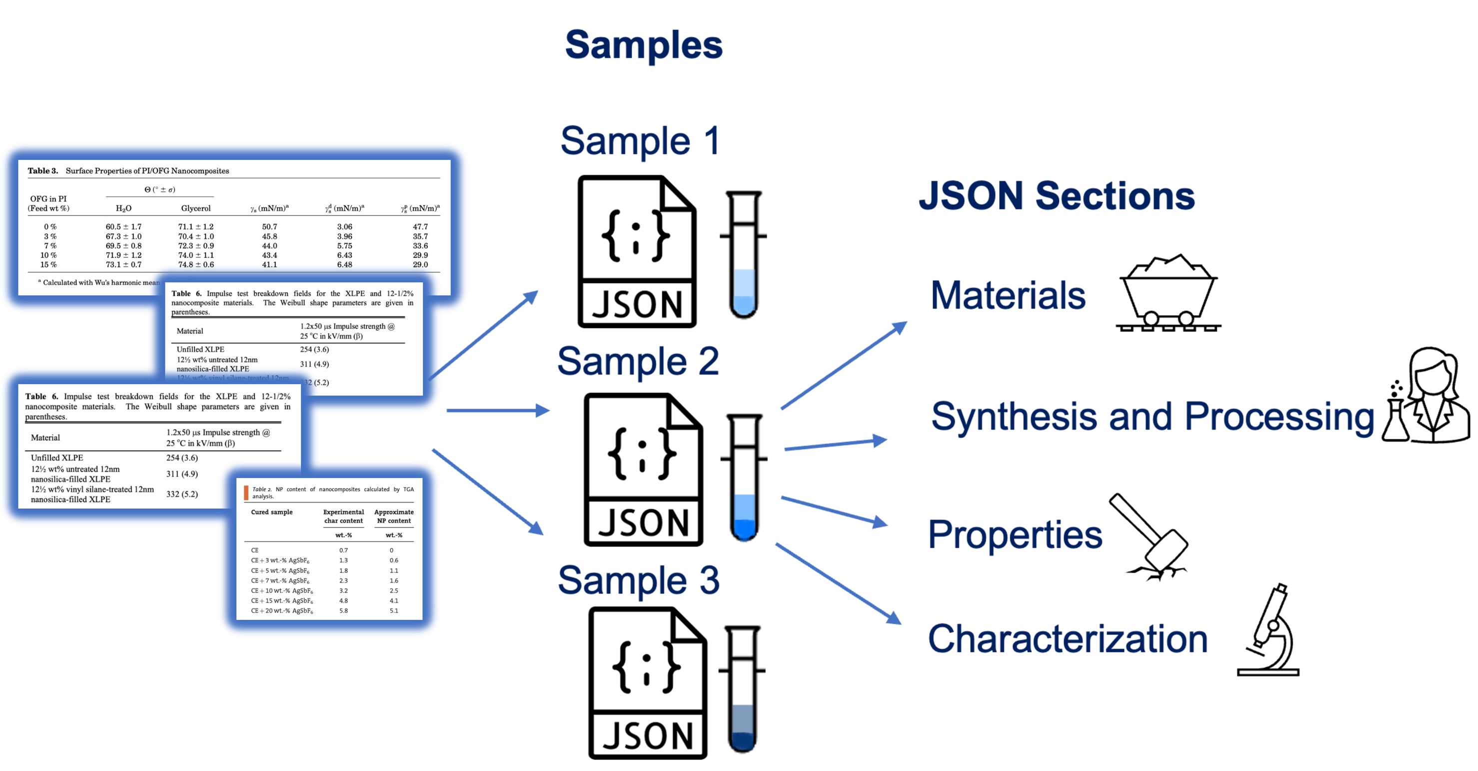
Duke UNIVERSITY

Unlocking Insights in Scientific Literature

- **Finding and using data from literature** is a common problem.
- We need to **search among many documents** for key information.
- Traditionally, data extraction is done **manually → time consuming** and **tedious**
- Collecting experimental data **at a scale** is critical.
- **Large language models (LLMs)** can make the information most important to scientists, such as material identification and properties readily available.
- Composition and properties of materials are predominantly **condensed in tables**.



1. Goal: Extracting multiple experimental samples per table

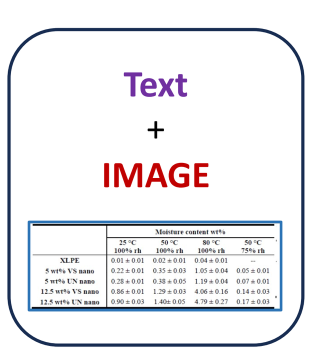


2. Dataset overview and ground truth samples

- Articles from **MaterialsMine** database
- **Table dataset**: 18 articles, 37 tables and their captions, 182 samples
- Sample size range from 2 to 15
- On average 3.1 material properties in each table

3. Choosing inputs of table data

Option 1: GPT-4-Vision on table image

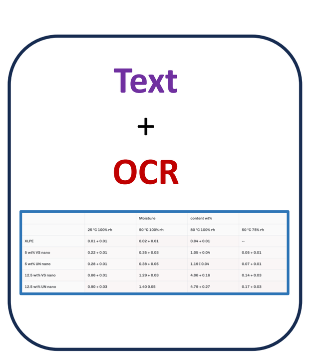


GPT-4(V)

Table 1 Scale parameters and shape parameters

Sample Type	η	β
Tritherm at 300°C	122	4
5 wt% untr silica at 300°C	181	4
10 wt% untr silica at 300°C	220	3

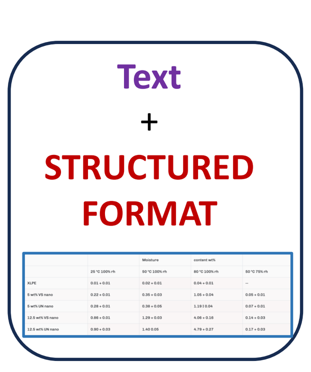
Option 2: GPT-4 on unstructured OCR extraction from table image



GPT-4

```
Table 1. Scale parameters
and shape parameters
n
Tritherm at 300°C
122
4
5 wt% untr silica at 300°C
181
4
10 wt% untr silica at 300°C
220
3
```

Option 3: GPT-4 on unstructured OCR extraction from table image



GPT-4

```
Table 1. Scale parameters and shape
parameters
0,1,2,3
,n
Tritherm at 300°C,122,4
5 wt% untr silica at 300°C,181,4
10 wt% untr silica at 300°C,220,3
```

4. Evaluation of LLM output

Composition level evaluation

Ground Truth

Sample id: 1,
matrix name: PP,
filler name: silica,
composition: {amount: 5%, type: wt},
particle surface treatment name: not specified,

match / not a match / partial match / match

Predictions

Sample id: 1,
matrix name: PP,
filler name: none,
composition: {amount: 0.0%, type: wt},
particle surface treatment name": not specified,

Accuracy scores of composition information extraction

Input type/Including missing samples	no	yes
Image	0.917 ± 0.036	0.910 ± 0.037
OCR	0.890 ± 0.065	0.790 ± 0.107
Structured Format (with captions)	0.948 ± 0.032	0.816 ± 0.113
Structured Format (without captions)	0.890 ± 0.056	0.832 ± 0.089

Property and condition level evaluation

Ground Truth

properties: {
Example Property Identifier: {value: 910, unit: MPa, conditions: [{type: temperatu -413, unit: K}],
Another Key Name: {...}
}

property match

Predictions

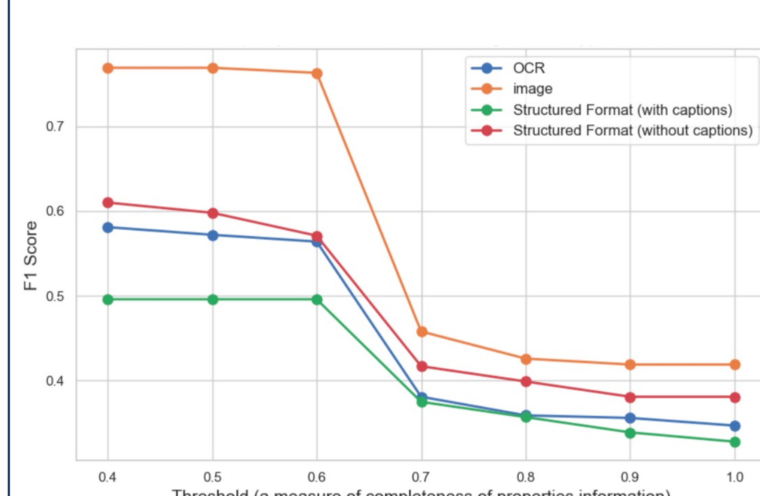
properties: {
Not Close Property Name: {...},
Example Property Name : {value: 910, unit: MPa, conditions: [[type: temperature, value: -413unit: K}],
}

F1 scores of property name information extraction

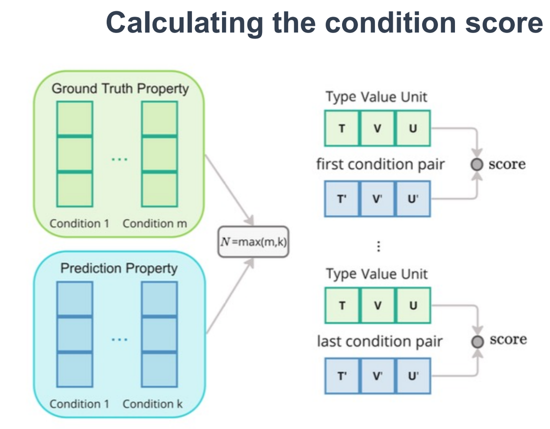
Input type/Including missing samples	no	yes
Image	0.869 ± 0.078	0.863 ± 0.078
OCR	0.766 ± 0.104	0.666 ± 0.117
Structured Format (with captions)	0.795 ± 0.107	0.682 ± 0.129
Structured Format (without captions)	0.617 ± 0.133	0.576 ± 0.134

F1 scores of property information considering value, unit and condition

- Calculated a matching score for each of the entities. The final score for a property is an average of these individual scores. Equality check is used for values and units



Calculating the condition score



$$\text{score} = \frac{\mathbb{1}_{[T=T']} + \mathbb{1}_{[V=V']} + \mathbb{1}_{[U=U']}}{3}$$

$$\text{condition score} = \frac{\sum \text{scores}}{N}$$

Findings

- Multimodal model with an image output yielded the most promising results.
- We introduced a flexible evaluation technique tailored to assess the accuracy and efficiency of these extraction methods, contributing to a nuanced understanding of their performance on this complex task.

Future work

- **Granular benchmarking** across entity and relationship types
- **Benchmarking** across commercial and open-source models
- Extracting sample information from **tables, figures and text**
- Scaling complex extraction and verification to **various materials domains**

This material is based upon work supported by the National Science Foundation under Grant No. DGE-2022040

defne.circi@duke.edu, ghazal.khalighinejad@duke.edu, anlan.chen@duke.edu
bdhingra@cs.duke.edu, cate.brinson@duke.edu