香港中文大學
The Chinese University of Hong Kong

Journal Club

https://doi.org/10.1038/s43246-024-00708-9

Data extraction from polymer literature using large language models

Sonakshi Gupta[1,4], Akhlak Mahmood[2,4], Pranav Shetty[1], Aishat Adeboye[3] & Rampi Ramprasad[2]

Zihan Li

23rd May 2025

Commun. Mater. **2024**, *5*, 269.

Content

1. Background for Research
2. Results and Discussion
3. Outlook

1. Background for Research

Challenge 1

data frequently exists in **unstructured natural language** text formats,

> We synthesized a haloalkane from an alkene at 50 °C.

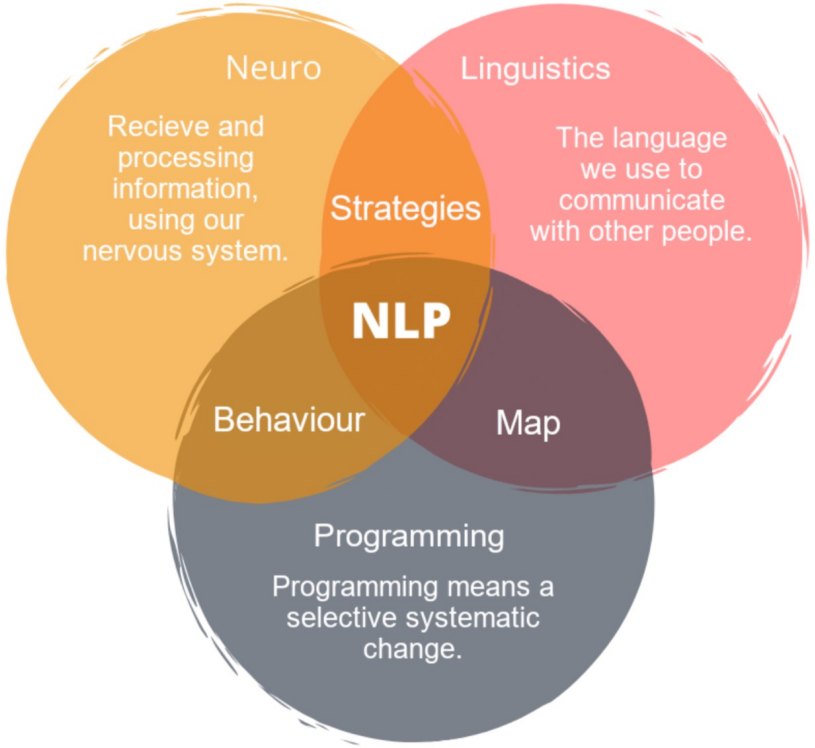
hard to be utilized by modern informatics that rely on **structured datasets**

> Temp = 50 °C
> Starting materials = alkene
> Product = haloalkane

Solution 1

Natural language processing (NLP)

Make computers understand human language

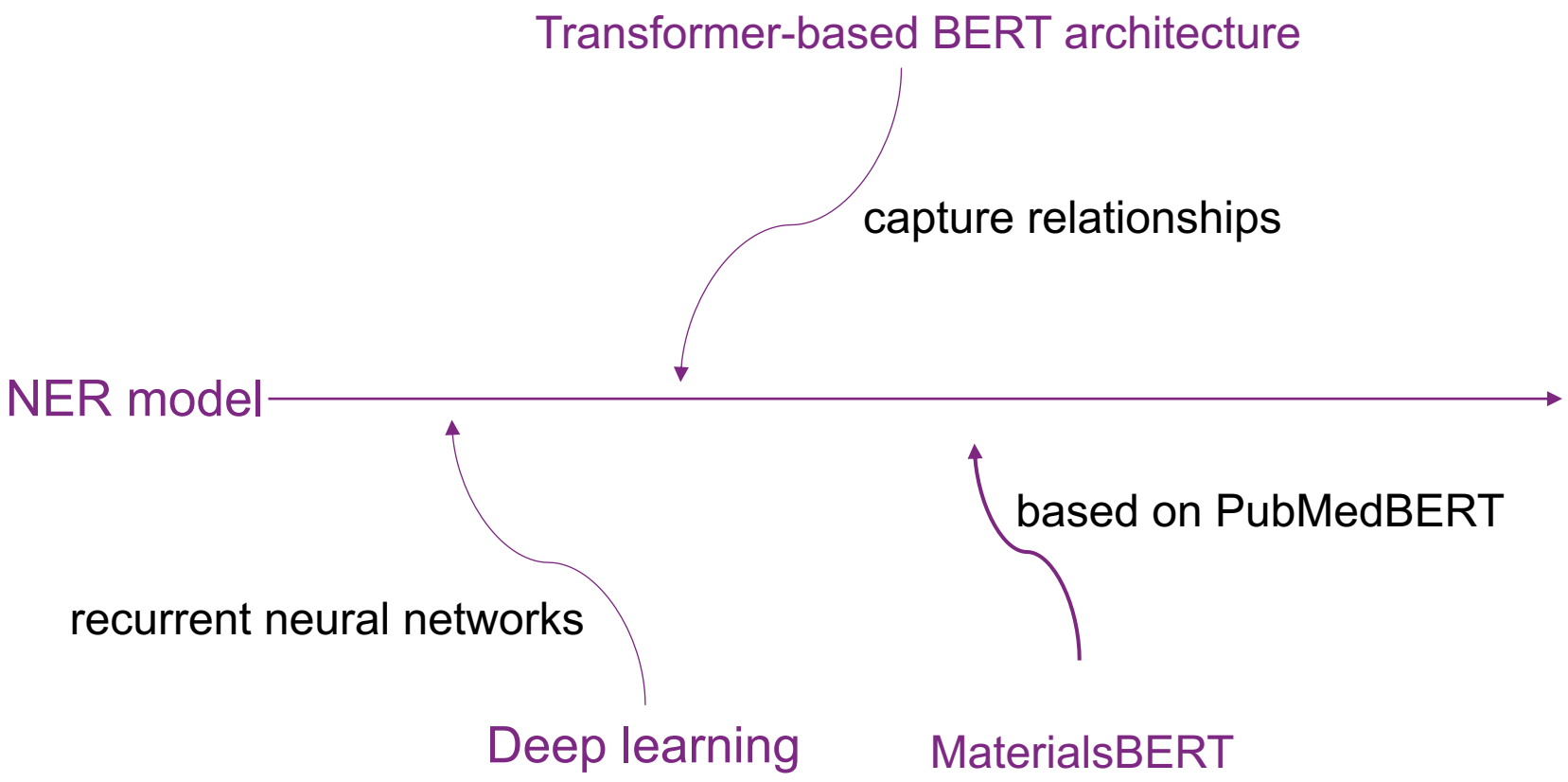


Corpus

- Turn texts to units (tokens)
- **Named Entity Recognition (NER)**
  Identify relationships
- Extracting information

Structured format

Research Progress for NLP



Challenge 2

- Discern entity relationships
- Non-standard names
- Complex expressions

Solution 2

Large language models (LLMs)

eg. Generative Pretrained Transformer (GPT) and Large Language Model Meta AI (LlaMa)

Advantages:

1. Pre-training to acquire better comprehension
2. Supervised fine-tuning to produce desired outputs

Disadvantages:

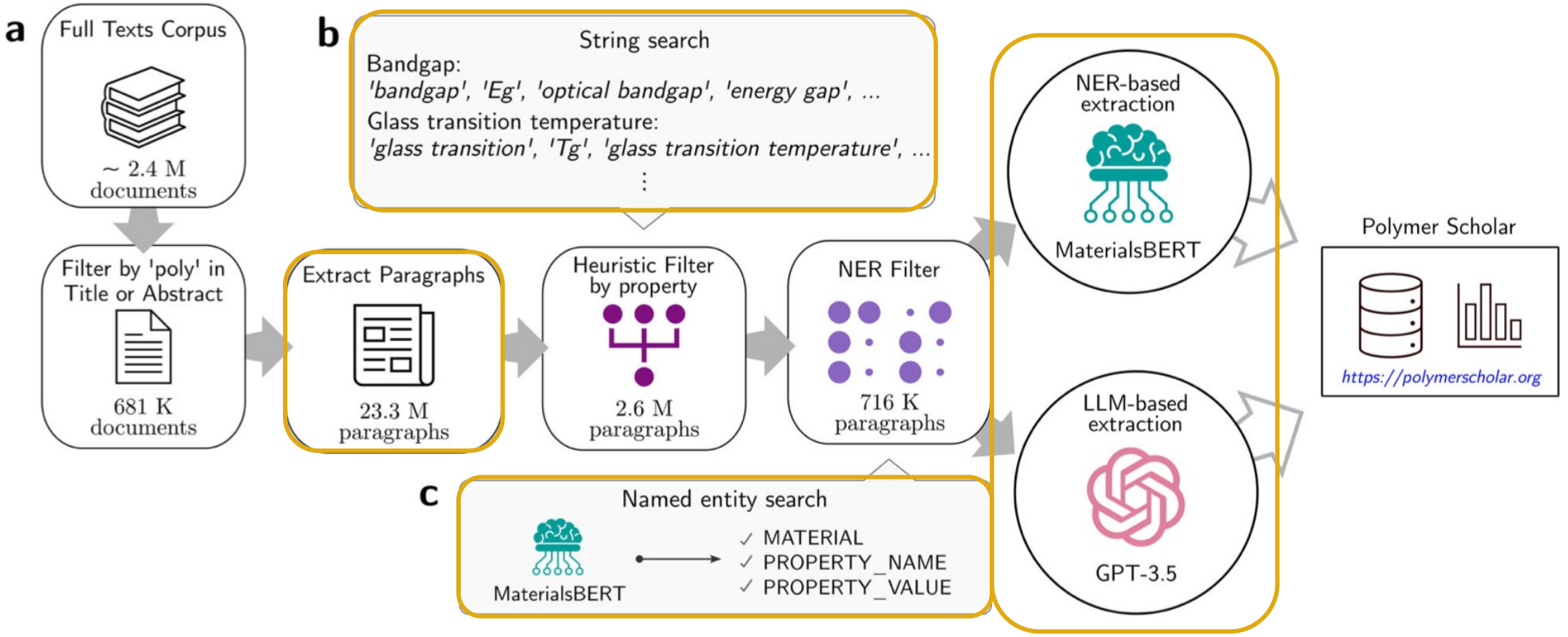
1. Depend on significant computational resources

Extract high-quality information

Reduce unnecessary model prompting

This Work

LLM- and NER-based pipelines for *property data of polymers*



2. Results and Discussion

Data extraction using large language models

manually pick keywords

↓

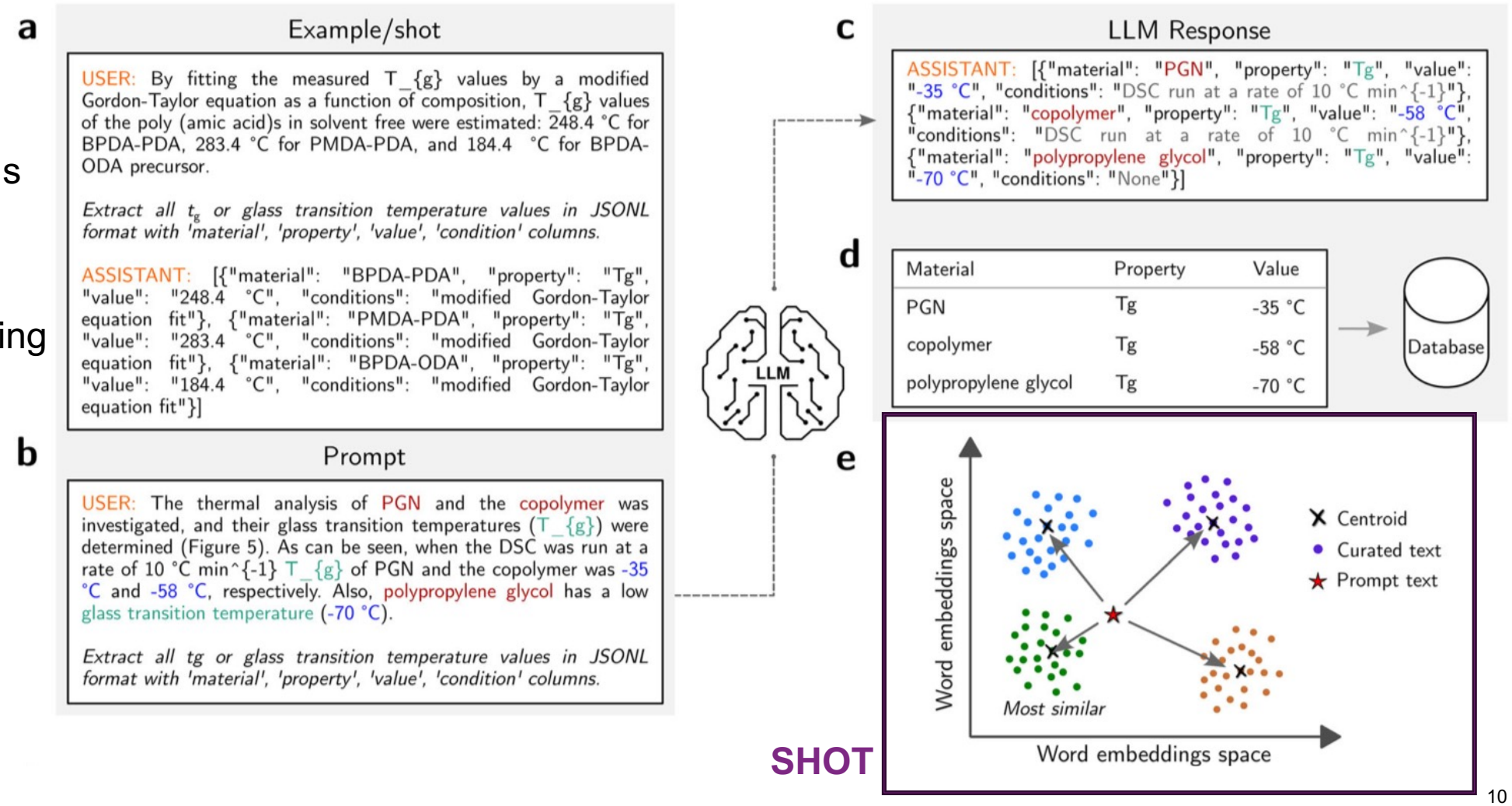
in-context few-shot learning

↓

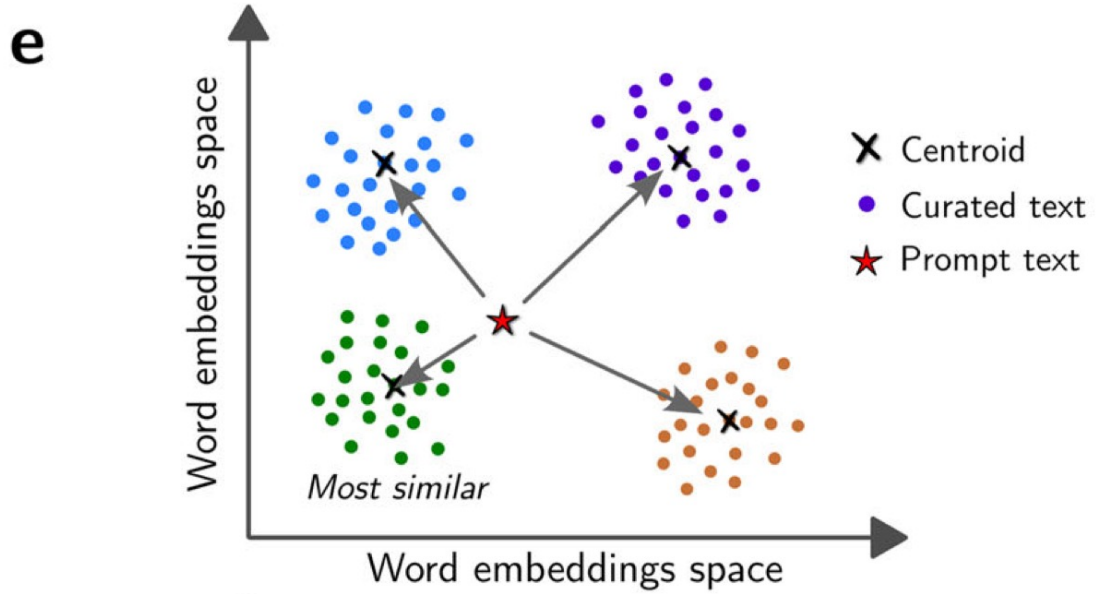
GPT-3.5 response

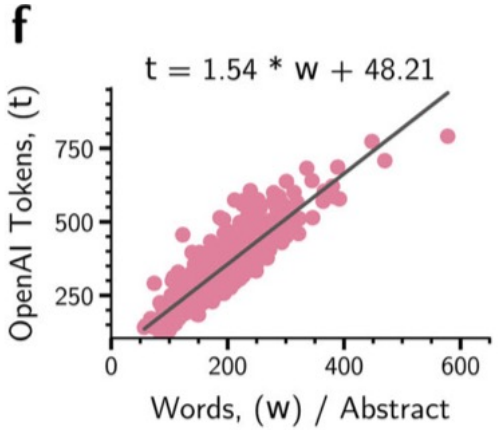
↓

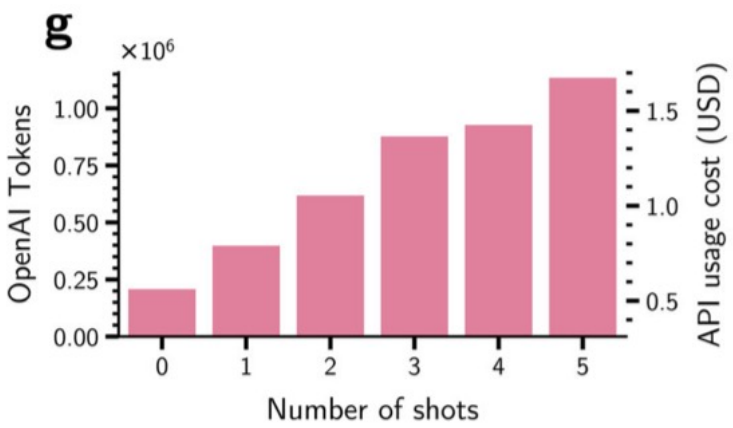
database

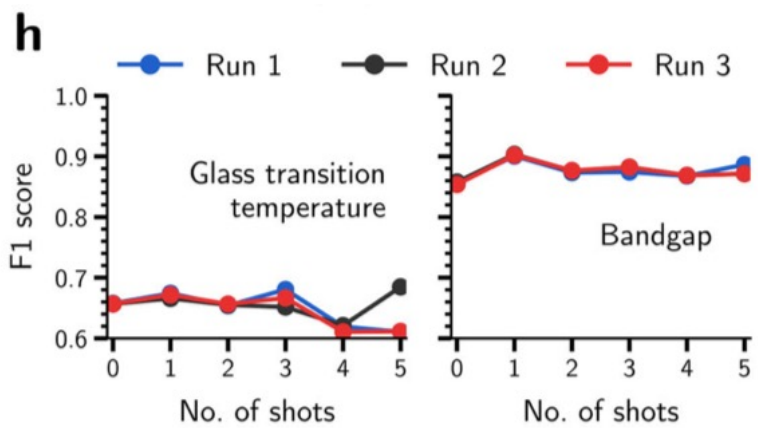


Data extraction using large language models

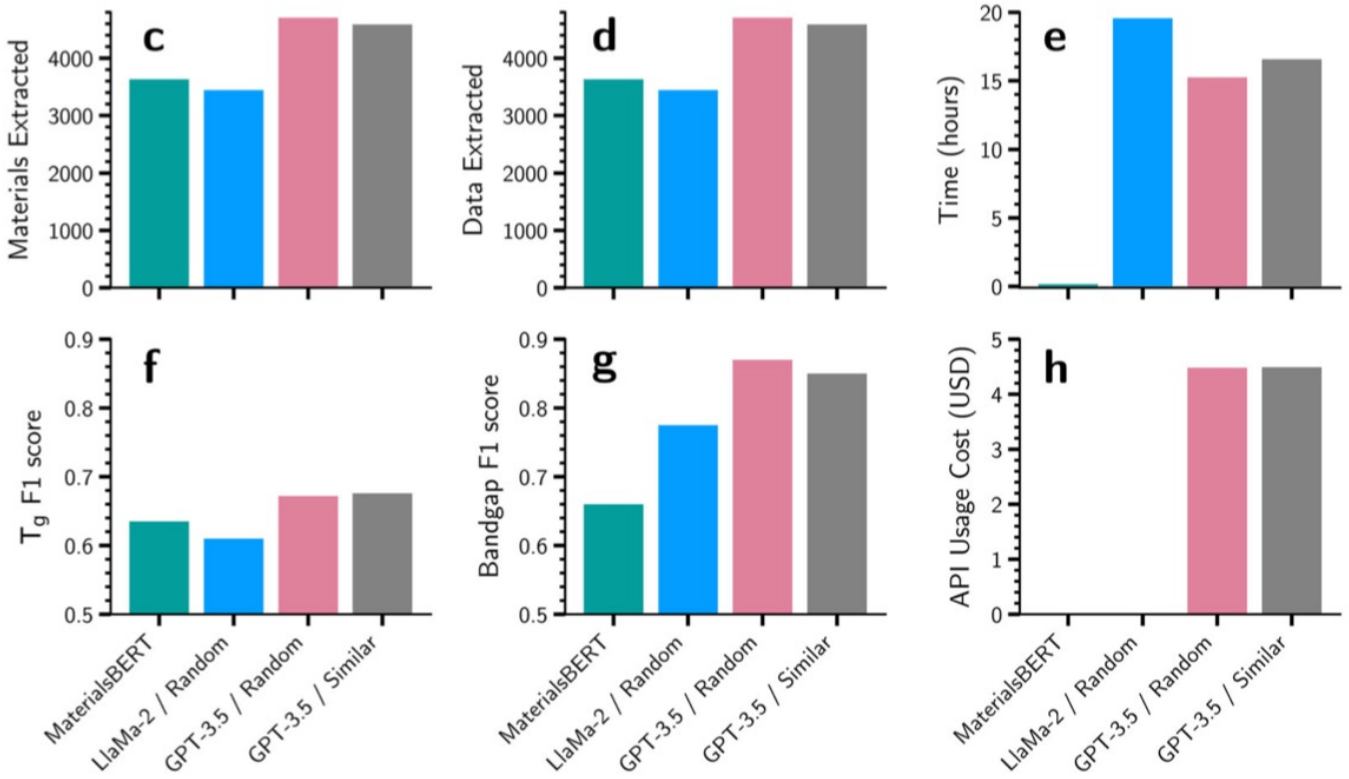


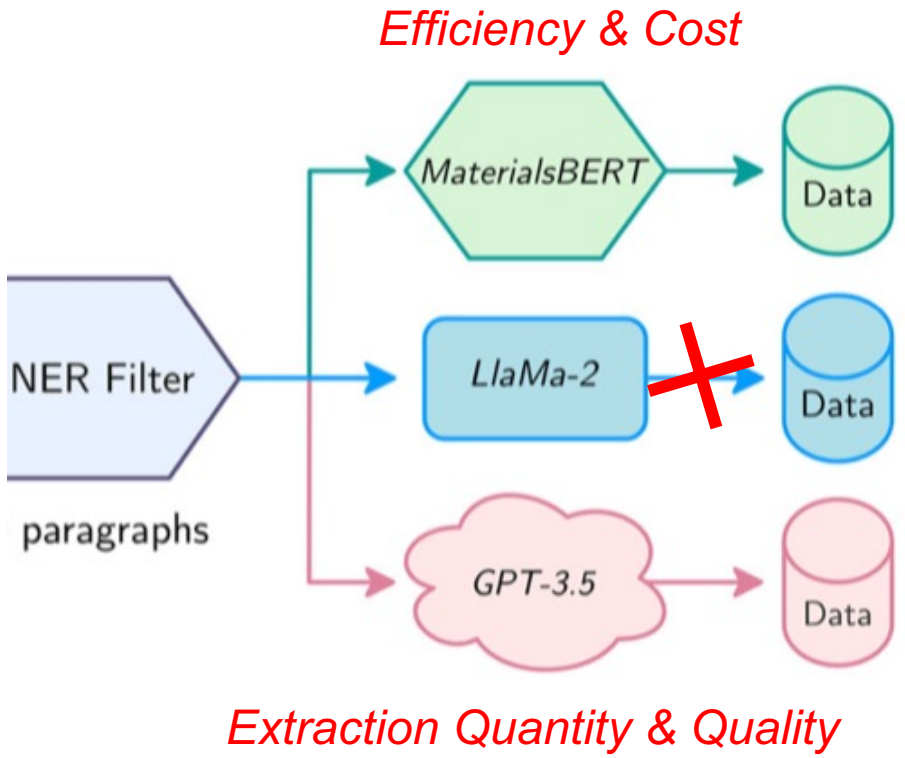




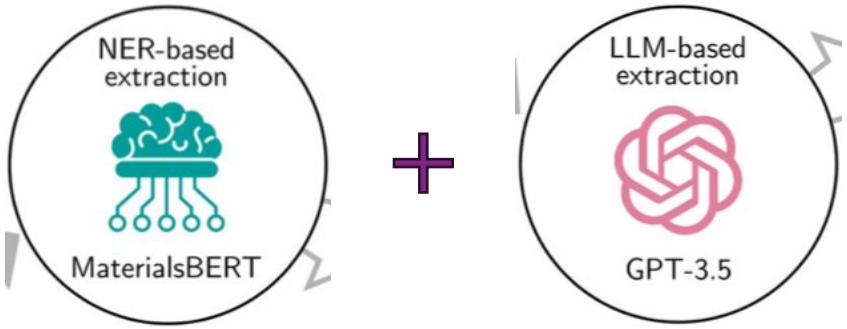


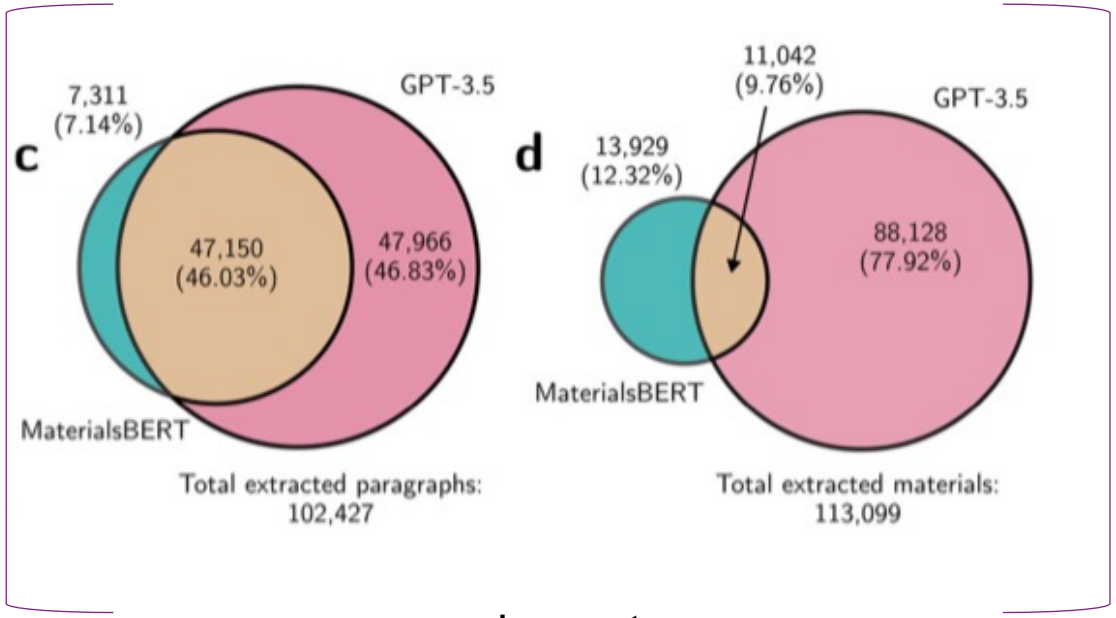
Performance benchmarking for a labeled subset of the full corpus





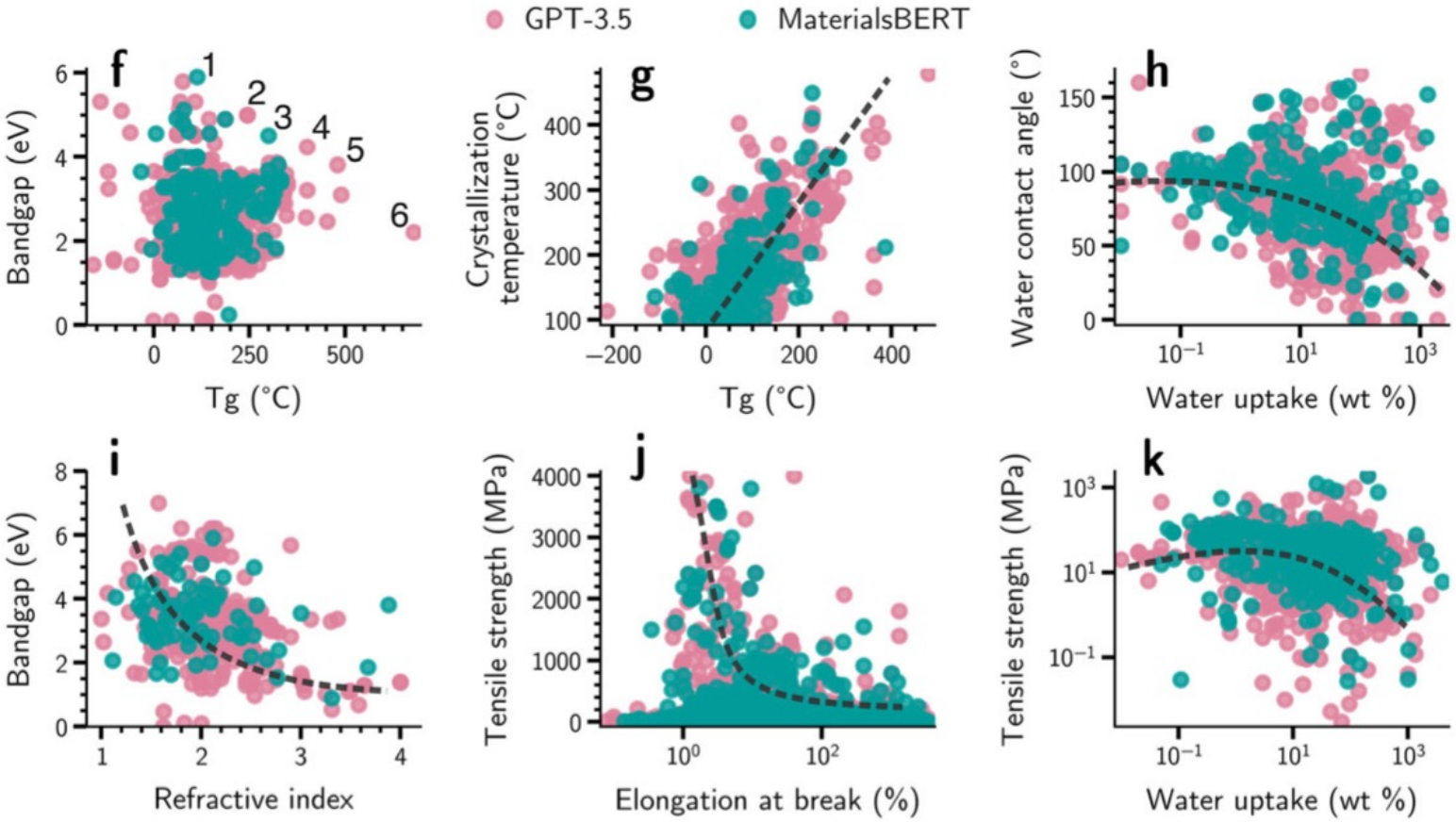
Data extraction from full texts





complementary

Correlations between extracted properties



For further tuning

For prediction

For double-check

3. Outlook

Outlook

A framework for automated extraction of polymer property data from full-text articles using large language models GPT-3.5 and NER-based model MaterialsBERT. From a corpus of 2.4 million articles, the method identified 681k polymer-related papers and extracted over 1 million records covering 24 properties of 106k unique polymers.

- Identify and extract polymer in figures via vision models + LLMs
- Difficulty in establishing cross-paragraph entity relationships
- Data in tables and supplementary materials remains difficult to extract
- The extraction of procedural tasks (such as synthesis routes)

Thank You