



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

Data extraction from polymer literature using large language models

Check for updates

Sonakshi Gupta^{1,4}, Akhlak Mahmood ⁰^{2,4}, Pranav Shetty¹, Aishat Adeboye³ & Rampi Ramprasad ⁰²

Zihan Li

23rd May 2025

Commun. Mater. 2024, 5, 269.



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)





Research Progress for NLP





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

а	Example/shot	С	LLM Response
manually pick keywords	USER: By fitting the measured T_{g} values by a modified Gordon-Taylor equation as a function of composition, T_{g} values of the poly (amic acid)s in solvent free were estimated: 248.4 °C for BPDA-PDA, 283.4 °C for PMDA-PDA, and 184.4 °C for BPDA-ODA precursor.	>	ASSISTANT: [{"material": "PGN", "property": "Tg", "value": "-35 °C", "conditions": "DSC run at a rate of 10 °C min^{-1}"}, {"material": "copolymer", "property": "Tg", "value": "-58 °C", "conditions": "DSC run at a rate of 10 °C min^{-1}"}, {"material": "polypropylene glycol", "property": "Tg", "value": "-70 °C", "conditions": "None"}]
in-context few-shot learning	Extract all t _g or glass transition temperature values in JSONL format with 'material', 'property', 'value', 'condition' columns. ASSISTANT: [{"material": "BPDA-PDA", "property": "Tg", "value": "248.4 °C", "conditions": "modified Gordon-Taylor equation fit"}, {"material": "PMDA-PDA", "property": "Tg", "value": "283.4 °C", "conditions": "modified Gordon-Taylor equation fit"}, {"material": "BPDA-ODA", "property": "Tg", "value": "184.4 °C", "conditions": "modified Gordon-Taylor equation fit"}]		Material Property Value PGN Tg -35 °C copolymer Tg -58 °C polypropylene glycol Tg -70 °C
b GPT-3.5 response database	Prompt USER: The thermal analysis of PGN and the copolymer was investigated, and their glass transition temperatures (T_{g}) were determined (Figure 5). As can be seen, when the DSC was run at a rate of 10 °C min^{-1} T_{g} of PGN and the copolymer was -35 °C and -58 °C, respectively. Also, polypropylene glycol has a low glass transition temperature (-70 °C). Extract all tg or glass transition temperature values in JSONL format with 'material', 'property', 'value', 'condition' columns.	e	Most similar
		SHOT	Word embeddings space





11



Performance benchmarking for a labeled subset of the full corpus









Correlations between extracted properties





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