Extracting accurate data from research papers

Using Large Language Models (LLMs)

Maciej P. Polak, Dane Morgan
Department of Materials Science and Engineering
University of Wisconsin - Madison

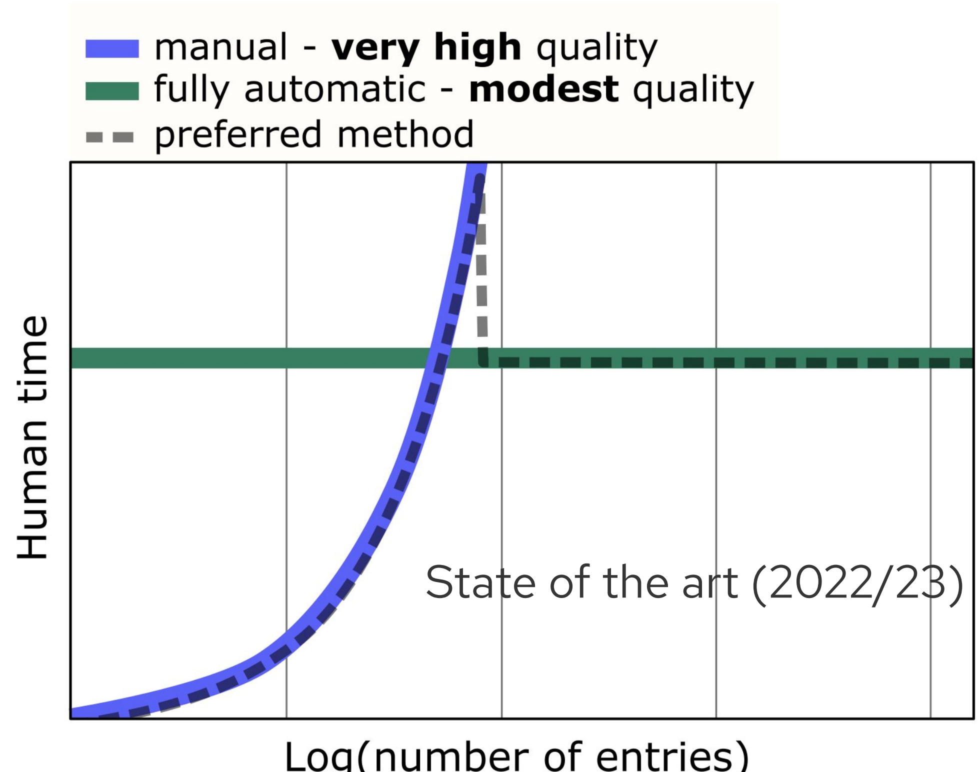
June 20th 2024

Our approach to data extraction

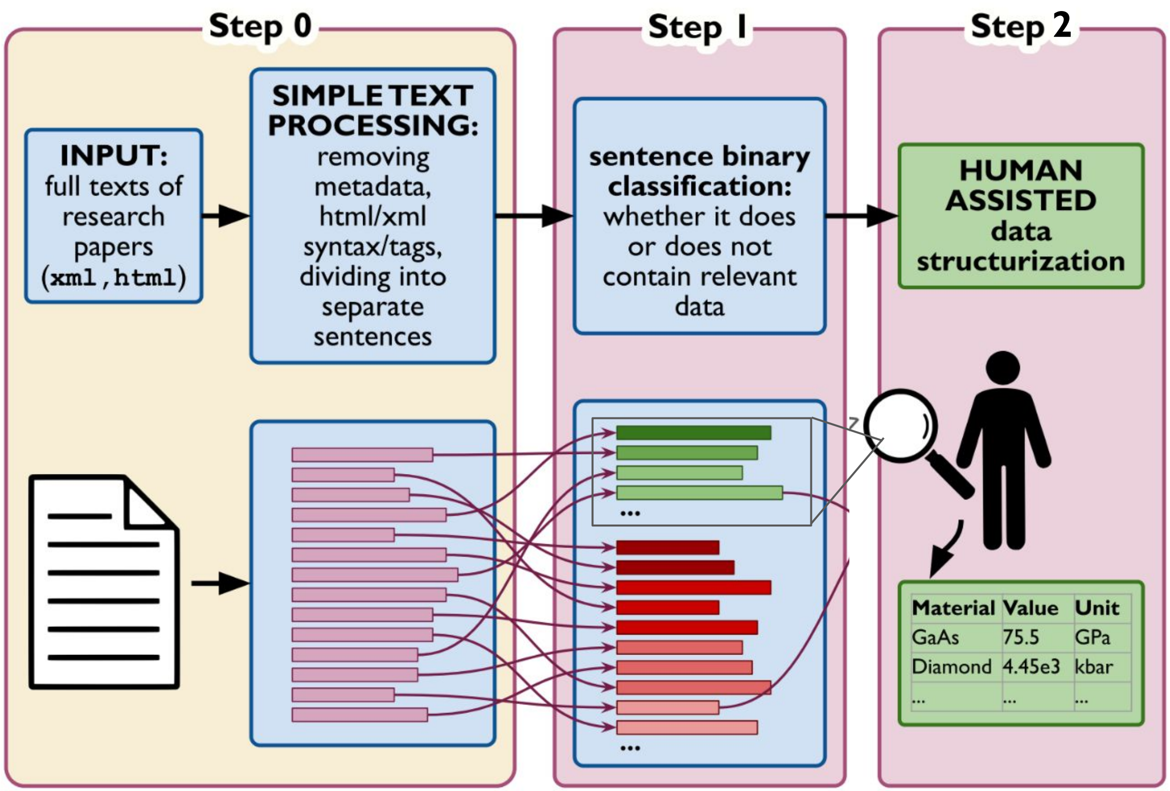
- Break up papers into sentences and **classify**
  - *Does it contain relevant data?*
- **Extract** the data
  - By hand
  - Automatically - with **LLMs**
- **Accurate** extraction is possible:
  - ~90% precision and recall

Our approach to data extraction

- **Problem:** high quality data needed for building machine learning models
- **Solution:** extract data from research papers
  - **Problems:**
    - low quality of auto-mated data extraction
    - too many papers to extract manually
  - **Solution:**
    - *use language models to extract data from research papers*



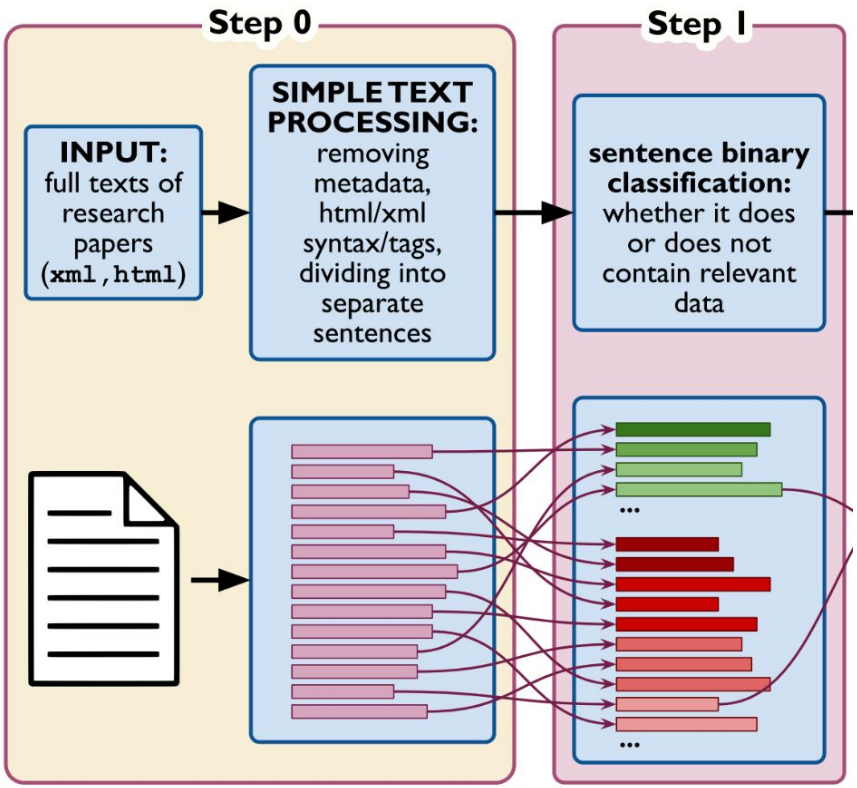
Using (L)LMs for sentence classification

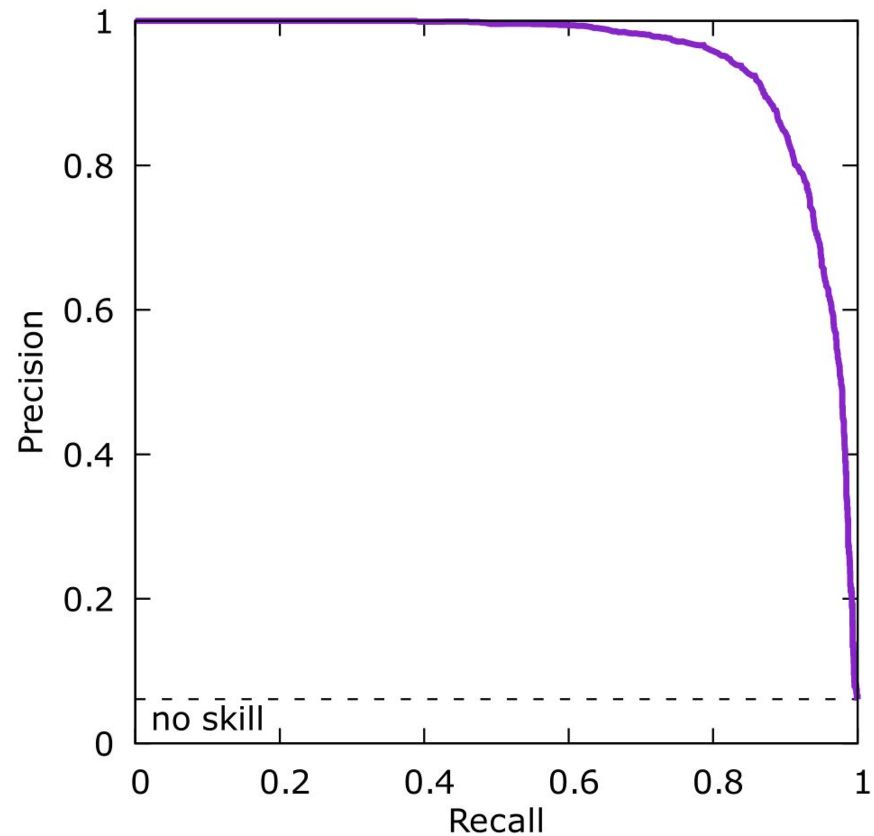


Data triplet:
Material, Value, Unit

Polak, M.P., Morgan, D., et. al., *Digital Discovery* 3, 1221 (2024) https://doi.org/10.1039/D4DD00016A

Using (L)LMs for sentence classification





Polak, M.P., Morgan, D., et. al., *Digital Discovery* 3, 1221 (2024) https://doi.org/10.1039/D4DD00016A

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

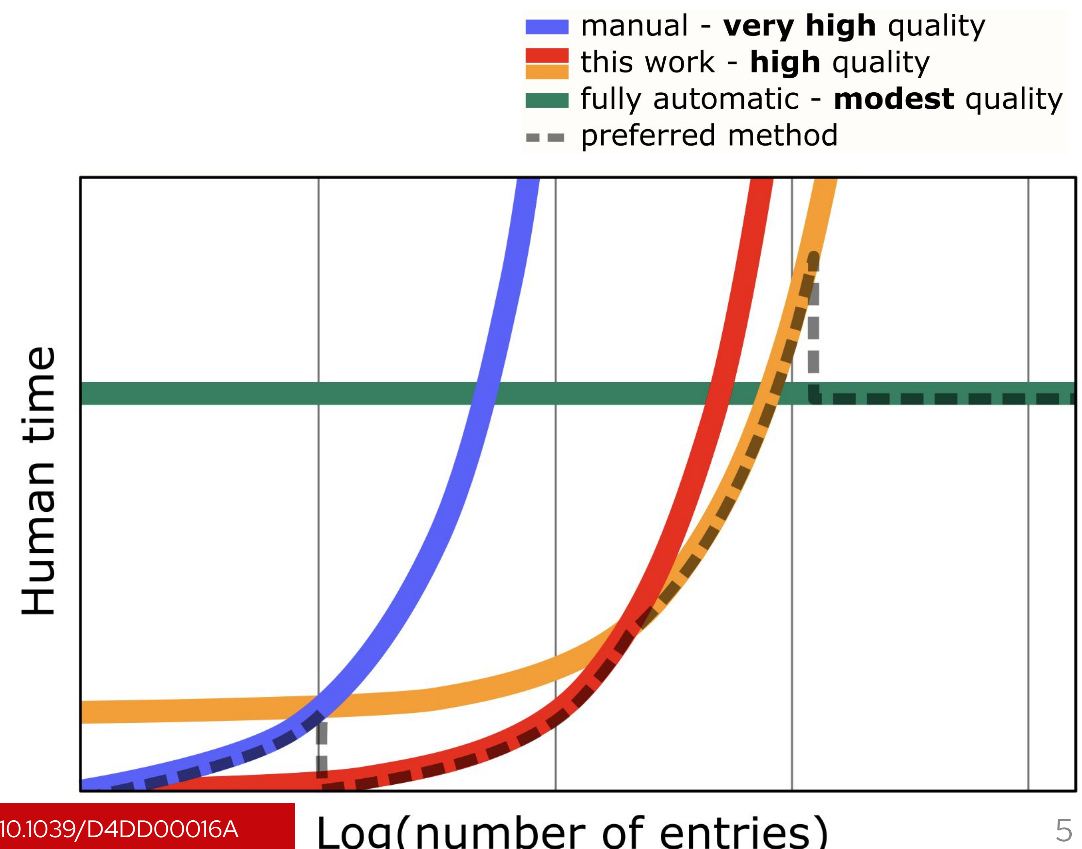
Check for updates

Cite this: *Digital Discovery*, 2024, 3, 1221

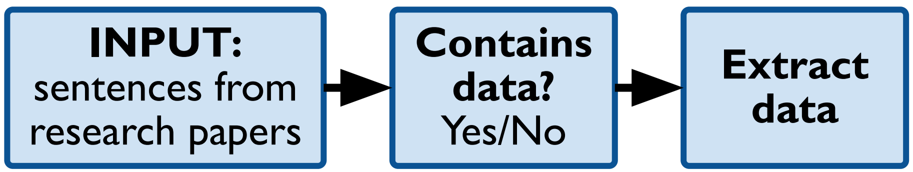
Flexible, model-agnostic method for materials data extraction from text using general purpose language models

Maciej P. Polak, * Shrey Modi, Anna Latosinska, Jinming Zhang, Ching-Wen Wang, Shaonan Wang, Ayan Deep Hazra and Dane Morgan*

- Extremely simple
- Needs minimal resources
- Almost no coding required
- Used to develop the most complete and largest to date database of critical cooling rates of metallic glasses



Automate data structurization with LLMs

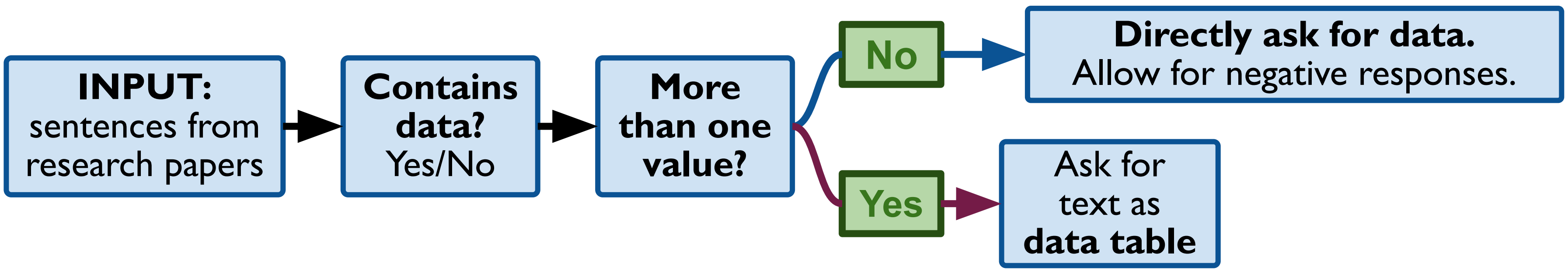


Using **GPT-4**:

- ~90% recall
- ~30% precision

Polak, M.P., Morgan, D., *Nat Commun* 15, 1569 (2024).
https://doi.org/10.1038/s41467-024-45914-8

Automate data structurization with LLMs

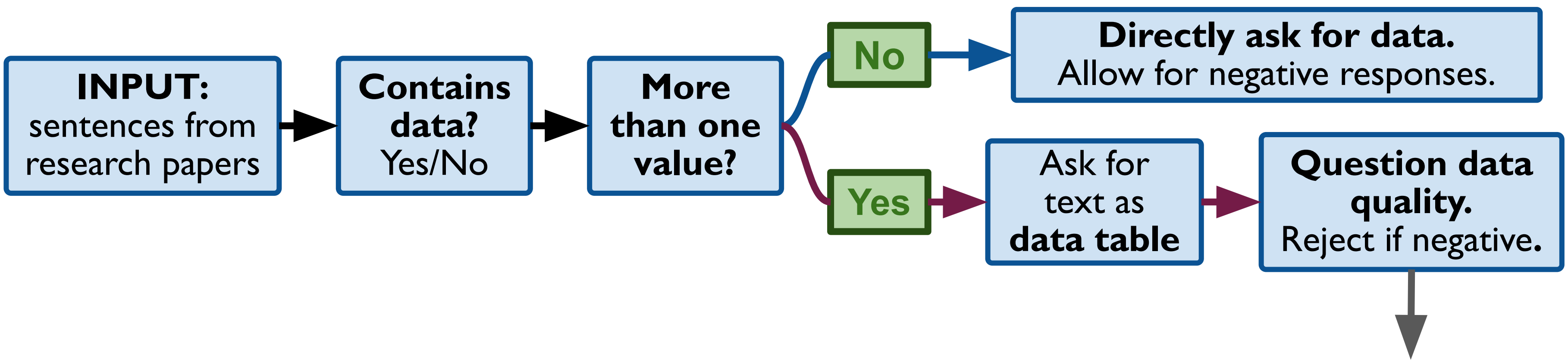


Using **GPT-4**:

- ~98% recall
- ~42% precision

Polak, M.P., Morgan, D., *Nat Commun* 15, 1569 (2024).
https://doi.org/10.1038/s41467-024-45914-8

Automate data structurization with LLMs



Using **GPT-4**:

- ~88% recall
- ~91% precision

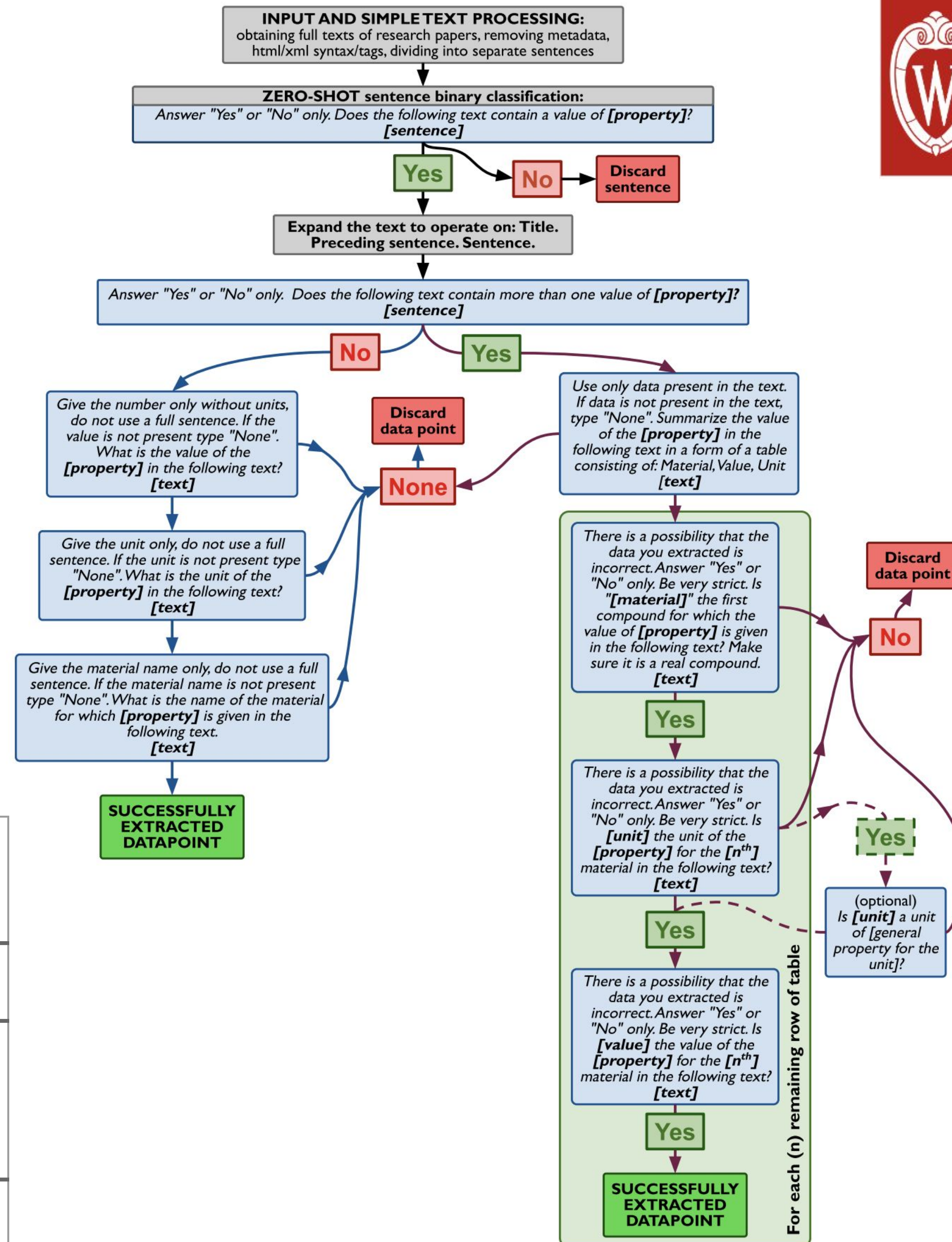
Error detection through follow-up questions:
- redundancy
- introducing doubt

Polak, M.P., Morgan, D., *Nat Commun* 15, 1569 (2024).
https://doi.org/10.1038/s41467-024-45914-8

ChatExtract

a workflow consisting of a series of engineered conversational prompts to a large language model (LLM), and actions based on the model's responses, with error detection.

	Precision (%)	Recall (%)
ChatExtract (GPT4)	**90.8**	**87.7**
Chain-of-thought (GPT4) (ChatExtract without error-detection)	42.7	98.9
Previous non LLMs	<50%	<50%



Polak, M.P., Morgan, D., *Nat Commun* 15, 1569 (2024).
https://doi.org/10.1038/s41467-024-45914-8

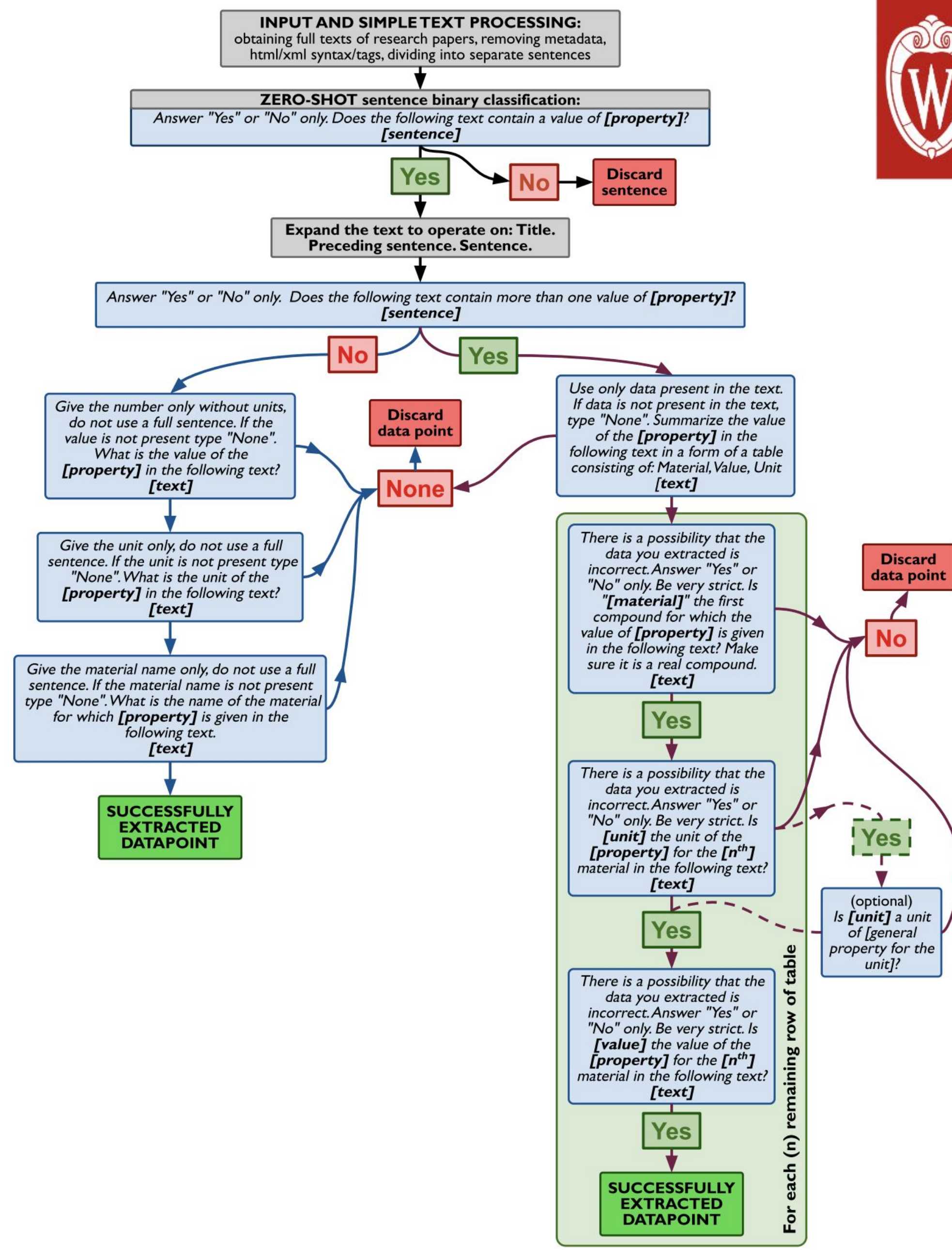
ChatExtract

Key concepts:

- Break up tasks into the simplest possible steps
- Strictly enforce output format
- Allow for negative answers
- When validating - be redundant, introduce doubt

Benefits of our approach:

- Flexible and adaptable
- No need for deep understanding of data to be extracted



Polak, M.P., Morgan, D., *Nat Commun* 15, 1569 (2024).
https://doi.org/10.1038/s41467-024-45914-8

nature communications

Article

https://doi.org/10.1038/s41467-024-45914-8

Extracting accurate materials data from research papers with conversational language models and prompt engineering

Received: 27 June 2023

Accepted: 5 February 2024

Maciej P. Polak [1] & Dane Morgan [1]

Metallic Glass Critical Cooling Rates

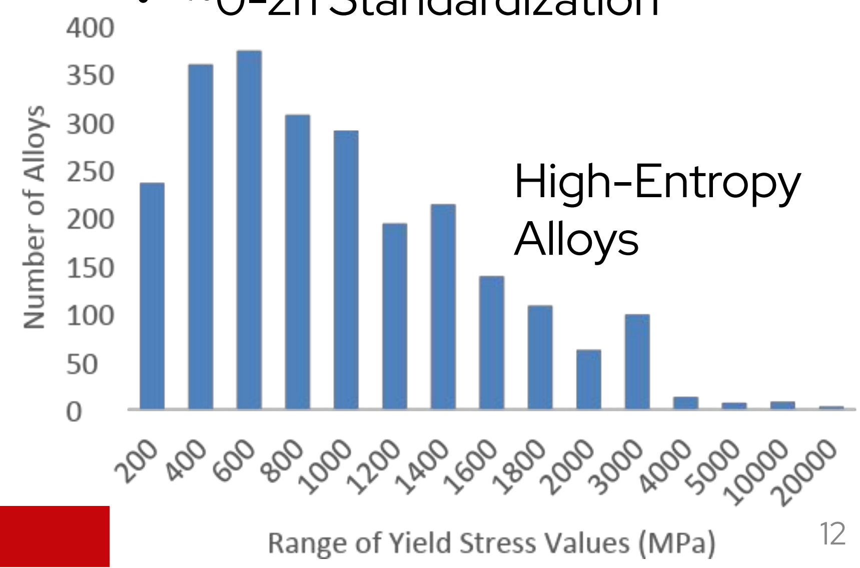
- Returned 684 papers, 110,126 sentences.
- ChatExtract (GPT-4) extracted 721 values.
- Standardized version had 280 values (120 unique compounds), 1.5x previous databases.

High-Entropy Alloy Yield Stress

- 4029 research papers, 840,431 sentences.
- ChatExtract extracted 8,961 values.
- Standardized version had 2416 values.
- Largest database up to date

Performance

- Precision and Recall ~90%
- ~10m human time
- ~3h compute time (ChatGPT)
- ~0-2h Standardization



Polak, M.P., Morgan, D., *Nat Commun* 15, 1569 (2024).
https://doi.org/10.1038/s41467-024-45914-8