MSR135

Using large language models (LLMs) for data extraction in literature reviews: an enhanced approach

Alexandrina Lambova[1], Kiril Matev[1], Julia Gallinaro[2], Ines Guerra[2], Ketevan Rtveladze[2], Suzanne Caverly[2]*

*[1]IQVIA, Sofia, Bulgaria, [2]IQVIA, London, UK, *presenting author*

BACKGROUND AND OBJECTIVE

- Systematic literature reviews (SLRs) are pivotal in making market access decisions regarding novel medical products. However, data extraction of clinical evidence for health technology assessment dossiers remains labour-intensive and error-prone.
- Last year, we developed a generative pre-trained transformer 4 (GPT-4)-based algorithm, suggesting that artificial intelligence (AI) large language models (LLMs) such as GPT[1,2], in conjunction with iterative algorithm engineering, could be used for generating the first version of the extraction file with good accuracy.
- We developed a new LLM-based multistep approach to overcome some of the challenges with complex clinical data extraction.
- The objective of the project was to enhance the previous algorithm, particularly for complex variables with historically low accuracy rates.

PREVIOUS APPROACH USING GPT-4 (2023)

Methods

- Last year, we developed a GPT-4-based algorithm that demonstrated the potential of LLMs for generating initial extraction from publications of clinical data (**Figure 1**).
- We assessed performance by measuring the accuracy of the GPT-based extraction compared to a manual extraction performed by humans (benchmark) and with a human quality check (QC) performed on the GPT outputs.

Results

- The extraction of clinical data, which included study details, patient characteristics, safety, efficacy, and quality of life outcomes, achieved an accuracy range of 45% to 83% using the GPT-4-based algorithm, with the highest accuracy in study details and the lowest in patient characteristics (**Figure 2**).

Motivation

- Low accuracy was found for variables present in tables or figures and variable extraction for specific subgroups. Current advancements in technology are expected to improve some of these aspects.

ENHANCED APPROACH (2024)

Methods

- We developed a new LLM-based multistep approach to overcome some of the challenges with complex clinical data extraction, such as subgroup variable extraction, long paper processing, and structured format generation (**Figure 3**).
- We leveraged LLM retrievers, an embedding model, and GPT-4 to extract relevant information for the variables in an unstructured format.
- Iterative prompt engineering, guided by subject matter experts, refined the information retrieval process. A combination of an LLM-based method and standard programming techniques was used to construct predefined extraction tables from the text.
- We measured the accuracy of the algorithm for 35 patient characteristic variables and 20 intervention details variables across thirteen studies by comparing the generated extraction to a manual extraction performed by humans.

Results

- The results showed an average accuracy of 81% for patient characteristics and 78% for intervention details (**Figure 4**).
- Notably, both the patient characteristic extraction and the intervention details extraction significantly improved compared to the previous results (patient characteristics: 45%, intervention details: 49%).
- The studies contain between one and five subgroups of interest, including historical controls. The algorithm identified all relevant subgroups correctly in 10 of 13 studies (**Figure 5**).

Conclusion

- Implementing a complex multistep approach enhanced LLM-based clinical data extraction.
- Independent improvements at each step contributed to overall precision.
- Our algorithm demonstrated promising results, paving the way for efficient clinical data extraction, even for complex variables and population subgroups.

SUMMARY AND CONCLUSION

The results suggest that the complex multistep approach enhanced LLM-based clinical data extraction and could be used for generating the first version of the extraction file with higher accuracy than the previous approach.

The current advancements in technology are expected to pave the way for efficient clinical data extraction, even for complex variables and population subgroups.

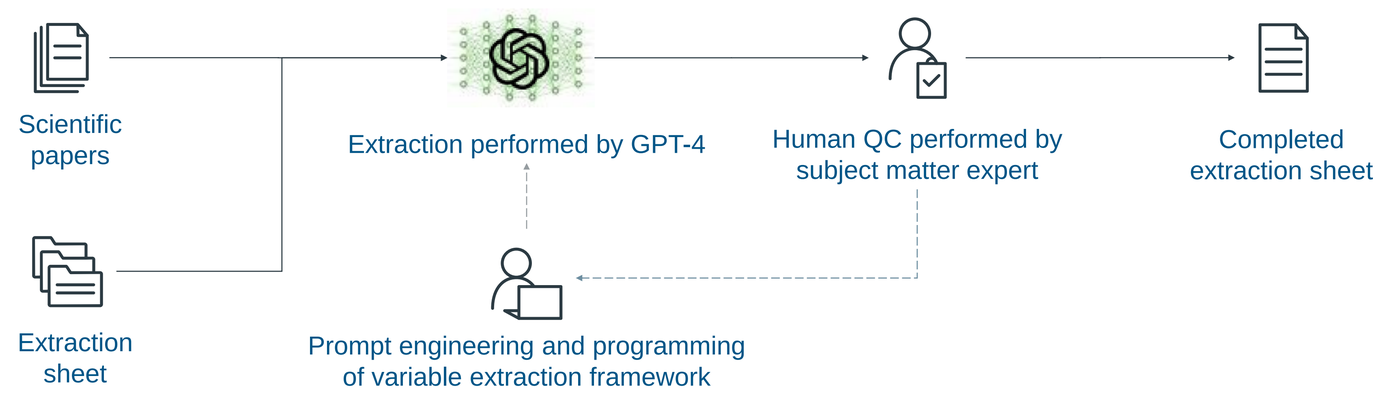
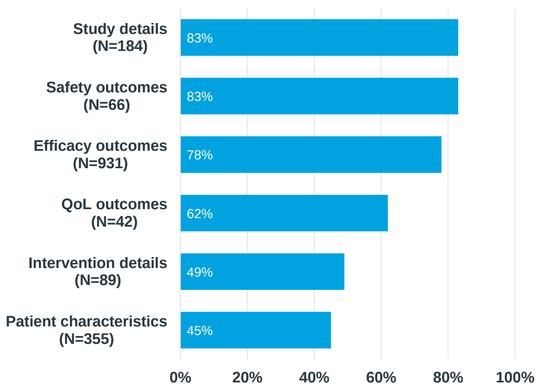


Figure 1: Overview of steps for extracting variables from scientific papers using GPT-4 (2023). Dotted grey lines indicate steps to be performed only during optimisation of the algorithm. Solid dark lines indicate steps performed during optimisation and application of the optimised algorithm for extraction.



High performance	
Variable	**Accuracy**
Unit of follow up duration	30 in 30
Description of event free survival	7 in 7
Definition of FLT3-ITD positive	8 in 8
Data cut-off	5 in 5
Number of patients assigned to treatment	11 in 12

Low performance	
Variable	**Accuracy**
Intervention dosing frequency	5 in 9
Route of administration	3 in 6
Median of event free survival duration	1 in 4
Number of patients with complete response	2 in 8
Definition of transplantation rate	1 in 8

Figure 2: Accuracy of GPT-4 extraction based on human QC (2023). Accuracy is given as percentage of variables that are not NA (not applicable) / NR (not reported) and that were assigned as correct during QC. *Left, chart:* Average accuracy across all papers and all variables by topic (accuracy) and number of not NA/NR variables across all papers per topic (N). *Right, tables:* Example of variables with high and low accuracy.

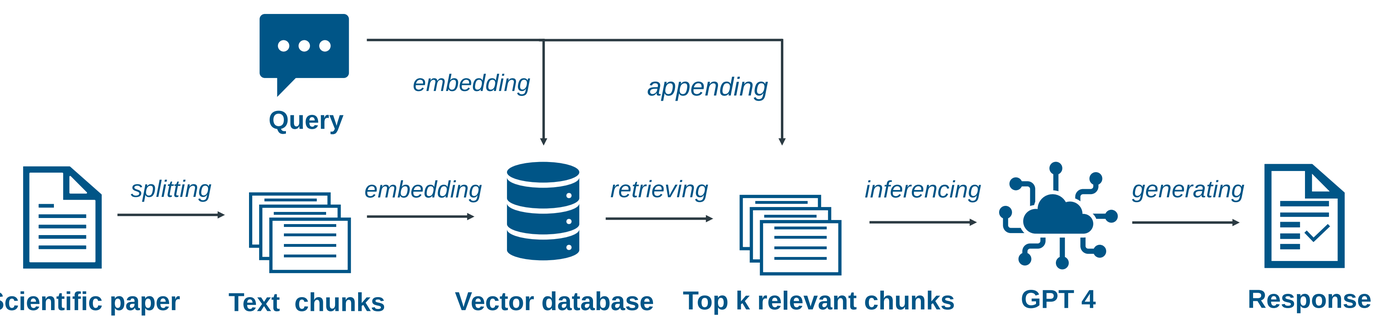
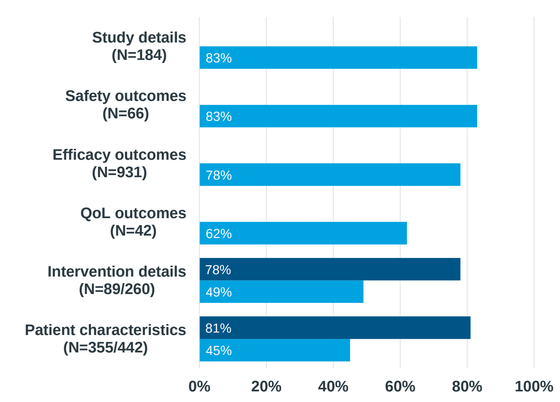


Figure 3: Overview of steps for extracting variables from scientific papers using enhanced approach: The pipeline shows the steps needed for pre-processing of a paper and extraction of one query from it. One query can correspond to more than one variable. Additional steps are performed after that to consolidate all variables in one final extraction table.



High performance	
Variable	**Accuracy**
Age in years (median)	23 in 25
Number of male patients	29 in 29
Number of female patients	24 in 28
Route of administration	18 in 19
Number of patients assigned to the group	34 in 34

Low performance	
Variable	**Accuracy**
Treatment duration timescale	3 in 5
Days of the cycle where the treatment is administered	7 in 16
Number of patients Ongoing / Continued treatment	3 in 6
Number of patients in survival follow-up	0 in 4
Age in years (Standard Deviation)	0 in 4

Figure 4: Accuracy of enhanced LLM-based extraction of "patient characteristic" and "intervention details" variables based on human QC: Accuracy is given as percentage of variables that are not NA (not applicable) / NR (not reported) and that were assigned as correct during QC. *Left, chart:* Average accuracy across all papers and all variables by topic and number of not NA/NR variables across all papers per topic (N). Light (dark) blue bars show the performance of the previous (enhanced) approach. *Right, tables:* Example of variables with high and low accuracy using the enhanced approach.

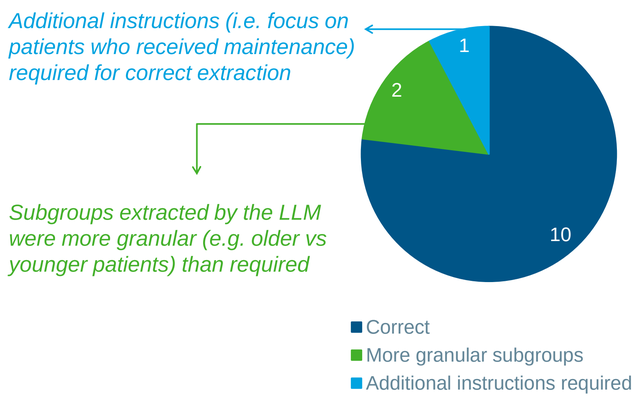


Figure 5: Performance of subgroup extraction: Number of papers for which subgroup extraction was correct, was more granular than expected or required additional information for correct extraction

References

1. Radford A, Narasimhan K, Salimans T, Sutskever I. "Improving Language Understanding by Generative Pre-Training." (2018).
2. Brown TB, Mann B, Ryder N, Subbiah M et al. "Language Models are Few-Shot Learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

ISPOR Europe 2024. 17-20 November 2024. Barcelona, Spain.