Key Considerations in the Use of Large Language Models for Data Extraction in Health Economics and Outcomes Research

MSR88

Elise Aronitz, MSc[1]; Jayson Brian Habib, MPH[1]; Christopher Olsen, BHSc[1]; Kevin Hou, PhD[1]; Nicole Ferko, MSc[1] • [1]EVERSANA, Burlington, ON, Canada.

Background and Objectives

- Applying large language models (LLMs) to improve data extraction is a relatively unexplored area. While LLMs have proven effective in making article screening more efficient and accurate, their potential to significantly enhance the data extraction process is still uncertain, especially in health economics and outcomes research.
- This research assessed several critical factors that affect how well LLMs can extract data proficiently. These factors include optical character recognition (OCR), redaction, table orientation, data stratification, and document format. By studying these elements, the research aims to identify the challenges and opportunities of using LLMs for data extraction.
- By exploring theses nuances in data extraction, the study provides valuable insights into the best practices and strategies to maximize the use and effectiveness of LLMs.
- This study contributes to a broader understanding of how artificial intelligence (AI) technologies can be leveraged to enhance research efficiency and accuracy.
- The findings aim to inform the implementation of LLMs for data extraction, laying the groundwork for developing efficient and reliable methodologies for their deployment.

Methods

- Articles of interest were identified based on the presentation of their data, with primary interest in simple tables (**Table 1**), complex tables (**Table 2**), or nested tables (**Table 3**).
- Using an application developed in R, articles were delivered to GPT[a] via an application programming interface (API) key.
- Since the developed application lacked image upload abilities, the OpenAI playground was utilized to extract data from images.

Accuracy Assessment

- Following extraction, a researcher checked the LLM output against the original articles. This review focused on assessing accuracy and allowed the identification of factors that might affect the extraction process.

Prompting

- Prompts were engineered to guide the LLM in extracting relevant data. These prompts underwent multiple iterations to refine them until the desired functionality was achieved. To ensure consistency, the same prompts were used in both the application and the OpenAI playground.

Pre-Processing

- When needed, documents uploaded to the LLMs were pre-processed using OCR or redaction. OCR processing was done within the application using the pdf_ocr_text function from the pdftools package, while redaction was done manually.

Results

- Based on current studies, multiple factors influence the ability of LLMs to accurately extract data, including document pre-processing, complexity of results, document format, and table orientation.

Document Pre-processing

- Pre-processing documents with OCR and redaction significantly enhances the accuracy of data extraction using LLMs.
- OCR is essential for converting images of text into plain text, enabling LLMs to extract data, which would otherwise be inaccessible due to the inability for our tools to read images. This allows LLMs to extract complex tables with near-perfect accuracy when columns are clearly laid out. However, when columns deviate from typical formats and introduce new variables, such as patient numbers not seen in earlier rows, a reduction in accuracy is observed.
- Redaction increases accuracy when extracting information from text as key information is uploaded to the LLM, although human intervention is required to redact the text. Redaction also improves accuracy when extracting from complex and nested tables, allowing the LLM to replicate the table format almost identically in Excel. When complex or nested tables were uploaded into our tools, the LLM accurately extracted the data but encountered difficulties with table labels. For example, when processing a table similar to Table 2, the LLM might repeat the header row name (e.g., smoking status) instead of listing individual statuses. Redaction can resolve this issue, resulting in an Excel output that closely mirrors the table's presentation in the article, with header rows followed by individual rows. This issue is less prevalent with simpler tables.

Table 1: Example of a Simple Table

Patients, n (%)	Total Screened (n=675)	Total Enrolled (n=650)
Disease 1	212 (29.9)	200 (30.8)
Disease 2	188 (26.5)	180 (27.7)
Disease 3	125 (17.6)	110 (16.9)
Disease 4	185 (26.1)	160 (24.6)

Table 2: Example of a Complex Table

Characteristics	Disease 1 (n=4500)	Disease 2 (n=4500)	Disease 3 (n=4500)
Smoking status			
Current smoker	1569 (33.3)	1784 (28.9)	1779 (34.1)
Ex-smoker	1080 (25.6)	1541 (29.3)	938 (24.8)
Non-smoker	1873 (40.0)	1879 (41.1)	2002 (39.5)
Unknown	10 (1.1)	45 (0.7)	23 (1.6)
Sex			
Men	2176 (53.2)	2164 (54.3)	2389 (53.7)
Women	2071 (46.8)	2397 (45.7)	2447 (46.3)
Ethnic group			
White or unknown	4097 (93.2)	3998 (95.0)	3831 (92.8)
Black	113 (4.1)	123 (3.9)	136 (4.2)
Asian	109 (2.5)	65 (2.7)	41 (2.1)
Mixed or other	15 (1.1)	28 (0.9)	26 (0.9)
Comorbidities			
Depression	371 (12.3)	325 (11.2)	481 (13.8)
Asthma	321 (16.0)	271 (13.5)	420 (16.6)

Table 3: Example of a Nested Table

Pattern n, %	Disease 1 (n=400)	Disease 2 (n=400)
Line of Therapy		
First Line	n=400	n=400
Drug A	120 (30.0)	130 (32.5)
Drug B	100 (25.0)	90 (22.5)
Drug C	180 (45.0)	180 (45.0)
Second Line	n=350	n=385
Drug D	100 (28.6)	160 (41.6)
Drug E	130 (37.1)	105 (27.3)
Drug F	120 (34.3)	120 (31.2)
Treatment		
Initial Treatment	n=400	n=400
Chemotherapy	300 (75.0)	300 (75.0)
Alone	200 (50.0)	210 (52.5)
In Combination	100 (25.0)	90 (22.5)
Radiotherapy	50 (12.5)	60 (15.0)
Chemoradiotherapy	50 (12.5)	40 (10.0)
Palliative Care	n=250	n=300
Surgery	150 (60.0)	100 (33.3)
Chemotherapy	50 (20.0)	25 (8.3)
Taxanes	75 (30.0)	25 (8.3)
Platinum	25 (10.0)	50 (16.7)
Combination	0 (0.0)	50 (16.7)
Radiotherapy	25	50
Chemoradiotherapy	0	50

Complexity

- Increased stratification complexity raises the risk of errors in data extraction. The LLM is more likely to produce incorrect results when more factors are included for extraction. For instance, extracting data based solely on treatment is more likely to yield accurate results than including both treatment and timepoint. Complexity also affects how the data is formatted in the output Excel, with redaction reducing the need for human intervention to replicate the format.

Document Format

- When uploading images of tables, the LLM shows better accuracy in extracting row names from complex and nested tables but often combines information into a single cell (e.g., Male: 2176 [53.2%], Female: 2071 [46.8%]), requiring human input to separate cells for analysis.

Table Orientation

- Table orientation, such as sideways tables, has minimal impact accuracy for simple tables, though the organization may differ from the original format as the information is processed row by row, repeating the table headers for each entry. Minimal human input is needed to correct orientation, resulting in significant time savings.
- Interestingly, when uploading an image of a table with a sideways orientation, the LLM had difficulty accurately extracting the data. This indicates that our tools are superior for extracting information from sideways tables. Rotating the table prior to uploading it to the LLM increased the accuracy, however, errors were still present.

Discussion

This research identified several factors influencing the accuracy of data extraction using LLM-based tools. OCR processing had the most significant impact, as our tools couldn't extract information from unprocessed documents. Redaction improved accuracy when converting plain text, enabling the LLM to easily identify the text to be extracted. For LLMs capable of analyzing images, OCR processing was unnecessary, and redaction had less impact since specific areas could be uploaded as images. However, using images increases the human burden because multiple data points are output into a single cell in Excel, and each section of interest must be uploaded individually.

Although the complexity of stratification had less impact when the LLM was prompted to extract tables in full, difficulties were still encountered due to nesting. These issues were purely related to labeling, and the data remained accurate. Therefore, the complexity of stratification should be a key consideration when extracting information from text. Careful prompt crafting and document pre-processing will be essential to ensure accurate output when the LLM is directed to extract outcomes based on a high level of stratification.

Overall, leveraging LLMs has the potential to significantly enhance the efficiency of data extraction; however, careful consideration of key factors that impact the accuracy of extraction is needed.

Abbreviations AI = Artificial Intelligence; API = Application Programming Interface; LLM = Large Language Model; OCR = Optical Character Recognition

Footnotes [a]GPT versions include 4o-2024-08-06 or 4o-2024-11-020 snapshot.

Presented at
ISPOR 2025
Montreal, QC, Canada
May 13 to 16, 2025

EVERSANA®