

## AI in Research Examples: Automatic data extraction

Using cutting edge language models such as ChatGPT and a technique called OCR (Optical Character Recognition) these researchers were able to automatically extract specific content from hard copies of histopathology reports and be put into a specified digital form.

Link to research paper: <https://pathsocjournals.onlinelibrary.wiley.com/doi/full/10.1002/path.6232>



Figure 1: AI generated image of a large pile of unstructured reports

### How did they achieve this?

The reports given were photos of handwritten or printed documents and therefore the first step was to convert the images into text using the method OCR.

Using a large language model, such as ChatGPT, it is possible to provide the model with an empty template and all extracted text and instruct it to fill in the template.

Putting these techniques together in sequence can automatically create structured digital templates from a series of handwritten documents.

### Why was this not possible previously?

Large language models are fairly new advances in the scientific community and therefore handling unstructured data to fill in a set template was never possible.

Text extraction from images has been around for a while and has continued to improve, however language models now provide us with a way to turn this unstructured text into a structured digital format that we can use.

### How is this applicable to you?

Being able to convert large amounts of handwritten text to a digital format can speed up the process of having to manually copy down information.

Using language models to automatically go through a large number of documents and find all relevant information is very beneficial and can save a researcher a lot of time in collecting data.

### Similar research areas



Digitising historical documents



Extracting useful information from forms, surveys, questionnaires



Extracting diagnoses, names, and treatments from health records to aid in clinical research



Extracting clauses and legal concepts from legal cases



Extracting key concepts to get a quick understanding of a large number of research papers