Generative AI for Research Data Processing: Lessons Learnt From Three Use Cases*

Modhurita Mitra[1], Martine G. de Vos[1], Nicola Cortinovis[1], and Dawa Ometto[1]

Utrecht University, Utrecht, The Netherlands
{m.mitra, m.g.devos, n.cortinovis, d.l.a.ometto1}@uu.nl

Abstract. We explore the feasibility of using generative AI for research data processing by applying this new technique to three different projects from three different scientific domains. We assess if generative AI is an appropriate tool for each data processing task, and come up with strategies to maximise the accuracy and consistency of the results. The projects involve complex data processing tasks: information extraction (of species names from seedlists from botanical gardens), natural language understanding (of Health Technology Assessment documents to extract data points of interest), and text classification (of Kickstarter projects to assign industry codes to them).

Keywords: Generative AI · LLMs · appropriateness · accuracy · consistency · reliability

1 Introduction

There has been enormous interest in generative AI since ChatGPT was launched in 2022. However, there have been concerns about the accuracy and consistency of the results produced by generative AI. In an exploratory study [5], we investigate the feasibility of using generative AI as a tool for research data processing. We share the lessons learnt and insights derived from this process.

The goal of our study is twofold:

1. To determine the conditions under which generative AI is an appropriate tool for a given research task, and
2. To determine strategies to maximise the accuracy and consistency of the results obtained using generative AI.

We focus on two aspects that are both crucial to research data processing methods [4], [3], [6], and about which concerns have been raised with regard to generative AI: *accuracy* and *consistency* of the results obtained using this technique. For generative AI to be a *reliable* data processing tool, the results must be both accurate and consistent.

[*] This is an extended abstract for a paper accepted to the IEEE International Conference on eScience, 16–20 September 2024, Osaka, Japan.

2 Method, use cases, and results

We used the Claude 3 Opus model from Anthropic AI, via its public API, to perform data processing in three use cases. We performed three runs in each case to test for consistency. We present qualitative results, via illustrative examples, for a small, representative dataset for each use case. The use cases are the following:

1. **Seedlists**: Extraction of plant species names from historical seedlists (catalogues of seeds) published by botanical gardens. This is an *information extraction* task.

   For four sample seedlist pages in a variety of different formats, containing a total of 125 plant species names, all the species names were extracted, and were extracted correctly. For documents obtained from OCR (Optical Character Recognition) of scanned seedlists, generative AI was even able to correct some OCR errors and report the correct species names.

2. **Health Technology Assessment (HTA) documents**: Extraction of certain data points (name of drug, name of health indication, relative effectiveness, cost effectiveness, etc.) from documents published by HTA organisations in the EU. This is a *natural language understanding* task.

   For HTA documents about one drug-indication combination, in three different languages (English, Dutch, French), 11 of the 14 desired attributes were correctly extracted from all three documents.

3. **Kickstarter**: Assignment of industry codes to projects on the crowdfunding website Kickstarter. This is a *text classification* task.

   For a sample of 540 representative projects assigned to six human raters in a staggered manner, the highest fraction of industry codes that matched between generative AI and a (single) human rater was 53%, over 145 projects. The highest fraction of codes that matched between two human raters was 60%, over 63 projects. Therefore, in this use case, the performance of generative AI is broadly comparable to that of a human rater.

3 Conclusions

We found that generative AI can be considered as a possible tool for a data processing task if the amount of data to be processed is large, no simple, rule-based method for performing the data processing can be found, and the results are of sufficiently high quality for the research purpose.

The temperature parameter in a generative AI model controls the randomness and variability of the outputs [2], [1]. Setting the temperature parameter to zero maximises the accuracy and consistency of the outputs. A clear, well-defined, unambiguous prompt helps in precise and accurate extraction of the desired attributes from the input data.

References

1. Anthropic AI: API reference (May 2024), `https://docs.anthropic.com/en/api/complete`, accessed: 2024-05-21
2. Anthropic AI: Glossary: Temperature (May 2024), `https://docs.anthropic.com/en/docs/glossary\#temperature`, accessed: 2024-05-21
3. Boisvert, R.F., Cools, R., Einarsson, B.: 2. Assessment of Accuracy and Reliability, pp. 13–32. SIAM (2005). https://doi.org/10.1137/1.9780898718157.ch2, `https://epubs.siam.org/doi/abs/10.1137/1.9780898718157.ch2`
4. Cong, G., Fan, W., Geerts, F., Jia, X., Ma, S.: Improving data quality: Consistency and accuracy. In: Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007. pp. 315–326 (2007)
5. Mitra, M., de Vos, M.G., Cortinovis, N., Ometto, D.: Generative AI for Research Data Processing: Lessons Learnt From Three Use Cases. In: IEEE International Conference on eScience. Osaka, Japan (September 2024)
6. National Academies of Sciences, Engineering, and Medicine: Reproducibility and Replicability in Science. National Academies Press, Washington, DC (May 2019), `https://www.ncbi.nlm.nih.gov/books/NBK547546/`, accessed: 2024-04-17