2023 IEEE 17th International Conference on Application of Information and Communication Technologies (AICT) | 979-8-3503-0356-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/AICT59525.2023.10313167

Effectiveness of Generative Artificial Intelligence for Scientific Content Analysis

Moritz Platt, Daniel Platt
King's College London, UK
moritz.platt@kcl.ac.uk, daniel.platt.berlin@gmail.com

***Abstract*—Generative artificial intelligence (GenAI) in general, and large language models (LLMs) in particular, are highly fashionable. As they have the ability to generate coherent output based on prompts in natural language, they are promoted as tools to free knowledge workers from tedious tasks such as content writing, customer support and routine computer code generation. Unsurprisingly, their application is also attractive to professionals in the research domain, where mundane and laborious tasks, such as literature screening, are commonplace. We evaluate Vertex AI 'text-bison', a foundational LLM model, in a real-world academic scenario by replicating parts of a popular systematic review in the information management domain. By comparing the results of a zero-shot LLM-based approach with those of the original study, we gather evidence on the suitability of state-of-the-art general-purpose LLMs for the analysis of scientific content. We show that the LLM-based approach delivers good scientific content analysis performance for a general classification problem ($ACC = 0.9$), acceptable performance for a domain-specific classification problem ($ACC = 0.8$) and borderline performance for a text comprehension problem ($ACC \approx 0.69$). We conclude that some content analysis tasks with moderate accuracy requirements may be supported by current LLMs. As the technology will evolve rapidly in the foreseeable future, studies on large corpora, where some inaccuracies are tolerable, or workflows that prepare large data sets for human processing, may increasingly benefit from the capabilities of GenAI.**

***Index Terms*—AI-Assisted Research, Literature Screening, Content Analysis, Prompt Engineering, Classification Performance**

I. Introduction

There is significant uncertainty surrounding the effectiveness of generative artificial intelligence (GenAI) and, specifically, large language models (LLMs) for academic tasks. This uncertainty is likely to become a much more salient issue as increasing numbers of academics are tempted to apply GenAI to improve productivity [1]. Recent literature has shown promising results for the effectiveness of LLMs for named entity extraction, classification and information extraction. In the present study, we analyse the effectiveness of LLMs in a scientific context by conducting a series of simple experiments using GenAI to replicate a study from the information and communication technology (ICT) literature that was conducted using conventional scientific content analysis methods. By comparing the results of the GenAI approach with the 'ground truth' of the manual approach, we gain valuable insight into the effectiveness of GenAI in content analysis. Our experiments indicate that general-purpose GenAI models achieve good performance in well-defined tasks such as recognition of named entities, but poorer performance in more complex tasks, such as those requiring a deeper understanding of the subject matter.

We propose the following hypothesis: *general-purpose GenAI provides sufficiently accurate results in common scientific content analysis tasks to be applied as a research tool.*

II. Background

LLMs are particularly large artificial neural networks (ANNs). The latter have historical precedence, with origins tracing back to studies conducted in the 1940s [2]. ANNs are inspired by the human brain and consist of many processing nodes that link inputs (e.g., data to be classified) and outputs (e.g., a classification label) via weighted connections. During training (e.g., providing the model with data and classification labels), the weights of an ANN are adjusted so that future inputs can be automatically classified based on previously processed examples. Early work was largely theoretical, but subsequent increases in computing performance made real-world applications of ANN feasible [3]. This allowed for applications on large bodies of text. Nonetheless, early ANNs did not achieve sufficient performance without extensive fine-tuning. The later LLMs, however, were able to achieve notable performance out-of-the-box, because of their ability to process sentences nonsequentially and to make use of the concept of 'attention', a mechanism which allows the model to selectively focus on specific parts of the input sequence, capturing long-term dependencies [4]. This contribution, combined with training on extensive corpora, led to performance improvements for many natural language processing tasks. An additional benefit of later commercialization of LLMs was the intuitive chat interface through which the models could be prompted: an idea that goes back to experimental software from the 1960s that allowed users to engage in dialogue with a computer [5] and makes interaction with a model feasible for laypeople.

III. Literature Review

Despite the relatively low maturity of GenAI, there is some relevant work regarding its use in content analysis. Yang, Li, Zhang *et al.* [6] show that general-purpose LLMs perform reasonably well when summarizing forum posts, news articles, meeting minutes and fiction. More generally, common commercial general-purpose LLMs perform well in classification tasks of unstructured data [7], [8] and structured data [9]. In addition to these promising results achieved with pre-trained models, it was found that fine-tuning can significantly improve

classification performance [10]. Further customizations, such as domain-specific LLMs, trained on domain data and configured for domain-specific tasks, outperform other artificial intelligence (AI) techniques even more clearly [11], [12]. The results of LLMs can be further improved by 'prompt engineering', a technique to improve input instructions that were found to have a strong influence on the output [13], [14]. With regards to the reliability of LLM in the context of critical tasks, some authors advise 'extreme caution' considering the limitations of the technology [15], specifically around the risk of generating inconsistent and hallucinated information, concluding that a 'human-in-the-loop' process is necessary to ensure quality [16].

IV. Methods

To investigate whether GenAI can be applied in a realistic research setting, we created a set of three experiments, representative of common tasks of analysis of scientific content.

A. Materials

We selected a paper from the ICT literature, specifically from the *International Journal of Information Management*, which according to the Scimago Journal and Country Rank[1] is the leading journal in the field of information systems and management. From the corpus of articles published in this journal, we chose an article that allowed for replication using AI technology, entitled 'The impact of knowledge management processes on information systems: A systematic review' [17]. This article uses typical systematic review methods and documents the results in such a way that they can be easily compared with the results produced by an AI workflow. The article constitutes a systematic review of 41 knowledge management studies related to information systems [17], of which 36 could be obtained[2] and were further processed as described in subsubsection IV-C1. These articles, combined with the results of the results of Al-Emran, Mezhuyev, Kamaludin *et al.* [17], constituted the materials for the study.

B. Study Design

We selected three subtasks from the manuscript [17] to form the three experiments of this study.

1) Experiment I: Detection of Study Country: In this experiment, we replicated work in the context of research question 4 of the study by Al-Emran, Mezhuyev, Kamaludin *et al.* [17, Sec. 4.4.1]. In the original study, the authors manually categorized research studies by their country of implementation. We performed this named entity recognition task using an LLM and compared the country names recognized by the LLM with those of the original article. We considered the classification to be successful when the LLM categorization matched that of the original article.

[1]Online at https://www.scimagojr.com/.

[2]The remaining five studies were not available digitally via King's College London Libraries.

2) Experiment II: Detection of the Research Method: In this experiment, we replicated work in the context of research question 2 of the original study [17, Sec. 4.2]. Here, Al-Emran, Mezhuyev, Kamaludin *et al.* [17] categorized studies by their research method; either questionnaire surveys or interviews combined with questionnaires. We instructed the LLM to decide for all presented papers whether they constituted a questionnaire survey or a combination of questionnaire survey and interview. We considered the categorization successful if the LLM output corresponded to the category defined original article [17].

3) Experiment III: Detection of Participant Type: This experiment is aligned with research question 1 of the study by Al-Emran, Mezhuyev, Kamaludin *et al.* [17, Sec. 4.1], in which the authors determined the types of participants within the studies. Here, the original authors defined the participants in free form, without adhering to predefined categories [17, Tab. 5]. Equally, we instructed the LLM to determine the type of participant without providing categories or examples (see subsubsection IV-B4). We considered a result to be successful if the result generated by the LLM was sufficiently similar to that of the original article [17], which introduces the need for interpretation (see subsubsection IV-C4).

4) Prompt Design: LLMs are commonly interfaced through natural language instructions, or 'prompts'. This is also true for the 'text-bison' model we used (see subsubsection IV-C1). As shown previously (see section III), the design of the instructions for any LLM has a significant influence on how well it performs. However, the prompts used in this study were deliberately simple to illustrate the intuitiveness of obtaining results from LLM. Therefore, they should be considered a starting point for optimizations and it can be expected that further improvements of the prompts might have a positive effect on accuracy [18]. The academic manuscript to be analysed was provided to the LLM prompt interface first, followed by a clear task. All tasks were written in basic American English and made direct reference to the text to be analysed. All prompts explicitly specified the desired output format.

For different tasks, different prompts were designed. For experiment I, in which the LLM was used to determine the country in which the study in question was conducted, the following prompt was used.

> Given this text on a research study related to knowledge management, identify and extract the country of implementation where the study was conducted. Prioritize accuracy and avoid making assumptions not present in the text. Answer with the country name only.

Experiment II used a more prescriptive prompt that explicitly outlines the expected classes, effectively turning this task into a classification problem with three classes:

> Carefully analyze the scientific text provided. Determine and classify if the research study described in the text uses: only a survey method, both a survey and interview methods, or neither of the methods.

Table I
THE MODEL SETTINGS OF THE 'TEXT-BISON' MODEL USED IN THE EXPERIMENTS.

Parameter	Value
Temperature	0
Max. output tokens	128
Top-k	1
Top-p	0

In experiment III, a very concise and unspecific prompt was used to test the performance of the LLM under uncertainty. Here, we simply specified that the model should return the 'role and corporate level of the participants', a very broad task:

> Analyze the text provided about a knowledge management research study. Identify the role and corporate level of the participants. Respond in one word.

C. Research Protocol

The set of experiments was controlled from a simple laptop computer and executed on Google Cloud Platform (GCP) Vertex AI. Cloud computing in general, and the Vertex AI platform in particular, is a suitable choice for AI research as it offers tools that facilitate the rapid and efficient building, deployment and scaling of machine learning models and cost-effective infrastructure. As the goal of the experiments was to generate results that reflected the level of sophistication of laypeople in AI technology, advanced techniques, such as fine-tuning or improving data quality through optimization of the input text, were omitted.

1) Model Configuration: To execute the experiments, we chose GCP Vertex AI[3], a Software-as-a-Service platform that makes available a range of foundational models, including 'text-bison'. The 'text-bison' model that we applied is positioned as a foundational tool, appropriate for a range of language tasks, from classification to concept ideation. As part of the PaLM 2 language model family, 'text-bison' represents one of the larger and more powerful configurations in the series, optimized for complex computational tasks and deep contextual understanding. This model has shown good performance characteristics for classifying colloquial English language [8, p. 12–13].

We configured the model as shown in Table I. The settings were chosen to make the results reproducible: the temperature, top-k and top-p parameters were set so that the least random responses were generated.

2) Preparation of Data: The complete manuscripts in Portable Document Format (PDF) were obtained from the respective publisher websites. Since common LLMs require unformatted text input, the PDF files were converted to plain text using 'IronPDF'[4], a commercial PDF manipulation library. No further document sanitization was performed. The resulting plain text files were stored locally for later processing.

[3]Online at https://cloud.google.com/vertex-ai.

[4]Online at https://ironpdf.com/.

3) Execution of Experiment: We executed the experiment through a Python script, building on the `aiplatform` package. The experiment consisted of a simple call to the `model.predict` endpoint with the input text and prompt (see subsubsection IV-B4) combined into a single message. The resulting response was stored locally for analysis.

4) Measuring Accuracy: Accuracy was measured by comparing the class labels generated by the LLM with the class labels from the original study [17]. Where the original study, for whatever reason, did not provide class labels, the corresponding manuscript was removed from the experiment, resulting in different sample sizes for the various experiments (see Table II). We did not expect the output of the LLM to correspond character-by-character to the results of the original study. Therefore, in experiments 1 and 2, we manually compared the generated labels with the expected labels. Minor semantic differences (e.g., 'United Arab of Emirates' [sic] vs. 'United Arab Emirates', or 'U.S.' vs. 'USA') were accepted. In experiment 3, the original study did not prescribe labels, but used free-form descriptions. Therefore, we manually compared the labels generated by the LLM with the descriptions in the original study, allowing small scope for interpretation (e.g., 'Firms' top management' and 'top management' were treated as a match).

V. RESULTS

As shown in Table II, our experiments delivered classification accuracy of $68.57 - 90\%$. The most accurate experiment ($ACC = 90\%$) was a simple named entity recognition task, for which no specific knowledge was required. A classification experiment with three target classes had an accuracy of 80%. The least accurate experiment ($ACC = 68.75\%$) constituted an information extraction problem that required a sufficiently deep understanding of the subject matter.

Table II
THE CLASSIFICATION PERFORMANCE OF THE LLM (ACC), COMPARED WITH THE PERFORMANCE OF A MAJORITY CLASS CLASSIFIER (ACC_{Maj}).

Experiment	n	ACC	$\mathbf{ACC_{Maj}}$
I Named Entity Extraction (Countries)	30	90 %	20 %
II Classification (Method)	35	80 %	70 %
III Information Extraction (Participants)	32	68.75 %	37.5 %

We therefore infer that our hypothesis is supported: GenAI may provide sufficient performance for well-defined and well-structured tasks that can be solved with general knowledge, such as named entity recognition or classification with a small number of target classes. GenAI may also, albeit less clearly, be appropriately performant for more complex information extraction problems.

VI. DISCUSSION

Our experiments show that GenAI performs best in well-defined tasks like named entity recognition, with varying accuracy in more complex tasks. This shows that GenAI has

strong potential for specific academic tasks. These findings underscore GenAI's potential in academic tasks, emphasizing the need for careful validation and responsible application in the evolving knowledge economy. Future studies should explore the boundaries of GenAI accuracy across different academic tasks and develop robust validation protocols to ensure ethical implementation.

A. Limitations

It is conceivable that the LLM model used could have been trained on the paper we are analysing. However, we note that LLMs are trained on vast amounts of data, which minimizes the influence of any single document. Additionally, our results are based on comparisons to a previous study, but we must acknowledge the potential ambiguity in their ground truth, as no study can guarantee absolute correctness and clarity.

VII. Conclusions

With the expected penetration of GenAI of many areas of the knowledge economy, academia will not be able to resist its great potential. Therefore, research must find an appropriate way of engaging with GenAI. Instead of demonizing this technology, academia must find clear rules for its use. The present paper has shown that GenAI can solve some tasks of scientific text analysis with high accuracy without requiring complex training or configuration. Therefore, it can be considered to have great potential for reducing workloads in the context of literature review and screening, and beyond.

However, this potential must be realized responsibly in order to use it ethically. In particular, the outputs of current versions of GenAI models should not be used without validation when high accuracy is required. Our results call for workflows that mitigate the anticipated inaccuracies of LLMs.

Declaration of Interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: M.P. reports a relationship with Google Germany GmbH that includes employment and equity or stocks.

The research activities culminating in this manuscript were undertaken by M.P. as part of his studies at King's College London. The opinions expressed in this article are the author's own and do not reflect the views of his employer.

References

[1] A. Birhane, A. Kasirzadeh, D. Leslie and S. Wachter, 'Science in the age of large language models', *Nature Reviews Physics*, vol. 5, no. 5, pp. 277–280, Apr. 2023. DOI: 10.1038/s42254-023-00581-4.

[2] W. S. McCulloch and W. Pitts, 'A logical calculus of the ideas immanent in nervous activity', *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943. DOI: 10.1007/bf02478259.

[3] H. Jaakkola, J. Henno, J. Makela and B. Thalheim, 'Artificial intelligence yesterday, today and tomorrow', in *Proceedings of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics*, Opatija, Croatia: IEEE, May 2019, pp. 860–867. DOI: 10.23919/mipro.2019.8756913.

[4] A. Vaswani, N. Shazeer, N. Parmar *et al.*, 'Attention is all you need', in *Proceedings of the 31st Conference on Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio *et al.*, Eds., Long Beach, CA, USA, 2017, pp. 5998–6008.

[5] J. Weizenbaum, 'ELIZA—a computer program for the study of natural language communication between man and machine', *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966. DOI: 10.1145/365153.365168.

[6] X. Yang, Y. Li, X. Zhang, H. Chen and W. Cheng. 'Exploring the limits of ChatGPT for query or aspect-based text summarization'. arXiv: 2302.08081 [cs.CL]. (Feb. 2023).

[7] G. Zhang, J. Wu, M. Tan, Z. Yang, Q. Cheng and H. Han, 'Learning to predict U.S. policy change using New York Times corpus with pre-trained language model', *Multimedia Tools and Applications*, vol. 79, no. 45-46, pp. 34 227–34 240, May 2020. DOI: 10.1007/s11042-020-08946-y.

[8] R. Anil, A. M. Dai, O. Firat *et al.* 'PaLM 2 technical report'. arXiv: 2305.10403 [cs.CL]. (May 2023).

[9] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang and D. Sontag, 'TabLLM: Few-shot classification of tabular data with large language models', in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, F. Ruiz, J. Dy and J.-W. van de Meent, Eds., PMLR, 2023, pp. 5549–5581.

[10] M. Khadhraoui, H. Bellaaj, M. B. Ammar, H. Hamam and M. Jmaiel, 'Survey of BERT-base models for scientific text classification: COVID-19 case study', *Applied Sciences*, vol. 12, no. 6, 2891, Mar. 2022. DOI: 10.3390/app12062891.

[11] X. Yang, A. Chen, N. PourNejatian *et al.*, 'A large language model for electronic health records', *npj Digital Medicine*, vol. 5, no. 1, 194, Dec. 2022. DOI: 10.1038/s41746-022-00742-2.

[12] A. H. Huang, H. Wang and Y. Yang, 'FinBERT: A large language model for extracting information from financial text', *Contemporary Accounting Research*, vol. 40, no. 2, pp. 806–841, Jan. 2023. DOI: 10.1111/1911-3846.12832.

[13] L. Giray, 'Prompt engineering with ChatGPT: A guide for academic writers', *Annals of Biomedical Engineering*, Jun. 2023. DOI: 10.1007/s10439-023-03272-4.

[14] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann and Q. Yang, 'Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts', in *Proceedings of the 2023 Conference on Human Factors in Computing Systems*, Hamburg, Germany: ACM, Apr. 2023, pp. 1–21. DOI: 10.1145/3544548.3581388.

[15] M. Sallam, 'ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns', *Healthcare*, vol. 11, no. 6, 887, Mar. 2023. DOI: 10.3390/healthcare11060887.

[16] A. Deroy, K. Ghosh and S. Ghosh, 'How ready are pre-trained abstractive models and LLMs for legal case judgement summarization?', in *Proceedings of the 3rd International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace*, J. G. Conrad, D. W. Linna, J. R. Baron *et al.*, Eds., Braga, Portugal, 2023, pp. 8–19.

[17] M. Al-Emran, V. Mezhuyev, A. Kamaludin and K. Shaalan, 'The impact of knowledge management processes on information systems: A systematic review', *International Journal of Information Management*, vol. 43, pp. 173–187, Dec. 2018. DOI: 10.1016/j.ijinfomgt.2018.08.001.

[18] C. E. Short and J. C. Short, 'The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation', *Journal of Business Venturing Insights*, vol. 19, e00388, Jun. 2023. DOI: 10.1016/j.jbvi.2023.e00388.