RESEARCH ON THE USE OF ARTIFICIAL INTELLIGENCE TO IMPROVE BIBLIOMETRIC ANALYSIS OF SCIENTIFIC ARTICLES

Marius SAVA[1], Gheorghe MILITARU [2]

Bibliometric analysis is a resource-intensive research activity, particularly when performing text analysis. This paper evaluates the reliability of using large language models, specifically GPT-4o, to improve the analysis of scientific papers. A total of 813 paper abstracts on improving organizational performance in the service industry using AI-based solution were analysed, comparing automated analysis with manual evaluation, and assessing the precision and level of agreement. The results show that GPT-4o is a powerful enhancer in the scientific research toolbox, significantly improving time efficiency and scalability of the process, but without providing enough context, the automation benefits are severely diminished.

Keywords: Bibliometric Analysis, Artificial Intelligence, GPT-4o, LLM

1. Introduction

Bibliometric analysis is defined as a technique to deal with large volumes of scientific data [1]. While the bibliometric analysis toolbox usually comprises citation count or co-authorship relationships, content analysis has become a useful part of it in recent years. For example, Bibliometrix [2], a popular choice among researchers, incorporates features for text mining of abstracts.

While various software was created to automate bibliometric analysis techniques, like the VosViewer [3], Bibliometrix [2] or CityExplorer [4], this activity of mapping the landscape of a domain scope remains a very time-consuming one [5].

Large language models (LLM) technology, a subset of machine learning field, exploded in recent years and is expected to enhance any type of digital activity [6]. pyBibX [7] presents an AI solution that offers features like “Abstractive Text Summarization” and uses GPT-4o for automated exploratory data analysis reports. The study, however, does not measure the accuracy of the reports.

Specialized AI research assistants, like Elicit [8] allow us to automate the usual research workflow, enabling searching, filtering and extracting key features

[1] PhD student, Faculty of Entrepreneurship, Business Engineering and Management, National University of Science and Technology POLITEHNICA, Bucharest, Romania, e-mail: savamarius14@protonmail.com

[2] Professor, Faculty of Entrepreneurship, Business Engineering and Management, National University of Science and Technology POLITEHNICA, Bucharest, Romania, e-mail: gheorghe.militaru@upb.ro

from paper abstracts. The assistant allows the creation of custom categories against which the abstracts can be categorized. The tool's main advantage is providing an out-of-the-box solution, but there is a low level of control over the model's instructions or output.

This article's intention is to verify in a case study whether LLM technology can enhance bibliometric analysis by using GPT-4o benchmarked against manual methods. The selected topic for the case study is the current state of research on improving organizational performance in the service industry using artificial intelligence-based solutions. LLM is an emerging technology that will accelerate how big data is understood in the scientific research field, showing high levels of precision closer to human performance. In the next sections, using the selected solution for investigation, GPT-4o, a set of 813 scientific article abstracts is analysed. The same analysis is performed manually by the authors. The results are then discussed with a focus on benefits and limitations.

2. Large language models

Large language models are models which are trained on vast amounts of data, capable of providing language generation and are mainly based on transformers architecture. The transformer architecture was firstly introduced with the article "*Attention is all you need*" (Vaswani et al. 2023) [9], in 2017, by eight scientists working at Google. As the time of writing this paper, the article has been cited by more than 135,000 times, making it one of the most cited articles in the field of computer science.

There are four main generations of LLMs as described in Fig. 1 [10]. There is an evolution throughout which the main goal is a model which can solve general problems, that is, tasks which are not specialized and most important, tasks on which the model had not been trained on before.

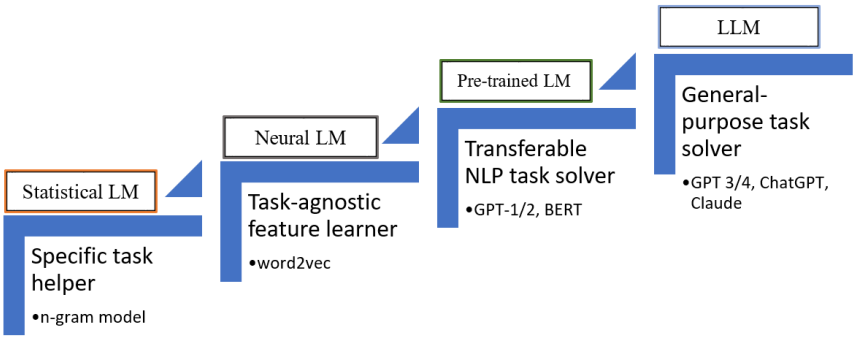

Fig. 1. An evolution process of the four generations of language models (LM) from the perspective of task solving capacity. Adapted from source: [10].

Key characteristics of Large Language Models

During this section, it is explained what characteristics define LLMs in order to understand their functionality and application.

Scalability refers to the ability of a system to deal with a growing amount of work [11]. The parameters which are involved into the scalability are the following [12]: model size, determined by the number of parameters, training dataset size, and the amount of computation used during training. One of the key results highlighted by the law is that there is a diminished return value in increasing each of these three factors at some point and balancing them is the optimal solution.

Contextual understanding: Previous models of machine learning had one meaning for each word, and problems in understanding jokes, for example, were unsolvable [13]. Context-aware representations [14] allow multiple meanings for each word, with long-range dependencies [15] to model rich semantics.

Generative capabilities: Large language models are able to generate any type of data, including music [16], or images like x-rays [17]. Basically, any digitized content could theoretically, be generated with the appropriate model.

Learning capabilities: Large language models are trained on vast amounts of data, generally in an unsupervised manner [14]. However, there is also a necessary and useful business case, to include some updated data for the model to remain valid. Few-shot learning [18] is a solution approach that enables updates in the model using few examples to solve tasks different from those in the initial dataset on which the model had been trained on. Zero-shot learning [19] aims to solve general problems based on the vast amount of data on which the model was trained. Since GPT-2, two post-training stages were critical for the model's performance: instruction fine-tuning, that is, providing to the model a set of pairs <instruction, answer>, and preference alignment using techniques like Reinforcement Learning from Human Feedback (RLHF) [20] and Direct Preference Optimization (DPO) [21], in which human judgement is used to reshape the output by selecting the preferred version from pairs of model answers.

Models fine-tuning: Large language models can be trained on general tasks like the open-source available LLaMA model [22], and depending on the specialized tasks that need to be fulfilled, fine-tuned to enhance their performance. Fine-tuning is not a recent technique, being used previously in fields like computer vision and transfer learning [23]. The most important benefit of it is that open-source models could be trained using a common pool of resources, enabling various consumers to fine-tune them at a lower cost for their applications.

Security and ethical considerations**:** From the beginning of the first wave of LLMs deployment, security and privacy considerations were raised [24]. Bias in data [25], privacy of data users [26], discrimination continuity by artificial

intelligence solutions [27], exploitation of data from LLMs [28] are main problems that need to find better solutions.

Hallucinations: The biggest performance hindrance for LLMs is the problem of inventing content when they do not have an answer. The main cause of hallucinations is source-reference divergence [29]. This can occur, for example, during the supervised learning step, when the model internalizes the pattern that adding plausible facts is acceptable as given in the training pairs. In this way, the model can generate outputs that are far from the training data.

For this paper, the most advanced general-purpose language model as of September 2024 is used, GPT-4o [30], according to LLM Leaderboards [31], which take into consideration standard evaluation datasets such as MMLU[3], GPQA[4] and MATH[5]. It has a context window of 128k tokens, and its main advantages are its versatility and state-of-the-art performance.

3. Application of ChatGPT in bibliometric analysis

The methodology used is a case study validated through statistical techniques during a comparative analysis. The selected case study is a common research activity: conducting a bibliometric analysis which defines the state of a research field for a particular study object – in this case, the current body of research performed on improving organizational performance in the service industry using artificial intelligence-based solutions.

A bibliometric analysis is performed on a dataset of scientific articles. The selection of this dataset depends on the topic, size of the research conducted, the scientific database queried. An initial dataset of research is selected and the main characteristics of it are presented. The LLM technology used to perform the automated analysis is also selected. Both automated and manual evaluations are performed on the dataset, using a predefined set of criteria. Using a statistical analysis the level of concordance between the two evaluations is presented. Finally, the interpretation of the results is discussed.

3.1. Dataset selection

A bibliometric analysis was conducted on the current state and emerging trends in the topic of improving organizational performance in the service industry using artificial intelligence. For "*performance*" query a number of 8,003,351 documents were found starting with 1831 to 2024, in the Scopus database.

The search was then restricted to "artificial intelligence AND performance" with 115,904 works. The volume of work was still large, so a more specific query

[3] Massive Multitask Language Understanding
[4] Graduate-Level Google-Proof Q&A Benchmark
[5] Measuring mathematical problem solving

by joining names of several industries within artificial intelligence's major building blocks produced a precisely 813 documents. The executed query was:

(("gpt" OR "artificial intelligence*" OR "machine learning" OR "deep learning") AND ("performance") AND ("customer service" OR "hospitality" OR "retail" OR "financial service*" OR " public service*" OR "IT service*" OR "business service" OR "consulting service*" OR "utility service*" OR "healthcare" OR "education") AND ("industry"))*

The main characteristics of the dataset are presented in Table 1. The most important data for our purpose is the abstract, which will be analysed according to the criteria.

Table 1

Main characteristics for selected query

Timespan	Sources	Documents	Annual growth rate	Authors	Authors of single-authored docs	International co-authorship	Co-authors per doc	References	Document average age	Average citations per article
1989:2024	478	813	13.02%	3207	58	30.63%	4.06	43683	2.32	21.11

3.2. Defining criteria for evaluation

The following dimensions, presented in detail, were selected as criteria for evaluation: AI categories, measurable improvements, scope of research, type of validation, challenges. These dimensions are based on principles of interpretation, requiring an understanding of *where* the article fits in the specific domain category, *why* the research was conducted, *what* the scope of research is, *how* the research is validated, and *what* challenges are encountered in the performed research.

Table 2

Dimensions analysed for the research dataset

Main category	Subcategory	
AI category	General AI	Edge AI
	AI applications into blockchain	Explainable AI
	Data mining	Federated learning
	Deep learning	Generative AI
	Digital twin	Machine learning
Measurable improvement	Accuracy and performance metrics	Healthcare and diagnosis
	Speed, efficiency, and cost	Productivity and operations
	Energy efficiency	Classification performance
Measurable improvement (cont.)	Sales and growth	Network and data
	Error rate reduction	Model performance
	User satisfaction	Security and privacy

Main category	Subcategory		
	Prediction and forecasting		
Scope	Simulation	Comparative study	Literature review
	Hands-on tutorial	Real-life application	Case study
	Experimental study	Survey	Methodology proposal
	System development	Proof of concept	Theoretical framework
	Empirical study	Comprehensive insight	
Validation type	Experimental results	Survey/questionnaire	
	Performance evaluation	Statistical analysis	
	Comparison study	Real-world application	
	Simulation	Theoretical analysis	
Challenges	Biosecurity and health	Data quality and management	
	Environmental and external influences	Ethical and social impact	
	Real-time and computational constraints	Implementation and integration	
	Prediction accuracy	Scalability and performance	
	Security and privacy		

3.3. Performing the evaluation

An in-depth analysis was conducted using GPT-4o by analysing the abstract of articles on the following dimensions AI techniques, measurable improvements, scope of research, type of validation and challenges. The query batches were exemplified in the Appendix of this paper. The abstracts, the predefined categories and the preferred output were integrated into the query as building blocks.

The overall accuracy for the articles in Table 3 is relatively good at 88.60%, being equally distributed across all categories, except for the '*General AI*' category. Fig. 2 exemplifies the main cause for this, which is that was that the term '*General AI*' is interpreted sometimes as AI in the most general sense, that is, a set of activities that are related to AI without going into specifics and sometimes refers to '*Artificial General Intelligence*' - AGI, which is a type of AI that is able to solve any type of problem. According to the available data, the number of articles that fail to indicate the type of AI or AI application is quite limited. Another misclassification is that of false positives, for example, in the '*machine learning*' or '*deep learning*' subcategories. There are abstracts that just mention '*machine learning*' without using it, but they are identified as if they are using it.

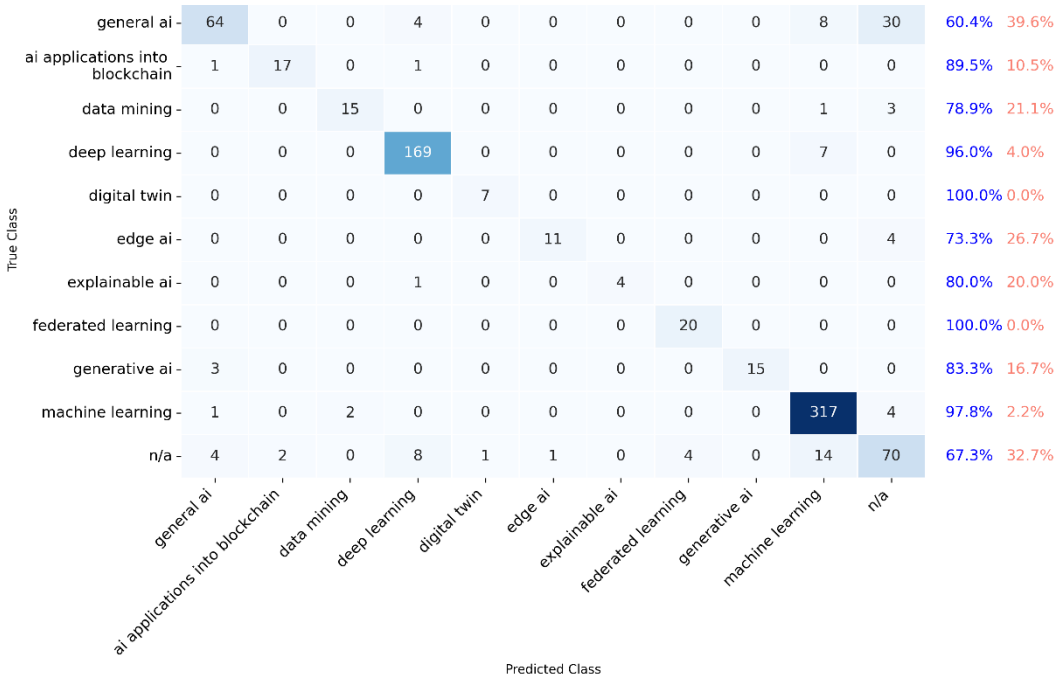

Fig. 2. Confusion matrix of the gpt-4o model for AI main techniques

Table 3

Evaluation metrics of the gpt-4o model for AI main technique

AI techniques subcategories	Overall accuracy	Precision	Recall	F1-Score
General AI	88.60%	87.67%	60.38%	71.51%
AI applications into blockchain		89.47%	89.47%	89.47%
Data mining		88.24%	78.95%	83.33%
Deep learning		92.35%	96.02%	94.15%
Digital twin		87.50%	100.00%	93.33%
Edge AI		91.67%	73.33%	81.48%
Explainable AI		100.00%	80.00%	88.89%
Federated learning		83.33%	100.00%	90.91%
Generative AI		100.00%	83.33%	90.91%
Machine learning		91.35%	97.84%	94.49%
N/A		63.06%	67.31%	65.12%

The overall accuracy for the '*Measurable improvement*' category was 63,19% as described in the Table 4. The confusion matrix, represented in the Fig. 3, shows that the distinction between '*accuracy & performance metrics*', '*model performance*' and '*classification performance*' was very blurred. Sometimes the answer could be a multi-class answer, but the test only assumes single-class responses. A notable number of unspecified entries, above one half, indicates the need for more detailed reporting on measurable improvements.

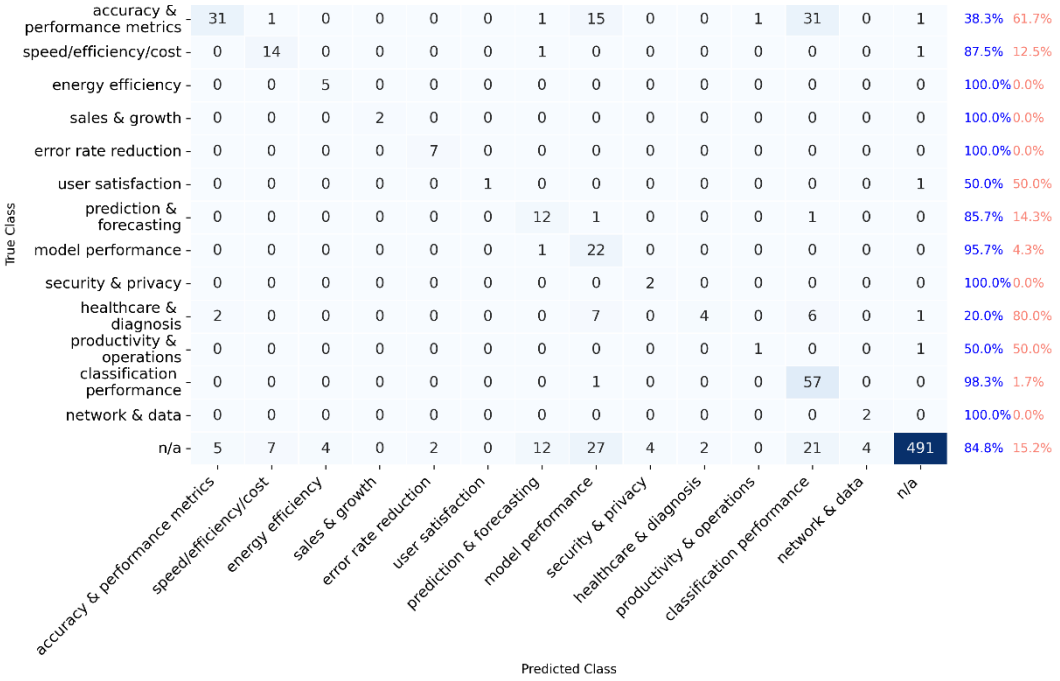

Fig. 3. Confusion matrix of the gpt-4o model for measurable improvement

Table 4

Evaluation metrics of the gpt-4o model for measurable improvement category

Measurable improvements	Overall accuracy	Precision	Recall	F1-Score
Accuracy and performance metrics	63.19%	81.58%	38.27%	52.10%
Classification performance		49.14%	98.28%	65.52%
Energy efficiency		55.56%	100.00%	71.43%
Error rate reduction		77.78%	100.00%	87.50%
Healthcare and diagnosis		66.67%	20.00%	30.77%
Model performance		30.14%	95.65%	45.83%
Network and data		33.33%	100.00%	50.00%
Prediction and forecasting		44.44%	85.71%	58.54%
Productivity and operations		50.00%	50.00%	50.00%
Sales & growth		100.00%	100.00%	100.00%
Security and privacy		33.33%	100.00%	50.00%
Speed, efficiency, and cost		63.64%	87.50%	73.68%
User satisfaction		100.00%	50.00%	66.67%
N/A		81.58%	38.27%	52.10%

The overall accuracy for the '*Scope of research*' category was 83,92% as described in the Table 5. The confusion matrix, represented in the Fig. 4, shows that the model did not perform well in identify papers whose only objective was on the theoretical level, and misidentified them as '*methodology proposal*' or '*system development*'. Many of the wrong labels for '*system development*' were in fact

'*experimental study*', the type of research that does not necessarily have a product as an output as would be expected from system development process. A third category of wrong categorization is the large number of items marked as 'N/A', the model not being able to identify the correct label from the abstract.

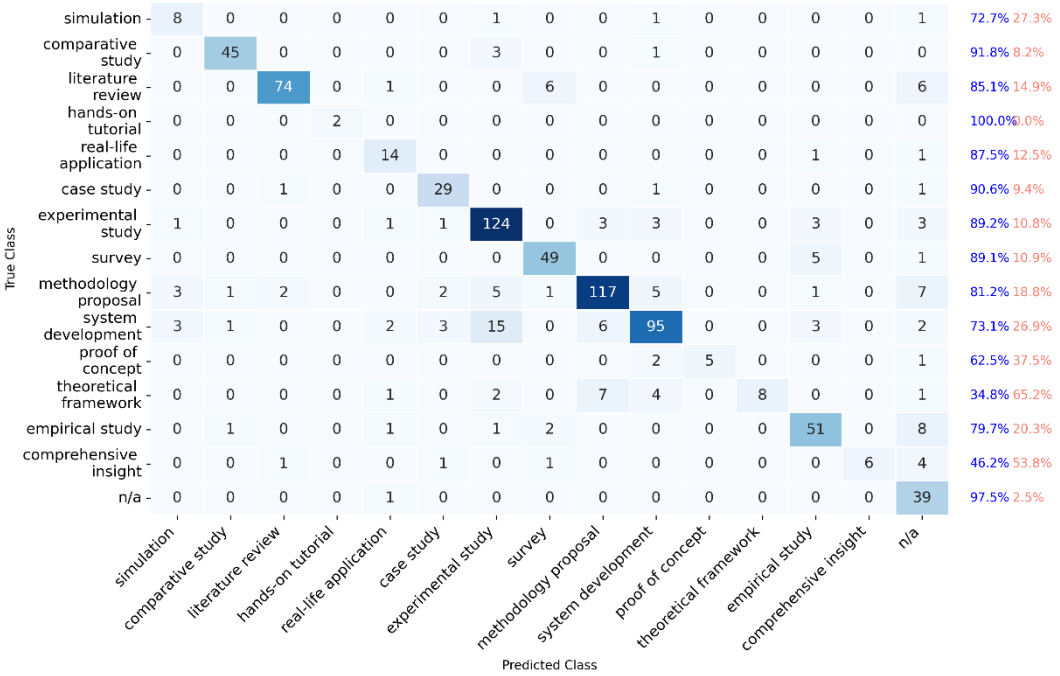

Fig. 4. Confusion matrix of the gpt-4o model for scope of research

Table 5

Evaluation metrics of the gpt-4o model for scope of research

Scope	Overall accuracy	Precision	Recall	F1-Score
Case study	83.92%	80.56%	90.63%	85.29%
Comparative study		93.75%	91.84%	92.78%
Comprehensive insight		100.00%	46.15%	63.16%
Empirical study		79.69%	79.69%	79.69%
Experimental study		82.12%	89.21%	85.52%
Hands-on tutorial		100.00%	100.00%	100.00%
Literature review		94.87%	85.06%	89.70%
Methodology proposal		87.97%	81.25%	84.48%
Proof of concept		100.00%	62.50%	76.92%
Real-life application		66.67%	87.50%	75.68%
Simulation		53.33%	72.73%	61.54%
Survey		83.05%	89.09%	85.96%
System development		84.82%	73.08%	78.51%
Theoretical framework		100.00%	34.78%	51.61%
N/A		52.00%	97.50%	67.83%

The overall accuracy for the '*Validation type of research*' category was 79,57% as described in the Table 6. The confusion matrix, represented in the Fig. 5, shows that the model had high rates errors when it needed to distinguish '*performance evaluation*' from '*comparison study*'. In this case, the confusion arises from using a comparison between multiple methods to test a model, instead of multiple models validated with a single method.

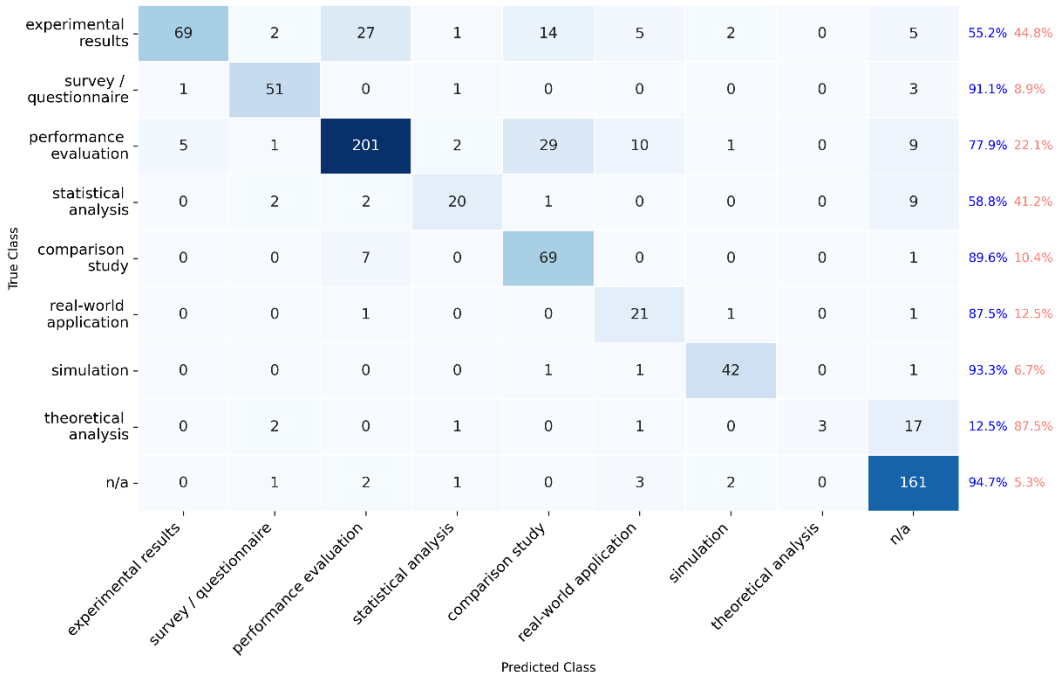

Fig. 5. Confusion matrix of the gpt-4o model for validation type of research

Table 6

Evaluation metrics of the gpt-4o model for validation type of research

Validation type	Overall accuracy	Precision	Recall	F1-Score
Comparison study	79.57%	60.53%	89.61%	72.25%
Experimental results		92.00%	55.20%	69.00%
Performance evaluation		83.75%	77.91%	80.72%
Real-world application		51.22%	87.50%	64.62%
Simulation		87.50%	93.33%	90.32%
Statistical analysis		76.92%	58.82%	66.67%
Survey/Questionnaire		86.44%	91.07%	88.70%
Theoretical analysis		100.00%	12.50%	22.22%
N/A		77.78%	94.71%	85.41%

The overall accuracy for the '*Challenges'* category was 70,05% as described in the

Table 7. The confusion matrix, represented in the Fig. 6, shows that the model had high error rates when '*prediction accuracy*' was found as a false positive. The cause was that the model could hardly distinguish between the problem that the researchers tried to solve and the challenges of the solution as they were presented in the abstract.

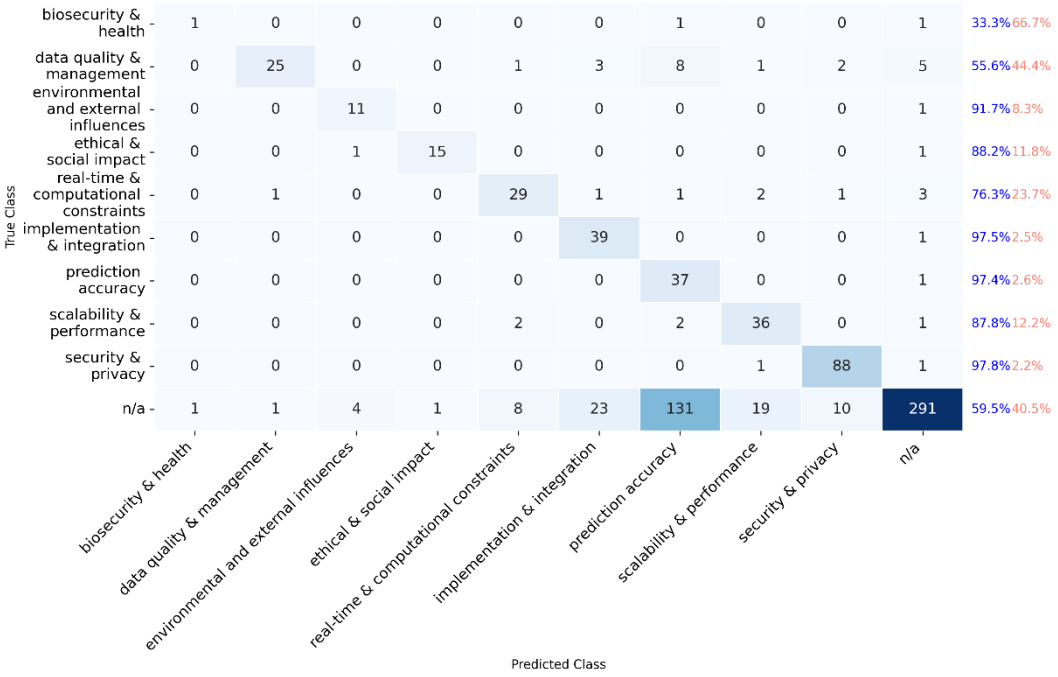

Fig. 6. Confusion matrix of the gpt-4o model for type of challenge

Table 7

Evaluation metrics of the gpt-4o model for type of challenge

Challenge type	Overall accuracy	Precision	Recall	F1-Score
Biosecurity and health	70.05%	50.00%	33.33%	40.00%
Data quality and management		92.59%	55.56%	69.44%
Environmental and external influences		68.75%	91.67%	78.57%
Ethical and social impact		93.75%	88.24%	90.91%
Implementation and integration		59.09%	97.50%	73.58%
Prediction accuracy		20.56%	97.37%	33.94%
Real-time and computational constraints		72.50%	76.32%	74.36%
Scalability and performance		61.02%	87.80%	72.00%
Security and privacy		87.13%	97.78%	92.15%
N/A		95.10%	59.51%	73.21%

This section provided a walkthrough of the results from the categorization of the GPT-4o model, based on standard metrics such as accuracy and F1-scores. During the next sections, results and conclusions are discussed.

4. Results analysis

Firstly, the overall accuracy varies from 63.19% to 88.60%, indicating inconsistency across the evaluation dimensions. The '*measurable improvement*' category has the lowest score, because of two factors, one being the lower base to which the accuracy is computed, less than half of the others, given the fact that 491 abstracts belong to just one subcategory – '*N/A*', and the other being that the abstract belong to multi-class labels, not single ones. Notably, the model has good performance in identifying the AI main technique (88.60%) and the scope of research (83.92%). This clearly shows that there is a strong degree of reliability in the automated evaluation. For the "*challenge type*" category, the model has a distinct problem in being unable to distinguish between the research problem and the problems or challenges of the proposed solution. All the cases are relevant in our evaluation and serve as proof of the necessity of including a human in the loop.

Secondly, there were two types of general errors causing the poor results on automated analysis:

a. Information was present but not found. In this instance, while the information was available, it was missed by the automated analysis.
b. The information was present but misinterpreted. In this instance, the information was related to a different context from the one in question. This causes the answer to be unrelated to the question.

Thirdly, the F1 score shows an imbalanced detection across the categories. The score varies between 22%-100% and the model favours recall instead of precision in 29 subcategories of 52, the rest of them being equal. This means that, in general the model is more prone to misclassify than to under-detect. Modelling the problem as a multi-class classification, instead of a single-class classification would improve the scores, as many subcategories were overlapping each other for various contexts.

This shows that while the results are not perfect, when compared with the manual evaluation which may also include misannotations, the results are very promising. It is worth mentioning that, even though the predefined output was clearly defined, the answer sometimes differed, mostly when the subcategory could not be identified. For instance, instead of "*N/A*", the query returned: "*Not specified*", "*Not applicable*" or "-". This translated into more manual processing work during the aggregation phase.

5. Conclusions

This paper shows how reliable is the LLM technology when integrated into research activities, particularly during the performance of bibliometric analysis.

One of the most common research activities, text classification (abstracts), was performed in a comparative study between manual and automated evaluation. The results were very promising, with an overall accuracy rate of 77%. This type of analysis is not limited to the research activity, but extends to the entire spectrum of data analysis, independent of any domain. At this stage, LLM technology still requires manual intervention across its entire life cycle of usage. Either researchers are using the tool as a black-box tool, and they will invest more time in finding explanations for the results, or a white-box solution is designed as in this paper, and this brings flexibility but requires human design effort. Designing the evaluation as a multi-class categorization would bring more complexity. Another approach would be to make a more delineated set of subcategories, and this could be done iteratively with a higher precision if the subcategories are updated based on data.

A major limitation of this study is that the research databases rarely include the conclusions of papers. This information would have been an important accelerator, creating data-readiness for the AI technology.

Further research could examine results from this paper using learning techniques like few-shot learning, adding descriptions for the subcategories or extending the categories to include study scale, details on the methodology, and main results.

R E F E R E N C E S

[1] ‘How to conduct a bibliometric analysis: An overview and guidelines - ScienceDirect’. Accessed: Oct. 31, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0148296321003155

[2] M. Aria and C. Cuccurullo, ‘bibliometrix: An R-tool for comprehensive science mapping analysis’, Journal of Informetrics, 2017, doi: 10.1016/j.joi.2017.08.007.

[3] VOSviewer, ‘VOSviewer - Visualizing scientific landscapes’, VOSviewer. Accessed: Jun. 03, 2024. [Online]. Available: https://www.vosviewer.com//

[4] ‘CitNetExplorer: A new software tool for analyzing and visualizing citation networks - ScienceDirect’. Accessed: Nov. 03, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1751157714000662

[5] J. P. Romanelli, M. C. P. Gonçalves, L. F. de Abreu Pestana, J. A. H. Soares, R. S. Boschi, and D. F. Andrade, ‘Four challenges when conducting bibliometric reviews and how to deal with them’, Environ Sci Pollut Res, vol. 28, no. 43, pp. 60448–60458, Nov. 2021, doi: 10.1007/s11356-021-16420-x.

[6] ‘JPMorgan pitches in-house chatbot as AI-based research analyst’. Accessed: Nov. 03, 2024. [Online]. Available: https://www.ft.com/content/96dfec5f-4d5f-4c3e-8f66-ebd0dfc8392d

[7] V. Pereira, M. P. Basilio, and C. H. T. Santos, ‘pyBibX -- A Python Library for Bibliometric and Scientometric Analysis Powered with Artificial Intelligence Tools’, Apr. 27, 2023, arXiv: arXiv:2304.14516. doi: 10.48550/arXiv.2304.14516.

[8] ‘Elicit: The AI Research Assistant’. Accessed: Jul. 17, 2025. [Online]. Available: https://elicit.com/

[9] A. Vaswani et al., ‘Attention is all you need’. 2023.

[10] W. X. Zhao et al., ‘A survey of large language models’, 2023, arXiv preprint arXiv:2303.18223.
[11] ‘Scalability’, Wikipedia. Nov. 02, 2024. Accessed: Nov. 03, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Scalability&oldid=1254966937
[12] J. Kaplan et al., ‘Scaling laws for neural language models’. 2020.
[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, ‘Distributed Representations of Words and Phrases and their Compositionality’, Oct. 16, 2013, arXiv: arXiv:1310.4546. doi: 10.48550/arXiv.1310.4546.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, May 24, 2019, arXiv: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.
[15] A. Vaswani et al., ‘Attention is all you need’. 2023.
[16] ‘A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends - ScienceDirect’. Accessed: Nov. 04, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422013537
[17] X. Yi, E. Walia, and P. Babyn, ‘Generative adversarial network in medical imaging: A review’, Medical Image Analysis, vol. 58, p. 101552, Dec. 2019, doi: 10.1016/j.media.2019.101552.
[18] T. B. Brown et al., ‘Language Models are Few-Shot Learners’, Jul. 22, 2020, arXiv: arXiv:2005.14165. doi: 10.48550/arXiv.2005.14165.
[19] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, ‘Zero-Shot Learning -- A Comprehensive Evaluation of the Good, the Bad and the Ugly’, Sep. 23, 2020, arXiv: arXiv:1707.00600. doi: 10.48550/arXiv.1707.00600.
[20] L. Ouyang et al., ‘Training language models to follow instructions with human feedback’, Mar. 04, 2022, arXiv: arXiv:2203.02155. doi: 10.48550/arXiv.2203.02155.
[21] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, ‘Direct Preference Optimization: Your Language Model is Secretly a Reward Model’, Jul. 29, 2024, arXiv: arXiv:2305.18290. doi: 10.48550/arXiv.2305.18290.
[22] H. Touvron et al., ‘LLaMA: Open and Efficient Foundation Language Models’, Feb. 27, 2023, arXiv: arXiv:2302.13971. doi: 10.48550/arXiv.2302.13971.
[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, ‘How transferable are features in deep neural networks?’, Nov. 06, 2014, arXiv: arXiv:1411.1792. doi: 10.48550/arXiv.1411.1792.
[24] M. Brundage and et.al, ‘The malicious use of artificial intelligence: forecasting, prevention, and mitigation’, Feb. 2018, Accessed: Aug. 23, 2024. [Online]. Available: https://dataspace.princeton.edu/handle/88435/dsp01th83m203g
[25] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, ‘A Survey on Bias and Fairness in Machine Learning’, Jan. 25, 2022, arXiv: arXiv:1908.09635. doi: 10.48550/arXiv.1908.09635.
[26] M. Veale and R. Binns, ‘Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data’, Big Data & Society, vol. 4, no. 2, p. 2053951717743530, Dec. 2017, doi: 10.1177/2053951717743530.
[27] K. Crawford, ‘Artificial intelligence’s white guy problem’, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:69337537
[28] J. G. Wang, J. Wang, M. Li, and S. Neel, ‘Pandora’s White-Box: Precise Training Data Detection and Extraction in Large Language Models’, Jul. 14, 2024, arXiv: arXiv:2402.17012. doi: 10.48550/arXiv.2402.17012.
[29] Z. Ji et al., ‘Survey of Hallucination in Natural Language Generation’, ACM Comput. Surv., vol. 55, no. 12, p. 248:1-248:38, Mar. 2023, doi: 10.1145/3571730.
[30] R. OpenAI and others, ‘GPT-4 technical report’, ArXiv, vol. 2303, p. 08774, 2023.

[31] 'AI Model & API Providers Analysis | Artificial Analysis'. Accessed: Nov. 16, 2024. [Online]. Available: https://artificialanalysis.ai

APPENDIX

Example of a query for analysing an abstract to categorize the AI technique:

```
You are a strict extractor.

- You will get one research abstract and a list of allowed labels for the field "AI_Category".
- Use ONLY the abstract text. No inference, speculation, or assumption.
- Choose a label ONLY if you are 100% certain it is explicitly supported by the text.
- If there is any doubt or ambiguity, output "N/A".
- Respond exactly as JSON with keys: "Article_ID","AI_Category","Reason"

Abstract:
"Industrial cloud computing and Internet of Things have transformed the healthcare industry with the rapid growth of distributed healthcare data. Security and privacy of healthcare data are crucial challenges in the healthcare industry. This article proposes a novel technique using deep learning and blockchain techniques for electronic health record privacy-preservation. The processed dataset classified normal and abnormal users using the convolutional neural network approach. Then, by using blockchain integrated with a cryptography-based federated learning module, the abnormal users have been processed and removed from the database along with the accessibility for the health records. The simulation has been done in the Python tool and experimental results show that the model's classification results and performance are better than other existing techniques.  2005-2012 IEEE."

Allowed AI_Category = [General AI  AI applications into blockchain  Data mining  Deep learning  Digital twin Edge AI  Explainable AI  Federated learning  Generative AI  Machine learning]

Output JSON:
{
  "Article_ID": "1",
  "AI_Category": "<one label | N/A>",
  "Reason": "<quoted snippet or N/A>"
}
```

Example of the query response for categorizing the AI technique in an abstract:

```
{
  "Article_ID": "1",
  "AI_Category": "Federated learning",
  "Reason": "\"...by using blockchain integrated with a cryptography-based federated learning module, the abnormal users have been processed and removed from the database...\""
}
```

Example of a query for analysing an abstract to categorize the measurable improvement:

```
You are a strict extractor.

- You will get one research abstract and a list of allowed labels for the field
"Measurable_Improvement".
- Use ONLY the abstract text. No inference, speculation, or assumption.
- Choose a label ONLY if you are 100% certain it is explicitly supported by the text.
- If there is any doubt or ambiguity, output "N/A".
- The improvement must be proved by NUMBERS!
- Respond exactly as JSON with keys: "Article_ID","Measurable_Improvement","Reason"

Abstract:
"Industrial cloud computing and Internet of Things have transformed the healthcare industry with the
rapid growth of distributed healthcare data. Security and privacy of healthcare data are crucial
challenges in the healthcare industry. This article proposes a novel technique using deep learning and
blockchain techniques for electronic health record privacy-preservation. The processed dataset
classified normal and abnormal users using the convolutional neural network approach. Then, by
using blockchain integrated with a cryptography-based federated learning module, the abnormal
users have been processed and removed from the database along with the accessibility for the health
records. The simulation has been done in the Python tool and experimental results show that the
model's classification results and performance are better than other existing techniques.  2005-2012
IEEE."

Allowed Measurable_Improvement = [Accuracy & performance metrics  Speed/efficiency/cost
Energy efficiency  Sales & growth  Error rate reduction  User satisfaction  Prediction & forecasting
Model performance  Security & privacy Healthcare & diagnosis  Productivity & operations
Classification performance  Network & data]

Output JSON:
{
  "Article_ID": "1",
  "Measurable_Improvement": "<one label | N/A>",
  "Reason": "<quoted snippet or N/A>"
}
```

Example of the query response for categorizing the measurable improvement in an abstract:

```
{
  "Article_ID": "1",
  "Measurable_Improvement": "N/A",
  "Reason": "\"The simulation has been done in the Python tool and experimental results show that
the model's classification results and performance are better than other existing techniques.\""
}
```