A Machine Learning and Large Language Model-Integrated Approach to Research Project Evaluation

Ma, Jian; Zheng, Zhimin; Zhu, Peihu; Liu, Zhaobin

Published in:
Journal of Database Management

Published: 15/06/2024

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.4018/JDM.345400

Publication details:
Ma, J., Zheng, Z., Zhu, P., & Liu, Z. (2024). A Machine Learning and Large Language Model-Integrated Approach to Research Project Evaluation. *Journal of Database Management*, *35*(1). https://doi.org/10.4018/JDM.345400

Journal of Database Management
Volume 35 • Issue 1

A Machine Learning and Large Language Model-Integrated Approach to Research Project Evaluation

Jian Ma, City University of Hong Kong, Hong Kong
Zhimin Zheng, National Natural Science Foundation of China, China
Peihu Zhu, InnoCity.com Ltd., Hong Kong*
Zhaobin Liu, City University of Hong Kong, Hong Kong

ABSTRACT

Research project evaluation upon completion is one of the important tasks for research management in government funding agencies and research institutions. Due to the increased number of funded projects, it is hard to find qualified reviewers in the same research disciplines. This paper proposes a machine learning and large language model integrated approach to provide decision support for research project evaluation. Machine learning algorithms are proposed to compute the weights of key performance indicators (KPIs) and scores of KPIs based on the evaluation results of completed projects, large language models are used to summarize research contributions or findings on project reports. Then domain experts are invited to consolidate the weights and scores for the KPIs and assess the novelty and impact of research contribution or findings. Experiments have been conducted in practical settings and the results have shown that the proposed method can greatly improve research management efficiency and provide more consistent evaluation results on funded research projects.

KEYWORDS

Large Language Models, Machine Learning Algorithms, Peer Review Assessment, Research Project Evaluation

INTRODUCTION

Research project evaluation is an important task in government funding agencies and research institutions (Wang et al., 2013). Key performance factors used in evaluating research projects include: 1) scientific merits on the novelty and impact of the research project; 2) relevance of the research to the mission and priorities of the funding scheme; and 3) output of the research project, that is, academic publications, awards, invention patents, and collaborations with other researchers, institutions, and industries.

Peer review methods are widely used in evaluating research projects, where domain experts are invited to evaluate the scientific merit, relevance, and research outputs of completed research projects (Bence & Oppenheim, 2004). Scientific review panels may be invited to make final decisions on the

DOI: 10.4018/JDM.345400 *Corresponding Author

undecided evaluation results. Hence, reviewer assignment plays a significant role in controlling the quality of project evaluation (Thushari et al., 2014), and many approaches have been proposed to support reviewer assignment (Liu et al., 2016), such as heuristic algorithm (Cook et al., 2005), hybrid knowledge, and model approach (Sun et al., 2008). As a main type of research project evaluation method, metrics-based approaches are also used to evaluate research projects based on specific metrics, such as the number and quality of publications, awards, and patents generated by the research project.

Research information systems, a decision support system for government funding agencies to support research management, are developed to track the progress of research projects, monitor research outputs, and evaluate the impact of research projects. These systems are used to streamline the evaluation process, ensure transparency, and facilitate knowledge sharing among researchers in universities and industries. With the increased number of funded research projects, it is becoming difficult to find relevant peer reviewers in the subject disciplines to evaluate the research projects, as a result, the evaluation results may be inconsistent. Current research information systems mainly focus on quantitative analysis that evaluates research projects based on statistical information of their research outputs (Donovan, 2007). The existing research evaluation approaches ignore the qualitative assessment on the novelty and impact of research projects and their relevance to the funding scheme.

This paper aims to propose an integrated approach that leverages machine learning techniques and large language models for the evaluation of research projects. Machine learning methods are employed to calculate the weights and scores of key performance indicators (KPIs) for quantitative analysis. Large language models are utilised to assist in processing text contents and then summarizing the research contributions or findings of the projects for qualitative assessment.

Experiments have been conducted in practical settings, and the results have shown that the proposed decision support framework can assist decision-makers in achieving more consistent evaluation results for funded research projects. It significantly improves the efficiency of research management work.

LITERATURE REVIEW

Research Project Evaluation Approaches

Methods for research project evaluation can be classified into quantitative and qualitative approaches. Quantitative evaluation methods evaluate research projects with objective measures using bibliometric analysis, citation analysis, and publication and citation counts (Bornmann & Marx, 2014; Glänzel & Schoepflin, 1999; Moed, 2006). The bibliometric approach conducts quantitative analysis with bibliographic data to examine patterns of publication, citation, and collaboration within a specific research discipline. It utilises bibliographic databases and citation indexes to identify relevant publications and track their citations over time (Haunschild & Bornmann, 2023). Hence, this method assesses the scholarly impact and influence of research projects by analysing publication patterns, citation networks, and author collaborations. Citation analysis focuses specifically on the citations received by a research project, a paper, or a thesis (Onwubiko et al., 2023). It involves identifying and analysing the references cited in scholarly articles, books, or other publications to assess the influence and visibility of a research project (Glänzel & Schoepflin, 1999). Citation analysis can reveal the extent to which a research project has been cited by other researchers, indicating its impact and contribution to the scholarly literature. Publication and citation counts refer to the quantitative measures of the number of publications and citations received by a research project. They can be derived from bibliographic databases, citation indexes, or academic search engines.

The previous methods mainly focused on indicators related to paper publication. However, these methods lack attention to the transfer of project achievements, talent cultivation, and other aspects. Additionally, machine learning methods are widely used to evaluate an item based on its many features.

For example, predicting whether a paper has influence (Weis & Jacobson, 2021) and predicting the emotions of a text (Chau, 2020). Currently, there are few machine learning methods used to predict the rating level of a scientific research project based on numerous indicators of the project. In particular, research project evaluation requires more interpretability and trust. Thus, machine learning-based methods for research project evaluation can be combined with experts.

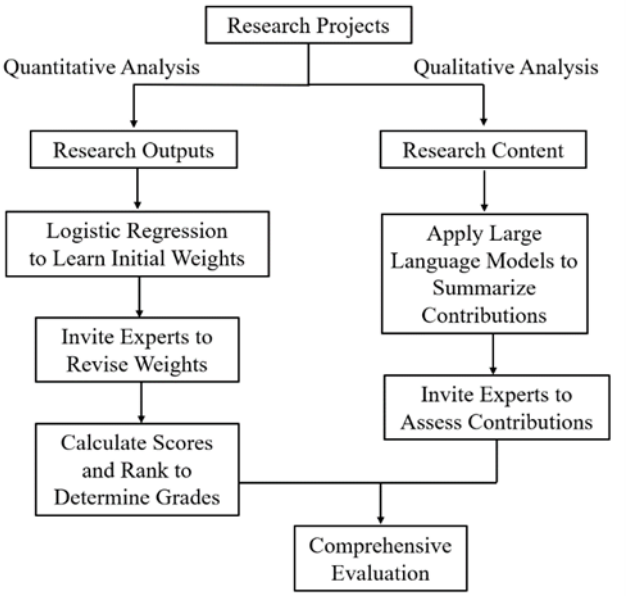
Qualitative evaluation methods primarily focus on subjective judgments, such as peer review (Franzoni & Stephan, 2023), expert judgment (Dahanayake et al., 2003), case studies (Baker et al., 2023; Merisalo-Rantanen et al., 2005), and surveys/interviews (Chang et al., 2023; Reale et al., 2007; Van den Besselaar & Sandstrom, 2016). The peer review method involves inviting domain experts to assess the quality of a research project. Similarly, the expert judgment methods rely on the insights and opinions of domain experts. Case studies involve conducting in-depth analyses of individual research projects. Surveys and interviews are conducted to gather direct feedback and data from researchers, stakeholders, or beneficiaries in order to assess a research project (Dobing & Parsons, 2008). However, the subjective judgments inevitably contain human bias, which can lead to inconsistent and unfair evaluation results. The increased number of research projects also bring a huge challenge to find experts to conduct research evaluations.

Large Language Models in Qualitative Analysis

Large language models have undergone significant development in recent years, revolutionizing the field of natural language processing and artificial intelligence. These models, such as generative pretrained transformer 4.0 (GPT-4), which is an emerging framework for generative artificial intelligence (Jovanovic & Campbell, 2022), have been trained on massive amounts of text data to acquire a deep understanding of language and the ability to generate coherent and contextually relevant text. The powerful ability of natural language processing (NLP), which enables computers to understand human language, such as language text, makes large language models become valuable tools in qualitative analysis. For text analysis, large language models can help analyse and categorise textual data, extract key themes or concepts, and identify patterns or trends within a data set (AlYahmady & Alabri, 2013; Jackson et al., 2019). For sentiment analysis, large language models can be employed to analyse the sentiment or emotional tone expressed in text. By inputting text data into the model, we can obtain insights into the overall sentiment of a document, identify positive or negative sentiments associated with specific topics, or track sentiment changes over time (Pang & Lee, 2008). For topic modelling, large language models can be utilised to identify topics or subject areas within a collection of texts. By applying techniques, such as clustering or latent Dirichlet allocation (LDA), researchers can extract underlying themes or topics from a corpus of qualitative data (Liu & Zuo, 2021; Wu et al., 2023). Large language models can aid in this process by providing topic suggestions or by generating representative examples for each topic (Blei et al., 2003). For text generation and summarization, large language models can generate human-like text, making them useful for generating summaries or paraphrases of qualitative data. This capability can be particularly helpful in synthesizing large amounts of qualitative data into concise summaries (See et al., 2017).

Current peer review methods heavily rely on experts from various disciplines. As the number of funded research projects continues to rapidly increase, finding relevant peer reviewers within the same discipline has become increasingly challenging. Additionally, the evaluation results may be inconsistent due to different preferences of experts in different research disciplines. The advancement of large language models with great power of automatically analysing text contents free from human interference and bias provides a great opportunity in assisting reviewers to conduct intelligent qualitative analysis. The aim of this paper is to propose an integrated approach for decision-makers in applying machine learning methods to perform quantitative analysis and using large language models to perform qualitative analysis in evaluating completed research projects.

Figure 1. A decision support framework with qualitative and qualitative analysis



INTEGRATION OF MACHINE LEARNING METHODS WITH LARGE LANGUAGE MODELS

Research Information systems for research project evaluation play a critical role in ensuring that the evaluation process is efficient, transparent, and effective. These systems support the work of reviewers, help to manage data related to research projects, and facilitate collaboration among various users. Machine learning algorithms have been shown to be good at identifying patterns and relationships within complex data sets and then making accurate predictions. They can contribute to the evaluation of research projects based on objective criteria. Large language models have advanced natural language processing capabilities to understand and summarise complex textual information, which can help identify key contributions and findings from research projects. Hence, a decision support framework is proposed in Figure 1. It uses machine learning methods for quantitative analysis of completed research projects and uses large language models for qualitative analysis of research projects.

Quantitative Analysis With Machine Learning

In accordance with the rules and regulations of a major government funding agency in China, a specific set of project output indicators has been selected for the funding scheme. This set includes five primary indicators and eight secondary indicators that are determined by experts of the government funding agency. The five primary indicators encompass scientific and technological awards, academic papers, patent technologies, academic activities, and talent cultivation. The secondary indicators encompass award quality, paper quality, paper citations, patent quality, technology conversion amount, activity numbers, participate numbers, and talent quality.

In the era of big data, machine learning methods have been widely adopted in various industry sectors, such as insurance (Jones & Sah, 2023) and music (Karthik et al., 2022). Specifically, machine learning can be used to predict a score for an item based on the features of this item. Research project evaluation requires more trust. Thus, besides machine learning-based methods, this paper also considers expert knowledge. First, we propose to use the supervised machine learning methods that learn the relationship of features and labels from the labelled data set to obtain the weights of

each feature, which is the output indicator of a research project in this paper. Then, domain experts are invited to adjust the weights assigned by the machine learning method.

Since the evaluation result of a research project is a 5-scale rating (dependent variables), and eight secondary output indicators (independent variables) are used for prerating prediction, we choose the multinomial logistic regression method to build the model and obtain the weights corresponding to each independent variable. The input of the model is an 8-dimensional vector $X = \left[x_1,\ldots,x_8\right]$, where $x_1,\ldots,x_8$ represent the values of the eight independent variables. The output of the model is a 5-dimensional vector $Y = \left[y_{(1)},\ldots,y_{(5)}\right]$. If the evaluation result of a completed project is 1, then $y_{(1)} = 1$ and the other dimensions being 0. Thus, the final output Y for a completed project is [1,0,0,0,0].

A logistic regression model shows an excellent performance of prediction and strong ability of explanatory. In this study, we use a multinomial logistic regression model (Menard, 2002) to assess the significance of research output indicators (independent variables) in relation to the evaluation result of a research project (dependent variables). The model is represented by the equation (1).

$$P(y_{(i)} = 1 \mid X, W) = \frac{\exp\left(W_{(i)}^T X\right)}{\sum_{j=1}^{5} \exp\left(W_{(j)}^T X\right)} \tag{1}$$

where $W_{(i)}$ is the estimated weight of output indicators to a research project with *i* as the evaluation results.

First, we can obtain the original observed values corresponding to each indicator from the project completion statistics table. Then, we train evaluation models for each discipline separately based on the above values. Figure 2 shows the initial weights computed by the multinomial logistic regression model based on the ratio of coefficients of the output indicators of a discipline's projects.

Figure 2. Initial weights of output indicators


Initial Setting of Metrics' Weight ×

Initial Setting of Metrics' Weight

Primary Metrics	Initial Weight	Secondary Metrics	Initial Weight
S1. Scientific Award	10%	S1-1. Award Quality	10%
S2. Scientific Paper	50%	S2-1. Paper Quality	30%
		S2-2. Citation	20%
S3. Patent	20%	S3-1. Patent Quality	10%
		S3-2. Transferred Amount	10%
S4. Academic Activity	10%	S4-1. Activity Expenses	5%
		S4-2. Number of Participants	5%
S5. Talent Cultivation	10%	S5-1. Talent Level	10%

Save Cancel

Based on the computed weights, subject panel experts are invited to contribute and consolidate the weights as shown in Figure 3, which adopts a voting approach to determine the final weights (Jiang et al., 2016).

The scores of key performance indicators for the research projects are also calculated from completed projects with evaluation scores given by the domain peer reviewers. For example, in the primary indicator 'Scientific Paper', the published journal ranking determines the scores of the indicator. As shown in Figure 4, the weight ratio of papers published in A-, B-, and C- category (classified based on Journal Citations Reports of Clarivate) journals is 6:3:1.

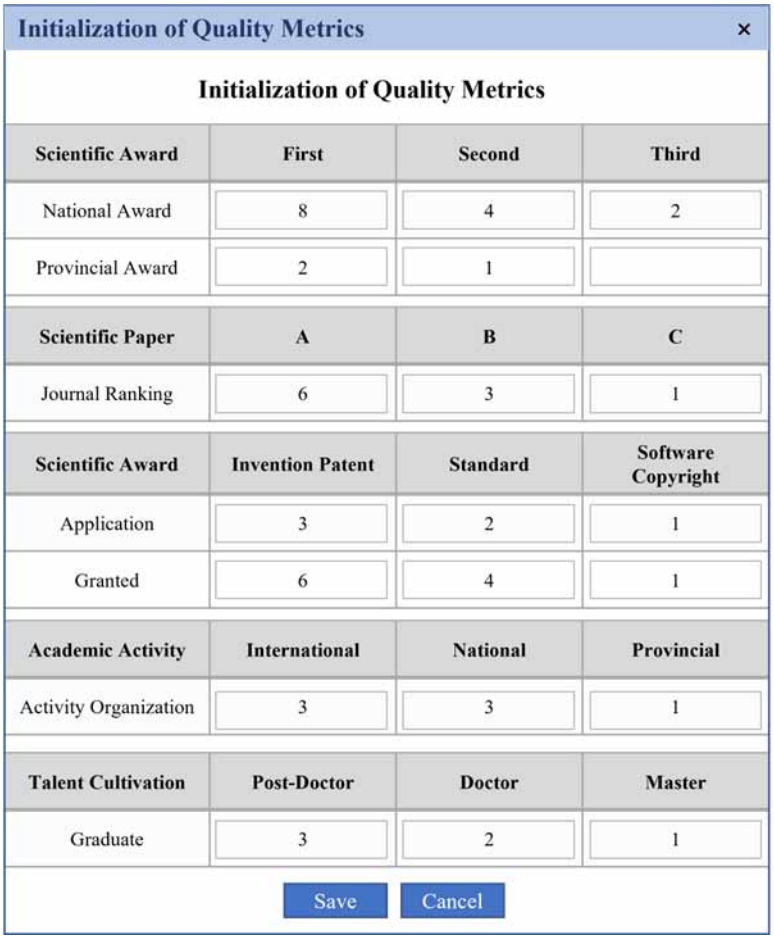
Finally, the weights assigned to KPIs and the scores obtained for each indicator are used to calculate the overall score for each completed research project. Using the evaluation results, performance scores can be plotted and visualised on a normal distribution curve. In accordance with requirements of research management, projects falling within the top 5% and bottom 15% are identified and selected for further action, such as arranging peer reviews. This approach aims to minimise the risks associated with potential misjudgements of project performance.

Qualitative Analysis With Large Language Models

To better evaluate research projects, this study also uses large language models to assist in summarising research contributions. The current large language models, such as GPT-4, are mainly developed based on general data and knowledge. They perform worse on specific tasks and specific domains, especially science, technology, and innovation that require a lot of domain knowledge. Hence, we will use big academic data to train and fine-tune a large language model, namely a special purpose large language model (STIGPT), which is a GPT model optimised for science, technology, and innovation using LIaMa 2, a powerful and open-source large language model. After developing the STIGPT with embedding domain knowledge, we design a specific prompt to train the STIGPT to summarise research contributions. The details of prompt design are presented in Figure 6. We first describe the context that defines the role of the model. A precise instruction is fed to the model, and output format is also defined. For data security and privacy concerns on research projects, we use data from academic papers with open access rights to illustrate how to construct the prompt, where the abstract/introduction of a paper is used as input, and contributions or findings discussed in the

Figure 3. Weights revision from experts

Figure 4. Weights of all output metrics



Initialization of Quality Metrics ×

Initialization of Quality Metrics

Scientific Award	First	Second	Third
National Award	8	4	2
Provincial Award	2	1	
Scientific Paper	**A**	**B**	**C**
Journal Ranking	6	3	1
Scientific Award	**Invention Patent**	**Standard**	**Software Copyright**
Application	3	2	1
Granted	6	4	1
Academic Activity	**International**	**National**	**Provincial**
Activity Organization	3	3	1
Talent Cultivation	**Post-Doctor**	**Doctor**	**Master**
Graduate	3	2	1

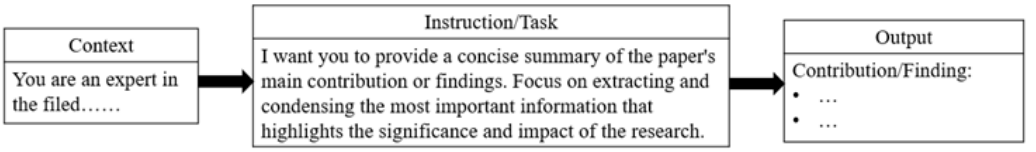
Save Cancel

Figure 5. Scores and assessment scales of each research project

Project Number	Scientific Award	Scientific Paper		Patent		Academic Activity		Talent Cultivation	Score	Grade Scale	Score Distribution
		Paper	Citation	Patent	Transfer	Activity	Participant				
PL00281	8	3	5	7	6	5	3	4	41	A+	Top5%
PL00102	5	6	8	4	3	4	4	3	37	A+	Top5%
...	...	...	...	...	...	...	...	...	...	...	...
PL00201	1	3	1	1	2	2	1	1	12	D	Bottom15%
PL00085	2	1	1	1	2	2	1	1	10	D	Bottom15%

conclusion part are used as answers. After obtaining the trained model, each output of a research project is qualitatively analysed by STIGPT with a summary of its main contribution or findings. The STIGPT can be implemented in local environment in government funding agencies to ensure data privacy and security. With the assistance of large language models, automatic summarisation saves

Figure 6. Prompt design of contribution summarizations



experts significant time and effort. Domain experts will be invited to assess the novelty and impact of research contribution or findings.

EXPERIMENT AND EVALUATION

In order to assess the reliability and accuracy of the proposed quantitative analysis approach, bibliographic data from Web of Science in the field of management science was collected for the past 5 years. From this data set, we were able to identify 1,421 completed projects that were funded by the National Natural Science Foundation of China (NSFC) based on the reported project numbers in the papers. To facilitate training and testing, a random selection process was employed, with 80% of the completed projects serving as the training data. This data set was used to learn and determine the weights assigned to the output indicators. The remaining 20% of the completed projects were designated as the testing data set, allowing us to predict the evaluation results of these projects.

Using the journal ranking data from the Chinese Academy of Sciences, the research papers were classified into five levels: A (Tier 1), B (Tier 2), C (Tier 3), D (Tier 4), and Other. To further evaluate the projects, we invited 10 experts in the same research discipline to rate these projects based on the quality of their published papers. Based on the ratings provided by these experts, the projects were ranked from highest to lowest scores and subsequently categorized into five levels: A+ (5%), A (20%), B (30%), C (30%), and D (15%).

In the training data set, the logistic regression machine learning method was used. The number of papers in each level was used as independent variable input, while the expert rating level served as the dependent variable output. The expert rating level was used to learn and analyse the impact of each indicator on the final evaluation score of the projects. Subsequently, the machine-learned weights were used as the default initial values. To fine-tune these weights, we invited the experts to make adjustments based on their expertise and insights.

Finally, the scores for each project were calculated in the test data set using the indicator weights confirmed by the experts, and they could be converted into evaluation levels. Precision, recall and F1-score were used to evaluate the effectiveness of the proposed method in this study. The specific definitions of these metrics were as follows (Taha & Hanbury, 2015):

$$Precision = \frac{\#\left\{\left(Prdicted\,Grading\right) \cap \left(Actual\,Grading\right)\right\}}{\#\left(Predicted\,Grading\right)}$$

$$Recall = \frac{\#\left\{\left(Predicted\,Grading\right) \cap \left(Actual\,Grading\right)\right\}}{\#\left(Actual\,Grading\right)}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}.$$

Table 1. Confusion matrix of prediction results

		Prediction				
		A+	**A**	**B**	**C**	**D**
Actual	A+	13	2	0	0	0
	A	2	48	7	0	0
	B	0	7	73	6	0
	C	0	0	6	80	0
	D	0	0	0	0	43

Table 2. Performance of the proposed method

Grading Level	Number of Projects	Percentage	Results		
			Precision	**Recall**	**F1-score**
A+	15	5%	0.87	0.87	0.87
A	57	20%	0.84	0.84	0.84
B	86	30%	0.85	0.85	0.85
C	86	30%	0.93	0.93	0.93
D	43	15%	1.00	1.00	1.00
Total	287	100%	0.90	0.90	0.90

According to the experimental results, the proposed method demonstrated excellent performance overall. Both precision and F1-score are 0.9, particularly for the last 15% of projects, the proposed method achieved a perfect accuracy of 100%. Additionally, the top 5% of projects rated as A+ also show outstanding performance, reaching an impressive performance at 0.87. Hence, the intelligent decision support framework to research project evaluation can effectively assist decision-makers in identifying the top 5% and bottom 15% of completed projects. Compared with an expert-based evaluation approach, the proposed method can obtain the evaluation results in a few minutes with high accuracy at 0.9. Hence, it can improve the work efficiency and provide more consistent evaluation results for funded research projects without human bias.

The evaluation of the proposed method and the system for research project evaluation in government funding agencies and research institutions was important to ensure that the systems are effective, efficient, and meet the needs of users. We invited experts to evaluate the proposed system across multiple dimensions, including usability, functionality, security, and performance. Based on their assessment, the system was deemed to be user-friendly, intuitive to navigate, and equipped with clear instructions on its usage. The data security and user privacy of the system are critical to the protection of sensitive data. The system has appropriate security measures in place to protect against unauthorized access, data breaches, and other security threats. The system can also be integrated with a research information system to streamline the evaluation process and reduce duplication of effort. The system has the capability of handling a large volume of data and users, and it can perform tasks quickly and efficiently.

CONTRIBUTIONS

This paper has made contributions in both theory and practice.

In terms of theoretical contribution, unlike previous studies that focused on automated evaluation methods based on the indicators related to the paper, this article pays extra attention to indicators, such as project achievement transformation and talent cultivation. We propose adopting an evaluation system that includes five primary indicators and eight secondary indicators. The five primary indicators encompass scientific and technological awards, academic papers, patent technologies, academic activities, and talent cultivation. The secondary indicators encompass award quality, paper quality, paper citations, patent quality, technology conversion amount, activity numbers, participate numbers, and talent quality. To integrate all the scores of a project under these indicators and enhance the interpretability of integration, we adopt the machine learning method to automatically assign weights to these indicators, thereby assisting experts in generating personalised understanding of each discipline and adjusting the weights assigned by the machine learning method for application in evaluation work. To evaluate from the perspectives of quantity and quality, we further utilise large language models to assist experts in summarising projects while analysing project quantity indicators. Based on the summary of STIGPT, experts can more efficiently analyse and judge the quality of a project. Finally, we proposed a method that integrates machine learning and LLM to assist in project evaluation.

In terms of practical contribution, the automation method proposed in this article can more efficiently assist project review experts in completing project evaluation work.

CONCLUSION

This paper presents a machine learning and large language model integrated approach to research project evaluation. Machine learning methods are used to calculate the weights of KPIs and scores of KPIs, aiming to assist decision-makers in evaluating completed research projects across various research disciplines. The STIGPT is also fine-tuned to summarise research contributions on completed project reports. Domain experts are invited to consolidate the weights and scores for the KPIs and evaluate the novelty and impact of the research contributions on the completed projects. The quantitative analysis and qualitative analysis can provide decision support services for decision-makers in evaluating completed research projects. Experiments are conducted in practical settings, and the results show that the proposed method achieves 0.9 at both precision and F1-score on quantitative analysis. The results have demonstrated that the proposed decision support framework can assist decision-makers in achieving more consistent evaluation results for funded research projects. It significantly improves the efficiency of research management work.

Research project evaluation plays an important role in assessing the effectiveness of research funding allocations. The current evaluation approaches are mainly based on experts, which inevitably bring human bias. The increased number of funded research projects also brings a huge burden for expert-based evaluation. This research proposes to use machine learning and large language model techniques to support research project evaluation. It streamlines the evaluation process, reduces administrative burden, enhances transparency, increases accountability, and builds trust in the evaluation process. This provides insights for government funding agencies on using artificial intelligence technologies to improve research management.

While experiments have been conducted and results evaluated using data sets from a specific funding scheme within one research discipline, future research will be conducted to expand the scope of experiments to include a diverse range of projects from different research disciplines. This will ensure that the machine learning models are robust and effective in evaluating completed research projects in different research disciplines. In addition, the proposed STIGPT is still being developed due to the limited data and computing sources. The future research will focus on fine-tuning the STIGPT

specialised for summarising research contributions and invite experts to evaluate the qualitative analysis. At last, other possible metrics will be explored and included in the model to conduct a comprehensive evaluation of research projects.

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING STATEMENT

No funding was received for this work.

PROCESS DATES

Received: December 13, 2023, Revision: March 21, 2024, Accepted: April 9, 2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Peihu Zhu, zhupeihu@gmail.com

REFERENCES

AlYahmady, H. H., & Alabri, S. S. (2013). Using NVivo for data analysis in qualitative research. *International Interdisciplinary Journal of Education, 2*(2), 181–186. https://www.semanticscholar.org/paper/Using-Nvivo-for-Data-Analysis-in-Qualitative-AlYahmady-Alabri/6a007b8b30daa0c55c775482792a9e6b019b1f9d

Baker, M. R., Jihad, K. H., Al-Bayaty, H., Ghareeb, A., Ali, H., Choi, J.-K., & Sun, Q. (2023). Uncertainty management in electricity demand forecasting with machine learning and ensemble learning: Case studies of COVID-19 in the US metropolitans. *Engineering Applications of Artificial Intelligence*, *123*, 106350. https://www.sciencedirect.com/science/article/abs/pii/S0952197623005341. doi:10.1016/j.engappai.2023.106350

Bence, V., & Oppenheim, C. (2004). The influence of peer review on the research assessment exercise. *Journal of Information Science*, *30*(4), 347–368. https://journals.sagepub.com/doi/10.1177/0165551504045854. doi:10.1177/0165551504045854

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937

Bornmann, L., & Marx, W. (2014). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgments of experts? *Journal of Informetrics*, *8*(1), 121–131. https://www.sciencedirect.com/science/article/abs/pii/S1751157715000073

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://dl.acm.org/doi/10.1145/3641289. doi:10.1145/3641289

Chau, M., Li, T. M., Wong, P. W., Xu, J. J., Yip, P. S., & Chen, H. (2020). Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *Management Information Systems Quarterly*, *44*(2), 933–955. https://www.researchgate.net/publication/343179922_Finding_People_with_Emotional_Distress_in_Online_Social_Media_A_Design_Combining_Machine_Learning_and_Rule-Based_Classification. doi:10.25300/MISQ/2020/14110

Cook, W. D., Golany, B., Kress, M., Penn, M., & Raviv, T. (2005). Optimal allocation of proposals to reviewers to facilitate effective ranking. *Management Science*, *51*(4), 655–661. https://pubsonline.informs.org/doi/10.1287/mnsc.1040.0290. doi:10.1287/mnsc.1040.0290

Dahanayake, A., Sol, H., & Stojanovic, Z. (2003). Methodology evaluation framework for component-based system development. [JDM]. *Journal of Database Management*, *14*(1), 1–26. https://www.semanticscholar.org/paper/Methodology-Evaluation-Framework-for-System-Dahanayake-Sol/9c40ecc0f6d06ac41f736388fc2a236b6d3704f9. doi:10.4018/jdm.2003010101

Dobing, B., & Parsons, J. (2008). Dimensions of UML diagram use: A survey of practitioners. [JDM]. *Journal of Database Management*, *19*(1), 1–18. https://www.semanticscholar.org/paper/Dimensions-of-UML-Diagram-Use%3A-A-Survey-of-Dobing-Parsons/17ccf01fe2fa027fb2c67d5d536b8019852ae526. doi:10.4018/jdm.2008010101

Donovan, C. (2007). The qualitative future of research evaluation. *Science & Public Policy*, *34*(8), 585–597. https://www.researchgate.net/publication/270188675_The_qualitative_future_of_research_evaluation. doi:10.3152/030234207X256538

Franzoni, C., & Stephan, P. (2023). Uncertainty and risk-taking in science: Meaning, measurement and management in peer review of research proposals. *Research Policy*, *52*(3), 104706. https://www.sciencedirect.com/science/article/abs/pii/S004873332200227X. doi:10.1016/j.respol.2022.104706

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management*, *35*(1), 31–44. https://www.sciencedirect.com/science/article/abs/pii/S0306457398000284. doi:10.1016/S0306-4573(98)00028-4

Haunschild, R., & Bornmann, L. (2023). Identification of potential young talented individuals in the natural and life sciences: A bibliometric approach. *Journal of Informetrics*, *17*(3), 101394. https://www.sciencedirect.com/science/article/abs/pii/S1751157723000196. doi:10.1016/j.joi.2023.101394

Jackson, K., & Bazeley, P. (2019). *Qualitative data analysis with NVivo*. Sage. https://us.sagepub.com/en-us/nam/qualitative-data-analysis-with-nvivo/book261349

Jiang, H., Yang, C., Ma, J., Silva, T., & Chen, H. (2016). A social voting approach for scientific domain vocabularies construction. *Scientometrics*, *108*(2), 803–820. https://link.springer.com/article/10.1007/s11192-016-1990-6. doi:10.1007/s11192-016-1990-6

Jones, K. I., & Sah, S. (2023). The implementation of machine learning in the insurance industry with big data analytics. *International Journal of Data Informatics and Intelligent Computing*, *2*(2), 21–38. https://www.researchgate.net/publication/371794479_The_Implementation_of_Machine_Learning_In_The_Insurance_Industry_With_Big_Data_Analytics. doi:10.59461/ijdiic.v2i2.47

Jovanovic, M., & Campbell, M. (2022). Generative artificial intelligence: Trends and prospects. *Computer*, *55*(10), 107–112. https://ieeexplore.ieee.org/document/9903869. doi:10.1109/MC.2022.3192720

Karthik, V., Chaudhary, S., & Radhika, A. (2022). Feature extraction in music information retrival using machine learning algorithms. *International Journal of Data Informatics and Intelligent Computing*, *1*(1), 1–10. https://www.researchgate.net/publication/370768696_Feature_Extraction_in_Music_information_retrival_using_Machine_Learning_Algorithms. doi:10.59461/ijdiic.v1i1.11

Leydesdorff, L., & Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the Association for Information Science and Technology*, *67*(11), 2766–2772. https://onlinelibrary.wiley.com/doi/10.1002/asi.21450

Liu, F., & Zuo, M. (2021). Learn from the rumors: International comparison of COVID-19 online rumors between China and the United Kingdom. [JDM]. *Journal of Database Management*, *32*(3), 46–68. https://www.igi-global.com/article/learn-from-the-rumors/282444. doi:10.4018/JDM.2021070103

Liu, O., Wang, J., Ma, J., & Sun, Y. (2016). An intelligent decision support approach for reviewer assignment in R&D project selection. *Computers in Industry*, *76*, 1–10. https://www.sciencedirect.com/science/article/abs/pii/S0166361515300610. doi:10.1016/j.compind.2015.11.001

Menard, S. (2002). *Applied logistic regression analysis* (No. 106). Sage. https://us.sagepub.com/en-us/nam/book/applied-logistic-regression-analysis-0

Merisalo-Rantanen, H., Tuunanen, T., & Rossi, M. (2005). Is extreme programming just old wine in new bottles: A comparison of two cases. [JDM]. *Journal of Database Management*, *16*(4), 41–61. https://www.researchgate.net/publication/220373787_Is_Extreme_Programming_Just_Old_Wine_in_New_Bottles_A_Comparison_of_Two_Cases. doi:10.4018/jdm.2005100103

Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer Science & Business Media. https://link.springer.com/book/10.1007/1-4020-3714-7

Onwubiko, E. C., Okeke, I. E., & Nwosu, O. (2023). Citation analysis of serials in graduate-students' thesis: A functional tool for effective serials management in university libraries. *International Journal of Library and Information Science Studies, 9*(4), 45–60. https://www.researchgate.net/publication/371398508_Citation_Analysis_of_Serials_in_Graduate-Students'_Thesis_A_Functional_Tool_for_Effective_Serials_Management_in_University_Libraries

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval, 2*(1–2), 1–135. https://ieeexplore.ieee.org/document/8187070

Reale, E., Barbara, A., & Costantini, A. (2007). Peer review for the evaluation of academic research: Lessons from the Italian experience. *Research Evaluation*, *16*(3), 216–228. https://academic.oup.com/rev/article-abstract/16/3/216/1588181?redirectedFrom=fulltext&login=false. doi:10.3152/095820207X227501

Silva, T., Jian, M., & Chen, Y. (2014). Process analytics approach for R&D project selection. [TMIS]. *ACM Transactions on Management Information Systems*, *5*(4), 1–34. https://dl.acm.org/doi/10.1145/2629436. doi:10.1145/2629436

Sun, Y. H., Ma, J., Fan, Z. P., & Wang, J. (2007, January 3–6). *A hybrid knowledge and model approach for reviewer assignment* [Conference session]. 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07) IEEE, Waikoloa, HI, USA. https://ieeexplore.ieee.org/document/4076467

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, *15*(1), 1–28. https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-015-0068-x. doi:10.1186/s12880-015-0068-x PMID:26263899

Van den Besselaar, P., & Sandstrom, U. (2016). Early career grants, performance, and careers. *Research Evaluation*, *25*(4), 363–374. https://www.sciencedirect.com/science/article/abs/pii/S1751157715300067

Wang, J., Xu, W., Ma, J., & Wang, S. (2013). A vague set based decision support approach for evaluating research funding programs. *European Journal of Operational Research*, *230*(3), 656–665. https://www.sciencedirect.com/science/article/abs/pii/S0377221713003597. doi:10.1016/j.ejor.2013.04.045

Weis, J. W., & Jacobson, J. M. (2021). Learning on knowledge graph dynamics provides an early warning of impactful research. *Nature Biotechnology*, *39*(10), 1300–1307. https://www.nature.com/articles/s41587-021-00907-6. doi:10.1038/s41587-021-00907-6 PMID:34002098

Wu, G., Miao, Z., Hu, S., Wang, Y., Zhang, Z., & Bao, X. (2023). Semi-supervised event extraction incorporated with topic event frame. [JDM]. *Journal of Database Management*, *34*(1), 1–26. https://dl.acm.org/doi/10.4018/JDM.318453. doi:10.4018/JDM.318453