Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.33527

Applications of Large Language Models in Data Analysis

Ruiqi Zou

School of Engineering and Technology, The University of Newcastle, Singapore, Singapore
c3543468@uon.edu.au

Abstract. With the background of the ever-increasing data volumes, data analysis becomes even more significant in the context of scientific studies and business decisions. Nevertheless, the conventional data analysis approaches usually consist of programming and statistical expertise, which to a certain degree restricts their fields of application. In this paper, the author explores the use of large language models in the analysis of data, performing a systematic analysis of the data analysis process in four phases: data processing, data planning, data reasoning and data feedback. The study reveals that big language models can automatically undertake actions like cleaning data and feature engineering, thus efficiently boosting the processing of data. Moreover, they are capable of breaking down more complicated analytical challenges to create systematic analytical processes. The reasoning ability of the models has been highly improved by utilizing the architecture of Transformer and prompt-based learning techniques. Moreover, with the addition of a reinforcement learning mechanism through human feedback (RLHF), the model is further optimized in performance and improved in the interpretation of results. Despite still having issues with inference stability, the accuracy of results and training costs, large language models show that they have enormous benefits in reducing the entry threshold to data analysis and improving the efficiency of the analysis process, which promises them a promising future.

Keywords: large language models, data analysis, RLHF, natural language processing, code generation

1. Introduction

The amount of data has literally gone skyrocketing in the era of big data and data analysis has become an essential instrument in supporting scientific research, business decisions and social governance. In financial forecasting, medical analysis, smart manufacturing or internet applications, the quality of made decisions and system performance directly depends on the possibilities of carrying out efficient data analysis. Nevertheless, conventional data analysis processes are often based on programming languages and statistical model development that require expertise and technical skills of the users. To non-expert users, the overall process, particularly data cleaning, to model building, will certainly have a steep learning curve, which partly restricts the general adoption of data analysis technologies.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.33527

As the technology of artificial intelligence continues to evolve, one of the most crucial intermediaries between the tasks of natural language processing and data analysis is slowly becoming large language models [1]. These models can be trained on large-scale corpora, using the Transformer architecture, and can learn complex text linguistic patterns and semantic relations, thus gaining strong text comprehension and text generation abilities [2]. This functionality can help the models not only to carry out the traditional tasks of natural language processing, but also be a major contributor in the field of code generation, intelligent question-answering, and data analysis support.

The use of large language models in data analysis has become an increasingly popular topic in the last few years. On one hand, large language models are able to perceive user needs by means of natural language and converting abstract analytical problems into particular data processing and modelling steps [3]; on the other hand, they can produce executable code, which automates such processes as data cleaning, feature engineering, and model training. This is a method of interaction based on natural language that has greatly reduced the entry barrier in the data analysis process allowing a greater number of non-expert users to be involved in the data analysis process. Simultaneously, it is possible to use large language models by users with certain technical knowledge as auxiliaries to enhance the effectiveness of analytics and minimize redundant work.

Despite the promising nature of large language models in data analysis, they have some issues related to the practical sphere. To illustrate, in complex problems, the reasoning process of the model can be unstable, logical fallacies can occur [4]; in processing structured data, the model can still be unable to comprehend data relations, and finally, the performance of the model is also affected by the training data and wordings of prompts, which can also influence the credibility of the analytical findings [5]. As such, there remains a need to conduct systematic review and analysis of the use of large language models in data analysis.

On this basis, this paper takes a data analysis workflow approach in classifying the use of large language models into four steps: data processing, data planning, data reasoning, and data feedback. It logically discusses their functions and approaches at every level and provides an overview of the significant development and current issues in the current studies. The paper will seek to offer recommendations towards further development of the large language models in data analysis through an analysis of the applicable technologies and application scenarios.

2. Data processing

Data processing is a core process in the data analysis process, which includes mainly data cleaning, feature engineering and conversion of data formats. Good quality data processing can go a long way towards improving the accuracy and reliability of the subsequent analysis and modeling.

Huge language models have shown good automatization of data processing tasks over the last few years. Studies show that intelligent agents powered by large language models have the ability to automatically clean tabular data, such as missing values, outliers, and standardizing data formats. Such approaches propel the data processing process with natural language instructions, in which users can accomplish the data preparation stage without composing intricate code, thus reducing the technical barrier to data analysis [6].

In massive data processing conditions, preprocessing of data and feature engineering can be often dependent on effective system support. Relevant studies have suggested distributed feature engineering pipelines that are specifically designed to be used in training large language models, which scale to large processing and feature manipulation of large data by building scalable, real-time data processing systems. Such a solution does not only enhance the data processing efficiency but

DOI: 10.54254/2753-8818/2026.33527

also facilitates dynamic data updates which give constantly optimized data inputs to train the model [7].

Overall, big language models, along with automated agent technology and distributed data processing systems, exhibit good potential applications in data cleaning and feature engineering. Nevertheless, when it comes to complex data, the stability and accuracy of processing results are still a point of interest that needs to be researched in the future.

3. Data planning

One of the important stages in data analysis is data planning. It mainly entails the definition of the analysis goals, decomposition of activities into steps, identification of suitable methods, and the structuring of the general sequence of analysis steps. Data planning, in comparison to data processing, puts more emphasis on the order of the tasks: what to do first and what to do next and decides whether the whole process of analysis is understandable, rational and effective.

Rahman et al. research indicates that data science agents that utilize large language models are progressively becoming capable of planning. These systems are able to automatically generate task objectives, using user-entered natural language queries, and break down challenging data analysis tasks into several executable sub-tasks, including data collection, data preprocessing, feature selection, model building, and result evaluation [8]. This feature shows that large language models can not only write code or respond to questions, but also engage in a more advanced way in the process of structuring and organizing data analysis processes.

The practical uses of data planning of large language models are shown in two aspects. First, the model allows creating a fairly detailed analysis roadmap on the basis of the description of the problem, which allows users to create an analytical framework rapidly. Second, the model has a certain level of coherence within multi-step tasks, which facilitates the analysis process closer to the real-world data science processes [9]. This planning ability can reduce the entry barrier to data analysis, as it allows users with no professional experience to plan a design; and it can also increase the efficiency of the early planning of the design, when experienced users are involved.

Nevertheless, there are some limitations of such methods. To begin with, the workflows of analysis produced by large language models are occasionally idealized and might not necessarily match the needs of real-world data situations. Second, the models can be ineffective in complex tasks or ones that demand high domain knowledge since they can be missing steps, illogical sequence, or lack depth in planning [8]. Thus, the existing data planning properties of large language models are more appropriate as a supplementary resource as opposed to a full substitution of manual analysis design.

All in all, big language models already show great potential to be used during the data planning stage. They can convert natural language requests to clear steps in analysis, which offer a starting point of future data processing, reasoning, and feedback. As the technology of agent systems, multi step reasoning and domain knowledge integration evolves further, the data planning service of large language models is likely to keep expanding.

4. Data inference

4.1. Mainstream research methods

The existing use of large language models in data analysis tasks is mainly based on the following types of methods: Transformer pre-trained models, prompt learning methods, and model

DOI: 10.54254/2753-8818/2026.33527

optimization methods.

4.2. Transformer pre-trained models

The majority of large language models are built on the Transformer model. This architecture relies on self-attention mechanisms to represent the relations among various words in text [10]:

This process allows the model to decode contextual information, and the performance in language comprehension and code generation is good [11]. As an example, the GPT-series of models have been shown to perform exceptionally well on a wide range of language tasks [10,12,13].

Transformer models, however, are computationally burdensome with respect to long texts. In order to solve this problem, certain researches have suggested sparse attention or model compression techniques in order to enhance efficiency [11].

4.3. Prompt learning methods

In recent years, prompt learning as a research direction in language modeling has become a popular focus which guides models to accomplish tasks through the design of input prompts. Wei et al. have suggested Chain-of-Thought (CoT) method that enhances the performance of the model on complex tasks by introducing reasoning steps in the prompts [14]. The experiments with the GSM8K dataset show that such an approach can be used to improve model accuracy.

The CoT approach however has unstable reasoning. To illustrate, in situations where the problem structure is complicated, the model can produce wrong reasoning steps hence influencing the ultimate outcomes.

To overcome this problem, Wang et al. suggested the Self-Consistency approach which improves the inference stability by producing multiple paths of inference and choosing the most consistent one [15].

5. Data feedback

5.1. RLHF training method

One of the most popular approaches that are currently employed to train large language models is RLHF (Reinforcement Learning from Human Feedback). Studies by Ouyang et al. have shown that the use of human feedback can greatly enhance the performance of a model in command-based tasks [16].

The general steps of RLHF are training a base language model, training a reward model, and optimizing model parameters with reinforcement learning. It has been experimentally demonstrated that models trained through RLHF significantly outperform them on complex tasks [16,17].

Nonetheless, RLHF training involves a lot of manual labelling of data, which is costly to train. Thus, certain studies have suggested using automatic feedback systems to make manual labeling cheaper and, consequently, enhance training effectiveness [17].

5.2. Standardized data and measurement scales

Researchers normally evaluate using standard datasets in data analysis activities. As an illustration, models are tested on their capacity to extract information in structured data with the use of the WikiTableQuestions and Spider datasets [3].

DOI: 10.54254/2753-8818/2026.33527

Besides, HumanEval is also popularly used in the task of code generation, where the goal is to assess the performance of models in generating program code. This data is used to measure the performance of the model by running the code generated and confirming the output.

The accuracy, code execution correctness, and text generation metrics like BLEU or ROUGE are commonly used as evaluation metrics.

6. Limitations and future research directions

Even though large language models have achieved certain advancements in the data analysis field, there are still a number of issues.

To start with, big language models can still be logically incorrect when confronted with complex reasoning tasks [4]. Second, the phrasing of prompts can easily affect model outputs, resulting in inconsistent analysis results [15]. Moreover, the training data is very sensitive to the performance of large language models, and the model performance can deteriorate when the tasks can be related to domain-specific knowledge [18].

The use of large language models in data analysis tasks can be further improved in future research in several respects. First, enhancing the reasoning mechanisms of the model the researcher can make the model logically consistent when performing the complex data analysis tasks. As an example, the combination of structured data processing techniques with the language model can enhance the model at the expense of grasping data relationships. Secondly, researchers can investigate the inclusion of knowledge graphs or domain-specific knowledge in the model training process to enhance the performance on specialized data analysis tasks. Moreover, further studies may help to decrease the cost of manual annotation by optimizing RLHF techniques or adding automated feedback processes, which will improve the effectiveness of training. With the further development of the related technologies, the large language models possess wide applicability opportunities in automated data analysis sphere.

7. Conclusions

In this paper, the systematic analysis of the application process of large language models to data analysis tasks is given with particular attention paid to four stages: data processing, data planning, data reasoning and data feedback. This paper shows that not only can large language models assist with simple data processing tasks but are also important to the design of analytical workflows, complex reasoning, and analysis of findings, greatly increasing the extent of automation and ease of access in data analysis.

Despite the fact that the current approaches have achieved certain progress in generating the code and supporting data analysis, they remain to be deficient in the stability of complex reasoning, result reliability, and the domain-specific knowledge ability. Moreover, the high cost of such a training method relying on human feedback can enhance the performance of the model but cannot be further expanded due to the expense.

Enhancing the usability of large language models in data analysis in future studies can be done in a variety of directions, including by improving the reasoning capabilities of the models themselves, improving their integration with structured data systems, and introducing automatic feedback mechanisms to decrease training costs. As the associated technologies continue to evolve, large language models will have a more significant role in intelligent data analysis and decision support systems.

DOI: 10.54254/2753-8818/2026.33527

References

[1] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv: 2108.07258.
[2] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv: 2303.18223.
[3] Shen, L., Shen, E., Luo, Y., Yang, X., Hu, X., Zhang, X., & Tai, C. L. (2022). Towards natural language interfaces for data visualization: A survey. IEEE Transactions on Visualization and Computer Graphics, 29(6), 3121–3144.
[4] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv: 2303.12712.
[5] Zhang, H., Li, J., Wang, Y., et al. (2023). Integrating automated knowledge extraction with large language models for explainable medical decision-making. In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE.
[6] Bendinelli, T., Dox, A., & Holz, C. (2025). Exploring LLM agents for cleaning tabular machine learning datasets. arXiv preprint arXiv: 2503.06664.
[7] Deva, S. (2025). Scalable real-time feature engineering pipelines for large language model training: A distributed systems approach. International Journal of Applied Mathematics, 38(6s), 206–229.
[8] Rahman, M., Bhuiyan, A., Islam, M. S., et al. (2025). Llm-based data science agents: A survey of capabilities, challenges, and future directions. arXiv preprint arXiv: 2510.04023.
[9] Raza, M., Jahangir, Z., Riaz, M. B., et al. (2025). Industrial applications of large language models. Scientific Reports, 15(1), 13755.
[10] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
[11] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv: 2302.13971.
[12] OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv: 2303.08774.
[13] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1–35.
[14] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.
[15] Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv: 2203.11171.
[16] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.
[17] Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2023). A survey of reinforcement learning from human feedback. arXiv preprint arXiv: 2312.14925.
[18] Liu, Q., Yang, R., Gao, Q., et al. (2024). A review of applying large language models in healthcare. IEEE Access, 13, 6878–6892..