Reduce AI Illusion Based on Data Science Technology and Prompt Engineering

Yannan Li

Brunel London School, North China University of Technology, Beijing, China

victor_ncut@outlook.com

Abstract. Since the advent of artificial intelligence (AI) systems such as ChatGPT, artificial intelligence has been widely used with its advantages of low cost and efficiency, but it has also brought negative effects such as academic plagiarism and fake news, especially the "AI illusion". Data science improves the performance of large language models and reduces the generation of error information through cleaning, preprocessing, and data enhancement techniques. The research direction of this paper starts from the data level and the algorithm level. At the data level, this paper discusses data cleaning and data preprocessing to demonstrate how to improve the accuracy, diversity and quality of data, optimize the performance of AI large models, and thus reduce the generation of AI illusions. At the algorithmic level, this paper introduces the role of prompt engineering, which guides the model to generate more accurate output by optimizing the problem design and prompt statements, and further reduces the risk of AI illusion. Finally, it is concluded that data science technology and prompt engineering play a key role in reducing the illusion of AI. Future research could delve deeper into data collection and model evaluation. At the same time, the algorithm level prompt engineering and alignment techniques also need to be further studied to enhance the robustness and reliability of the model. By multiple techniques, the performance of AI systems will be continuously optimized, effectively reducing the generation of illusions.

Keywords: AI Illusion, Data Science Technology, Data Cleaning, Data enhancement, Prompt Engineering.

1. Introduction

Since the advent of ChatGPT and other artificial intelligence (AI) systems, artificial intelligence has been widely used in many fields, and its ability to generate content has been widely recognized with its advantages of low cost, efficiency and convenience. However, the wide application of AI systems has also brought many negative effects, such as: 1. Academic plagiarism. For example, when ChatGPT is used to write an academic article, ChatGPT will copy the sentences of other academic articles, resulting in infringement of intellectual property rights of others, legal disputes and affecting the academic reputation of individuals [1]. 2. False news, such as: some criminals will use ChatGPT to write rumors about floods, and publish and spread this false information on the Internet. Profiting from it and being widely viewed [2]. 3. Impact on scientific research and ethics of science and technology. For example, the frequent wrong answers of ChatGPT reduce the rigor and science of academic research, and further promote the spread of false information to bring great harm to society. At the same time, plagiarism and other behaviors involved in the information generated by ChatGPT also affect the entire academic ecology. In particular, the use of the monopoly position of ChatGPT's creators by large corporations raises questions of technological ethics and value penetration [3].

These problems not only bring potential dangers to society, but also may be exploited by lawbreakers, resulting in the spread of large-scale misinformation. The false information generated by AI systems, the so-called "AI illusion," has attracted widespread attention from academia and society [4]. At the same time, with the advent of the era of big data, data science technology has been deeply integrated into many fields such as information science, economics, and network science, and has become a core component of the field of AI. Especially in the application of large language models, data science technology can effectively analyze model performance and improve the performance of the model through data cleaning and data pre-processing, data enhancement, and diversification, thereby reducing the occurrence of AI illusions [5]. This provides an important technical means for reducing the error information generated by AI systems and has a wide range of application significance.

Research status: According to the existing research, AI illusion is mainly caused by the error and noise of training data, so data quality plays a key role in the output accuracy of the model. To reduce the production of AI hallucinations, the researchers optimized the output of large language models in two ways, 1. Data level; Improve data quality through data cleaning and preprocessing and data enhancement. 2. Algorithm level: Through prompt engineering, optimize problem design and guide the output of large language models. At the same time, alignment techniques ensure that the AI's output meets expectations. The above several approaches are the key measures to solve the AI illusion [6].

The work of this paper: This article is mainly from the data science and technology (data level) and prompt engineering (algorithm level) two levels. At the data level, through data cleaning, data preprocessing and data enhancement, these two methods improve the data quality and diversity, and improve the model performance of large language models (taking ChatGPT as an example), so as to reduce the generation of AI illusion. From the algorithm level, through the prompt statement prompting the project, the large model is guided to output relevant information, so as to reduce the AI illusion.

2. Using data science technology to reduce the generation of AI illusions and related cases

2.1. Data cleaning and data preprocessing

Data quality control: Discuss how to remove noise and erroneous data through data cleaning to ensure that the model does not hallucinate with bad data during training. (Manual annotation to improve data quality) The importance of manual annotation to improve data quality: ChatGPT improves the performance of large language models and reduces hallucinations by feeding "new data" into its base model. These "new data" are mainly our academic and life knowledge and through manual annotation, these "new data" can help AI systems better understand the core ideas of human beings. And reflect the quality of human answers to AI. In InstructGPT, the annotator will annotate the optimal answers from prompt messages and template answers, in order to improve the accuracy, relevance, and authenticity of AI answers, and reduce the generation of irrelevant information by AI. After the "new data" is input into the model, the model will be optimized according to the input manually labeled data, and the new model has better "authenticity, beneficial, and no harm", and the probability of AI illusion is lower than that of the previous model.

2.1.1. Case studies of data cleaning and preprocessing to reduce AI illusions. Investigative News reported that OpenAI hired a group of low-paid, unstable workers in Kenya to complete the data annotation. Therefore, the quality of these data may not be able to meet the goals and expected results set by the work. Secondly, manual annotation requires personnel to concentrate their efforts, make full use of their cognitive functions to make the best judgment, and maintain the consistency of the quality of manual annotation information as much as possible. Therefore, scientists or scholars in the field of data science are required to carry out this work. Data quality directly affects the quality of the AI system to answer the "best answer", and directly affects the authenticity, accuracy and relevance of the information generated by ChatGPT, AIGC and other AI systems. Without high data quality, AI is more likely to hallucinate. [4]

Preprocessing strategy: describes how to process data (such as normalization, de-duplication, etc.) before training the data to reduce hallucinations.

2.1.2. Case study of data cleaning and preprocessing to reduce artificial intelligence illusions. In the medical field, the artificial intelligence model is based on medical images (x-rays, nuclear magnetism, etc.) and medical text data, such as unstructured data such as medical reports and disease statements, and removes irrelevant and harmful information through data cleaning and data annotation. Combined with medical papers to improve the quality of data. [7]

2.2. Data Enhancement

Data enhancement plays a key role in model training, especially in emotion classification task, which can effectively improve the generalization ability of the model. By introducing diversified training data, data enhancement helps large language models to be exposed to more information such as language expressions and affective categories during training, so as to have stronger adaptability in cross-domain tasks. In general, when a model is trained on a limited data distribution, it is prone to overfitting problems, resulting in poor accuracy of the content generated when it deals with tasks in a new domain. By using data-enhancement methods such as back-translation, synonym substitution, and random insertion or deletion of words, large language models are able to capture more detail in multiple forms of expression, effectively reducing the generation of AI illusions.

By introducing multi-domain data into a large language model, data enhancement not only increases the diversity of training data, but also enables the model to make more accurate judgments in different domains, avoiding the phenomenon of "AI illusion" due to domain bias. In addition, data enhancement can also improve the robustness of the model, making it more stable when dealing with complex or ambiguous emotional information, thus further improving the performance of the model in the generation task. [8]

3. Reduce the generation of AI illusion at the algorithm level

Prompt engineering concept: Method of guiding large models to output accurate, relevant information through prompt statements. There is no need to adjust model parameters.

The algorithm flowchart is shown in Figure 1:

Step 1: Enter an initial prompt, such as: Tell me about the multiple regression model.

Step 2: Evaluate the generated content of the model.

Step 3: Combine the evaluation results and optimization tips to add some specific direction requirements, such as: Please use multiple regression model for modeling and analysis according to the data set I gave you.

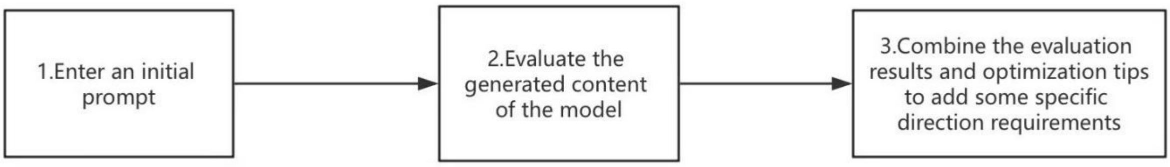
1.Enter an initial prompt
2.Evaluate the generated content of the model
3.Combine the evaluation results and optimization tips to add some specific direction requirements

Figure 1. algorithm flow chart

Optimization prompt engineering method: you can clear the direction of the problem, provide the corresponding scenario, guide the relevant example, gradually guide, set the scope, etc. These methods can effectively help AI understand and output accurate and true information. From the perspective of large models, these methods can improve the performance of large models from the algorithmic level and reduce the generation of AI illusions. There are several ways to further optimize prompt engineering.

(1) Clear the direction of the problem:

It can help ChatGPT better understand the results that humans expect, for example: "Explain the mathematical formula of Artificial Neural Network (ANN) forward propagation through the network and explain the meaning." ChatGPT generates a mathematical formula for the forward propagation of the ANN neural network, along with an explanation of the meaning of each step.

(2) Provide relevant scenarios:

Artificial intelligence such as ChatGPT can help us enrich or refine the language, for example: "I'm going to intern at a company, please write me a brief introduction and enrich the part of my personal experience." Artificial intelligence will generate a more experienced self-introduction template according to our needs.

(3) Related example guidance:

It can help AI better understand which areas of information are needed. For example, according to the previous question, "Please describe the development of AI in various fields," the second question, "Please describe the role and role of AI in the financial field."

(4) Gradual guidance method:

A large task can be divided into logical blocks, which form a logical chain, and each part has a desired output. For example: "I am going to write a report, please make a template for me" so that the AI will generate a report template, cleaner and clearer.

(5) Setting range method:

The content generated by AI is more precise and refined by setting the scope. For example: "Based on the above, please help me refine it. In 200 words, the main idea is clear." [9]

The above methods are widely used in large AI models such as ChatGPT, which is an important method in large language models, suggesting that the above methods can significantly improve the model performance of large language models in various tasks and effectively reduce the generation of AI illusion.

4. Challenges and Prospects

In terms of data, the future research direction can be mainly focused on how to quickly and efficiently collect high-quality data and diversified data to improve the performance of the model. However, the problem now is that access to large amounts of high-quality data is expensive, data is unbalanced, samples are scarce, and a host of data privacy, ethical and other issues remain unresolved. In terms of large models, researchers can discuss the role of cross-validation, leave-one validation and other evaluation methods in identifying and reducing model illusion. However, the use of these methods requires high computing resources and may not be efficient when processing large amounts of data. In terms of big data, analyze how to improve the robustness of models and reduce illusions by processing large amounts of data. In terms of algorithm, more in-depth research can be carried out from the aspects of Prompt engineering and alignment technology to further improve the performance of the model and reduce the generation of illusion. However, designing flexible prompt engineering strategies is a complex task. The prompt engineering can fine-tune different tasks, but information that is too detailed may not be understood by the model, resulting in content that is not as expected.

5. Conclusion

This paper demonstrates the key role of data science technology and prompt engineering in reducing the illusion problem of AI large models. By improving the accuracy, authenticity, and relevance of data quality, data science provides a solid foundation for AI models to generate accurate and high-quality information for complex tasks. With the widespread use of AI systems such as ChatGPT, data science is becoming increasingly important in optimizing model performance and reducing error generation. At the same time, prompt engineering improves the quality of AI-generated content and reduces the generation of irrelevant or incorrect information by optimizing the design of questions and guiding the direction of AI answers. In addition, alignment techniques better align the AI's output with the desired goals of humans to better conform to social and ethical standards. Creating specialized AI systems in specific domains further enhances the model's ability to handle specialized tasks, thus reducing the occurrence of AI illusions. Through the collaborative application of these technologies, the accuracy and reliability of AI systems will continue to improve, providing more trusted information support for various fields.

References

[1] Li, C. K. (2024). The Application and challenges of large language models in academic writing: A case study of ChatGPT. Research on the Integration of Industry and Education.

[2] Lu, J. P., Dang, Z. Q. (2024). Analysis on AIGC False Information Problem and Root Cause from the Perspective of Information Quality. Documentation, Information & Knowledge.

[3] Zheng, S. L. Yao, S. Y. Wang, C. F. (2023). ChatGPT The Economic and Social Impact of the development of New generation artificial intelligence technology Industrial economics review.

[4] Mo, Z. Y. P., Liu, D. Q., et al. (2023). Analysis on AIGC False Information Problem and Root Cause from the Perspective of Information Quality. Documentation, Information & Knowledge, http://dik.whu.edu.cn/jwk3/tsqbzs/CN/10.13366/j.dik.2023.04.032

[5] Yang, J., Wang, X. Y. Bai, R. J., Zhu, N. (2015). Institute of Science and Technology Information, Shandong University of Technology, Zibo 255049, China.

[6] Wang, Y. Z., Li, Q., Dai, Z. J., Xu, Y. (2024). Current status and trends in large language modeling research. Chinese Journal of Engineering DOI: 10.13374/j.issn2095-9389.2023.10.09.003

[7] Kang, Y. L., Guo, Q. Y. Zhang, W. Q. et al. (2023). Medical language model based on knowledge enhancement: current situation, technology and application. Journal of Medical Informatics

[8] Li, S. C., Wang, Z. Q., Zhou, G. D. (2022). A large language model-driven cross-domain attribution-level sentiment analysis. Computer Application and software.

[9] Huang, J., Lin, F., Yang, J. et al. (2024). From prompt engineering to generative artificial intelligence for large models: the state of the art and perspective Chinese Journal of Intelligent Science and Technology