# Robustness of Big Language Modeling in Finance

*Mohan* Yao

Economic and Management, Tiangong University, 300380, Tianjin, China

**Abstract.** With the gradual entry of artificial intelligence into all aspects of people's lives, people begin to use big language models to solve problems in various fields. In the financial field, people use financial big prediction models to solve problems such as stock prediction, risk assessment, etc., but the big language models can be incorrect due to model hallucination and adversarial attacks. Therefore, investigating the robustness of large language models in finance is the main topic of this article, and searches the literature using the keywords "large language model", "adversarial attack", "model illusion", etc. in recent years. We searched the literature in recent years. The existing literature explains the causes of adversarial attacks and model illusion, and methods that enhance the robustness of large language models are come up. It is shown that an attacker can trigger the model illusion of a large language model through an adversarial attack to reduce the reliability of the large language model. There is a lack of specific datasets of big language models in the financial domain to get a solution to improve the big language models in the financial domain in a better way. Future research should be specific in the financial domain for further adversarial training and robustness optimization of big language models in the financial domain.

## 1 Introduction

Whether the Big Language Model is reliable or not has become a matter of great concern. With the continuous development of artificial intelligence technology, the Big Language Model has entered everyone's life, and it can answer the questions people ask involving various fields and give its point of view. As the Big Language Model continues to grow, people are starting to rely on it more and more. People will let it predict stock movements, project risk assessment, and other problems in the financial field. However, the reliability of big language models is still questionable, big language models can come up with wrong answers due to adversarial attacks and modeling illusions, and how to improve the robustness of big language models in the financial field has become a key concern. Research on the robustness of large language models is currently focused on the following aspects: adversarial attack: literature [1] proposed the reason for the generation of adversarial samples, refuting the previous view that adversarial samples are due to the highly nonlinear nature of neural networks, and instead proposing that adversarial samples are mainly generated because of the linear behavior of the model in high-dimensional space.

Corresponding author: 1812010523@stu.hrbust.edu.com

The sensitivity of different models to adversarial samples is compared, and it is found that both linear models and neural networks are vulnerable to adversarial samples, while radial basis function (RBF)-based models exhibit greater robustness; Model illusion: the literature [2] investigates a number of different retrieval-enhanced generative architectures, including a retriever, a sequencer and an encoder-decoder, in two knowledge-driven dialog tasks( Wizard of Wikipedia and CMU Document Grounded Conversations), tested their proposed models on two knowledge-driven conversation tasks, and proposed an effective approach based on retrieval-enhanced generation that significantly reduces modeling illusions and improves the knowledge base and factual accuracy of large language models; Robustness: literature[3] discusses how to properly assess the defensibility of machine learning models against adversaria samples, first noting the importance of assessing defensibility as well as common motivations, emphasizing the importance of threat models in the assessment, and noting that correct assessing defensibility is critical to advancing research in the field of adversarial samples. Researchers should carefully consider the threat model and adopt a rigorous assessment method to ensure the effectiveness of the defense and better improve robustness. Since adversarial attacks are input samples intentionally designed by an attacker to produce erroneous results or irrational behavior in deep learning models through small but targeted modifications in financial modeling, if a model is given a wrong stock data, the results it gets are wrong, so what is the relationship between adversarial attacks and model illusions and how do they work together to make reliability decrease in large language models. However, it is not clear how model illusion and adversarial attack can make financial large language models less reliable and how to better improve the robustness of financial large language models, in order to clarify this problem, this paper will systematically analyze and review the relevant literature in this area recent years, this paper starts from two aspects of the model illusion and adversarial attack to explore the reasons that cause the reduction of model robustness and to find ways to improve the robustness of large language models.

## 2 Modeling illusion

### 2.1 Classification of modeled hallucinations

Model illusion refers to a situation where a model generates information or content that seems plausible but is not actually real when processing input. This usually occurs in generative models (e.g., language models, image generation models, etc.), especially when the model processing complex, fuzzy or ambiguous inputs, Literature [3] categorizes modeling illusions into four types 1. text 2. image 3, video 4. audio (as shown in fig. 1).

Literature [4] explores various factors that generate model hallucinations, including biases in training data, lack of real-time information in the model, and inherent limitations in model comprehension and generating contextual accuracy. The literature reviews existing strategies for mitigating model hallucinations, including the use of external knowledge bases, cue engineering techniques, and reinforcement learning based on user feedback. it concludes with a discussion of potential directions for future research, such as developing more robust evaluation metrics building more robust model architectures, and exploring new training approaches.

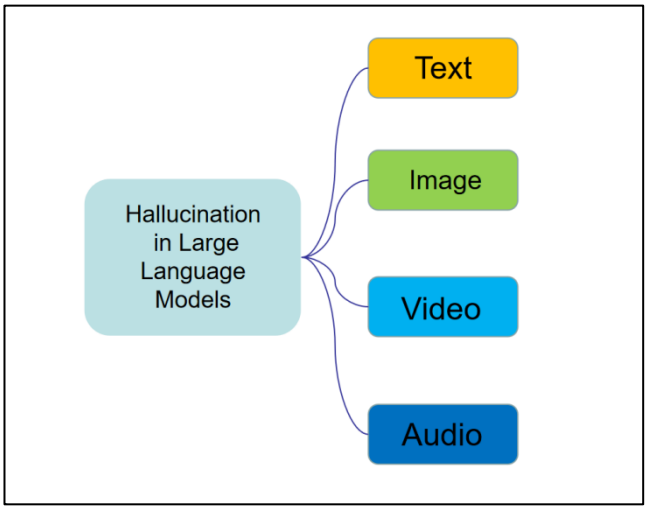


**Fig. 1.** Four classification of model hallucinations.

### 2.2 Modeling Illusions in Large Scale Visual Language Models

Large visual language models are one of the categories of large language models, and large visual language models, like all large language models, are susceptible to be cued, and in idealized hallucinatory scenarios, the assessment results do not match the hallucinations in real scenarios. Literature [5] argues that existing object based hallucination assessment methods do not accurately reflect the true hallucination situation of large visual language models. This literature proposes a hallucination assessment framework based on large language models (HaELM). HaEllis less costly and more reproducible than ChatGPT, and better protects privacy when it can be deployed locally. In terms of accuracy, HaELM still achieves about 95% of the accuracy of ChatGPT with the aforementioned advantages. However, the model still has some problems, for example, when HaELM recognizes certain types of hallucinations, it does not perform as well as ChatGPT because the simulated hallucination data used to train HaELM cannot fully cover all types of hallucinations in the real world. Although the subsequent evaluation cost of HaELM is low, the initial data collection and model training still require a certain amount of time and computational resources.

### 2.3 Metheds to mitigate modeling illusions

Literature [6] provides a comprehensive survey of more than 32 model illusion techniques aimed at mitigating large language models and categorizes them based on parameters such as data set utilization, common tasks, feedback mechanisms, and retriever types. The literature highlights a number of techniques for mitigating modeling illusions, such as, Retrieval Generation (RAG): an approach that enhances the factual accuracy of a large language model by allowing it to access an external knowledge base; and Knowledge Retrieval: a technique that focuses on detecting and correcting illusions timely during the course of generation. CoNLI and Cove: These methods utilize natural language feedback and self-improvement techniques to improve the reliability of the output that large language models suppose to get a result.

Whereas, literature [2] proposes a new approach based on retrieval enhancement generation to mitigate model illusion.S.M Towhidul Islam Tonmoy et al. investigated a

number of different retrieval enhancement generation architectures including Retrievers, Sequencers, and Encoder-Decoders, and proposed some new variants such as: using Poly-encoder Transformers for finer context-candidate document scoring; proposed an iterative retrieval scheme to enhance retrieval by repetition; used an end to end trained retriever in the Fusion-in-Decoder technique; and constructed a retrieval mechanism based on conversation rounds to avoid the problem that the standard retriever ignores most of the conversation contexts. S.M Towhidul islam Tonmoy et al. tested their proposed model on two knowledge driven conversation tasks (Wizard of Wikipedia and CMU Document Grounded conversations), and the experimental results showed that their proposed model significantly reduces the occurrence of hallucinatory phenomena especially when dealing with out-of-distribution topics and test data is more effective. But like RAG, FiD also suffers from the lack of relevance of the retrieved documents to the context of the conversation.

# 3 Countering attacks

## 3.1 Explanation of counterattacks

Adversarial attacks are deliberately designed input samples that are modified by small but targeted modifications in order to produce erroneous results or irrational behaviors in machine learning models. Literature [7] describes the great success of Deep Neural Networks (DNNs) in various machine learning tasks, but also points out their vulnerability to the security risks of adversarial sample attacks. Deep learning has achieved great success in the field of computer vision, but at the same time, research has shown that deep neural networks are susceptible to adversarial attacks, i.e., small perturbations to the inputs that cause the model to predict incorrect outputs. Adversarial attacks are a major threat that leads to errors in large language models. The paper defines common terms related to adversarial attacks such as adversarial samples, adversarial perturbation, black-box attack, white-box attack, etc.

## 3.2 Methods to counter attacks

Deep learning has achieved great success in the field of computer vision, but at the same time, research has shown that deep neural networks are vulnerable to adversarial attacks, i.e., small perturbations to the inputs that cause the model to predict incorrect outputs. Adversarial attacks are a major threat that leads to errors in large language models and literature [8] defines common terms related to adversarial attacks such as adversarial samples, adversarial perturbations, black-box attacks, white-box attacks, etc. Various adversarial attack methods for image classification tasks are reviewed, including optimization-based attacks (e.g., L-BFGS), gradient-based attacks (e.g., FGSM, BIM), saliency map-based attacks (e.g., JSMA), etc.

In the following, we will focus on an approach to countering attacks based on gradient attacks Literature [1] suggests that the main reason for the vulnerability of neural networks to confrontation samples is their linear nature, rather than nonlinearities or overfitting, as was previously thought.Ian J. Goodfellow, Jonathon Shlens et al. proposed a fast method for generating confrontation samples: the Fast Gradient Symbol Method (FGSM) ), which provides an efficient method for generating adversarial samples(as shown in fig. 2).

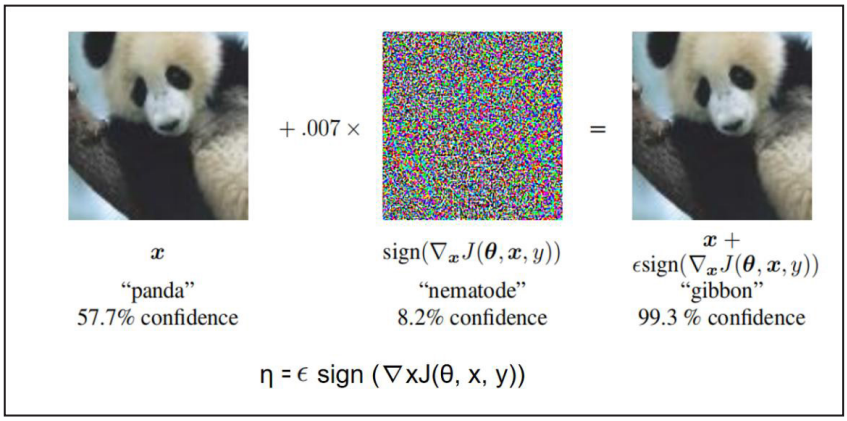


**Fig. 2.** Fast gradient notation method [1].

Literature [1] compares the effects of adversarial training and weight decay, and concludes that adversarial training is more effective. And the adversarial training is applied to deep networks as a regularization method, which improves the robustness of the model to adversarial samples.

### 3.3 Robustness optimization

Having understood what an adversarial attack is and how to perform it, this paper moves to the next stage - how to optimize for robustness? Literature [9] transforms the problem of attacking and defending against adversarial samples into a saddle-point (min-max) optimization problem, where the internal maximization problem corresponds to the attacker's search for adversarial samples and the external minimization problem corresponds to the training of robust classifiers. By optimizing this saddle-point problem, a model that is robust to a specific threat model can be obtained. Although the loss function of the saddle-point problem is non-convex and non-concave in nature, in practice, gradient-based optimization methods (e.g., Projected Gradient Descent PGD) are effective in finding a solution that is close to the maximum value, thus generating strong adversarial samples. Literature [9] conducted experiments on MNIST and CIFAR10 datasets, and successfully trained neural network models robust to various attack methods, including white-box and black-box attacks, through adversarial training methods.

## 4 The relationship between modeling illusions and confrontational aggression

Through the above literature review, this paper argues that the difference between model illusions and adversarial attacks is that model illusions are more related to internal logic or reasoning errors that may occur in the model when processing the available data, whereas adversarial attacks are inputs intentionally created against the model by an external attacker. But they have the similarities. This paper concludes by finding that despite the fact that adversarial attacks and model illusions have different sources and modes of influence, both model illusions and adversarial attacks are indicative of vulnerabilities and misbehavior that may occur in models when processing complex or unusual inputs. Both problems require techniques and approaches to enhance the robustness and security of models against possible errors or attacks. In order to address how the adversarial attacks and modeling illusions presented in this paper can lead to reliability degradation on financial large language models,

the methodology of the literature [10] is used to investigate adversarial attacks and modeling illusions to carry out experiments on the currently dominant large language models. Using ReEval, an evaluation architecture proposed in the literature [10], which dynamically generates new data for evaluating big language models by interfering with evidence in the cues in order to measure the reliability of the big language model in arriving at the correct response based on a provided set of contexts, ReEval accordingly provides two ways to synthesize the evaluation dataset(as shown in fig.3): (1) Answer swapping (category 1): On the premise of ensuring that context-dependent conditions do not change , a valid answer is replaced by the previous answers; (2) Context enrichment (category 2): enrich the context that it means add more information which has relationship with original text based on the basis.

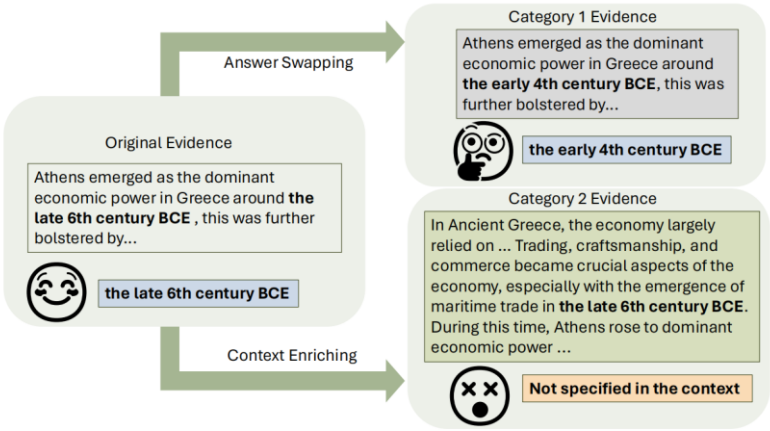


**Fig. 3.** Two approaches for ReEval synthetic evaluation dataset [10].

Category 1 is a situation in which the conditions related to the answer are changed, while Category 2 is a case where adding more information and making the context richer. Using ReEval can generate new test cases that make the large model create model illusion easily. By using the test cases generated by ReEval, the performance of various large language models was evaluated on two popular open-domain QA datasets (Natural Questions and Realtime QA). With this experiment it was found that models that performed well on static data produced unsupported answers on scrambled evidence and showed significant accuracy degradation on all LLMs including GPT-4.

## 5 Conclusion

In this paper, by reviewing the existing work on model illusion and adversarial attacks, we explain in detail the principles of model illusion and adversarial attacks, respectively, as well as how to improve the robustness of a large language model by mitigating model illusion and adversarial training, and propose that the relationship between the two, in a specific financial large language model, is to reduce the reliability of the financial large language model by eliciting a series of responses to model illusion through adversarial attacks. reliability of the model.

The current paper only assesses the reliability from the perspective of two different evidence perturbations, and there is currently no very well-developed dataset in the financial

domain to test the reliability of the financial grand prognosticator model, which still only stays with a few mainstream grand language models at present.

Future work needs to construct higher quality datasets in the financial domain, which currently lacks publicly available datasets dedicated to the training and evaluation of large language models in the financial domain. Future research could focus on constructing financial datasets that contain multi modal data such as stock prices, news events, company financial reports, etc., and cover the features of different markets and financial products to support more comprehensive and closer to real-world application scenarios for model training and evaluation. Explore the construction of financial datasets containing adversarial samples and model phantom annotations for specifically training models to recognize and resist such attacks. Study how to combine external knowledge bases such as financial domain knowledge graph and rule engine.It is important to enhance the ability of understanding that models can have and the financial knowledge about inference, so as to reduce the generation of model illusion and improve the interpretability and credibility of the model.

In conclusion,the fields that employ the large financial language model is more extensive, but also faces great challenges. Future research needs to focus on key issues such as model robustness and trust worthiness, and actively explore more effective solutions to promote the healthy development of FinTech.

## References

1. I. J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
2. S. Kurt, et al. "Retrieval augmentation reduces hallucination in conversation." arXiv preprint arXiv:2104.07567 (2021).
3. C. Nicholas, et al. "On evaluating adversarial robustness." arXiv preprint arXiv:1902.06705 (2019).
4. R. Vipula, A. Sheth, and A. Das. "A survey of hallucination in large foundation models." arXiv preprint arXiv:2309.05922 (2023).
5. J. Y. Wang, et al. "Evaluation and analysis of hallucination in large vision-language models." arXiv preprint arXiv:2308.15126 (2023).
6. S. M. Tonmoy, et al. "A comprehensive survey of hallucination mitigation techniques in large language models." arXiv preprint arXiv:2401.01313 (2024).
7. X. Han, et al. "Adversarial attacks and defenses in images, graphs and text: A review." International journal of automation and computing 17: 151-178. (2020).
8. A. Naveed and A. Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." Ieee Access 6: 14410-14430 (2018).
9. M. Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).
10. X. D. Yu, et al. ReEval: Automatic Hallucination Evaluation for Retrieval-Augmented Large Language Models via Transferable Adversarial Attacks. Findings of the Association for Computational Linguistics: NAACL 2024. (2024).