Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.34274

Pre-emptive Prediction of Inference Hallucinations in Large Language Models Based on Machine Learning Algorithms

Jingchen Yang

School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China
Jingchen_Yang23@163.com

Abstract. The reasoning illusion of large language model is the core bottleneck restricting its safe landing. The existing research mostly focuses on post-event detection and lacks pre-prediction mechanism. In order to improve the accuracy of illusion prediction and generalization performance, a BiLSTM-Attention classification algorithm based on RBMO optimization is proposed in this paper. The model uses a two-way long-short-term memory network to capture the pre-and post-sequence dependencies, focuses on key features with the help of attention mechanism, and introduces the red-billed blue magpie optimization algorithm to adaptively optimize the network weights and hyperparameters, so as to alleviate the local optimal problem of gradient descent. A standardized data set containing 1729 samples was constructed in the experiment, and a variety of machine learning algorithms were selected for comparison and verification. The results show that the proposed algorithm is optimal in accuracy, recall, accuracy, F1 value and AUC value, and all indexes are 0.975 or above, which is overall superior to traditional machine learning and single deep learning models, and can provide effective technical support for large model hallucination pre-prediction and safe deployment.

Keywords: Reasoning Illusion, RBMO, BiLSTM, Attention

1. Introduction

With the in-depth application of large language models in intelligent question answering, knowledge retrieval, professional assisted decision-making and other scenarios, the illusion problem in the process of model reasoning gradually becomes the key bottleneck restricting its reliability and practicality. The phenomenon of hallucination is that the model generates content that seems reasonable but contradicts objective facts, is not logically consistent or has no basis, which not only reduces the interaction credibility, but also may cause serious risks in professional fields such as medical, legal, and industrial control [1]. Current research shows that the generation of large model hallucinations is closely related to factors such as the parameter scale of the model itself, the length of inference context, the input task domain, the timeliness of knowledge, and the configuration of inference deployment. How to accurately predict the probability of hallucinations through quantitative features has become a core research topic in the fields of natural language processing and large model security [2]. The existing research mostly focuses on the post-event detection and optimization of hallucinations, lacking the pre-prediction mechanism for reasoning behavior

© 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: 10.54254/2755-2721/2026.34274

characteristics, and it is difficult to avoid risks in advance in the reasoning stage. Building an efficient hallucination prediction model has important practical significance to improve the landing safety of large models [3].

Machine learning algorithm provides an effective quantitative analysis path for large model hallucination prediction. By modeling model input features, inference configuration and context information, it can automatically mine the nonlinear correlation between features and hallucination phenomena, and realize the classification and discrimination of inference behavior [4]. Traditional machine learning methods such as random forest and support vector machine can complete basic feature classification, but their performance is limited when dealing with sequential features and long dependencies, and it is difficult to fully capture the influence of context length changes and dynamic features of inference process on hallucinations [5]. Deep learning models can better adapt to high-dimensional complex data by virtue of their powerful feature extraction capabilities. However, a single network structure often has problems such as insufficient feature coding and unreasonable weight allocation of key information, which makes it difficult for prediction accuracy and generalization performance to meet the actual needs. Therefore, designing a hybrid deep learning model that integrates optimization strategy and attention mechanism to improve the accuracy and stability of illusion classification has become an important direction to break through the limitations of existing methods [6].

In order to solve the problems of insufficient feature utilization and low classification accuracy of existing models in illusion prediction, a BiLSTM-Attention classification algorithm based on RBMO optimization is proposed in this paper. This method takes bidirectional long-short-term memory network as the core architecture, fully captures the before-and-after dependence between context length and inference behavior sequence, and automatically focuses on key features that have a significant impact on hallucination through attention mechanism, so as to enhance the model's perception ability of important information. At the same time, RBMO optimization algorithm is introduced to adaptively optimize network weights and hyperparameters, which avoids the defect that traditional gradient descent is easy to fall into local optimum, and improves the convergence speed and generalization performance of the model. The proposed algorithm organically integrates feature learning, weight optimization and classification decision-making, and can more accurately discriminate the probability of hallucination according to the input features and reasoning configuration of the model, providing an efficient and feasible technical solution for the pre-prediction of hallucination in large model reasoning, and also providing new research ideas for safe reasoning and trusted deployment of large models.

2. Data sources

In this paper, data collection is carried out for the task of large language model inference illusion classification and prediction, and then various features are standardized and outliers are eliminated. Finally, a standardized dataset containing 1729 real samples is formed. The dataset covers input features such as model parameter scale, context length, input domain, problem type, inference batch size, inference accuracy, knowledge deadline, whether to enable retrieval enhancement, and classification labels to identify whether the inference process produces hallucinations, which can be directly used for machine learning model training and hallucination classification and prediction research. First, a correlation analysis was performed, and a correlation heat map was plotted, as shown in Figure 1.

DOI: 10.54254/2755-2721/2026.34274



Figure 1. The correlation heat map

Violin plot analysis was performed on each variable and a violin plot was drawn as shown in Figure 2.

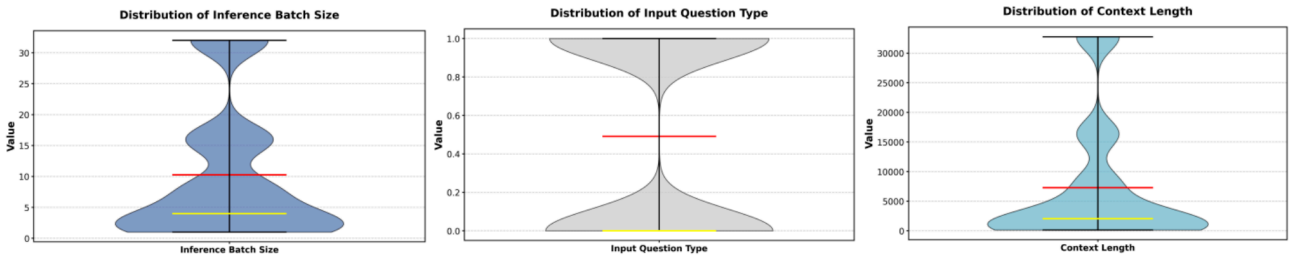


Figure 2. The violin plots

3. Method

3.1. RBMO

The red-billed blue magpie optimization algorithm is a novel meta-heuristic optimization algorithm proposed in 2024, which is inspired by the natural behavior of red-billed blue magpie colony cooperative predation, prey attack and food storage. The core of the algorithm is to simulate the cooperative mode of birds foraging in small groups or large groups, and find the optimal solution through information sharing and position iterative update. In the initialization stage, population

DOI: 10.54254/2755-2721/2026.34274

individuals are randomly generated, corresponding to candidate solutions in the search space, and parameters such as population size and maximum number of iterations are set [7]. Location update is divided into two core stages: foraging and attacking. Small swarm mode focuses on local fine search, large swarm mode strengthens global scope exploration, and dynamic balance algorithm exploration and development ability. In the attack stage, individuals move towards the current optimal solution, adaptively adjust the moving step size, expand the search range in the early stage of iteration, and focus on the deep development of high-quality solutions in the later stage. In the food storage stage, the optimal solution in the historical iteration is preserved, so as to avoid losing high-quality information and improve the convergence stability of the algorithm. Individual performance is continuously evaluated through the fitness function. If the termination condition is not reached, the position update and fitness calculation are repeated until the maximum number of iterations or convergence threshold is satisfied, and the global optimal solution is output [8].

3.2. BiLSTM

Bidirectional long-short-term memory networks are extended variants of long-short-term memory networks, and the core breakthrough is to capture both forward and reverse context dependencies of sequence data simultaneously. The basic LSTM unit solves the gradient disappearance problem of traditional recurrent neural networks through three gating structures: forget gate, input gate and output gate, and effectively maintains long-term sequence memory. The forgetting gate controls the information retention ratio of the cell state at the previous time, the input gate filters the written content of new information at the current time, and the output gate controls the degree of cell state conversion into hidden layer output. BiLSTM deploys two independent LSTM layers in parallel, and forward LSTM processes sequences from start to end in time, capturing historical to current dependency information. Backward LSTM processes in reverse time order to obtain the associated features from the future to the current, and the hidden states in the two directions do not interfere with each other. At each time step, the bidirectional hidden state vectors are stitched and fused to form a comprehensive feature representation containing both context and context information [9].

3.3. Attention

Attention mechanism is a dynamic weight allocation technology that simulates human visual attention. The core is to make the model automatically focus on key parts and weaken irrelevant contents when processing information. The mechanism constructs a computing framework based on three core vectors: query, key and value. The query vector represents the focus target of the current task, the key vector is the index identifier of the input feature, and the value vector is the original feature information to be weighted. The calculation is completed in three steps. First, the degree of correlation of the query with all keys is calculated through dot product or cosine similarity to obtain the original attention score [10]. Secondly, the score is scaled and divided by the square root of the key vector dimension, so as to avoid the gradient disappearing caused by excessive high-dimensional dot product result. Then normalize the score to weights between 0 and 1 through the Softmax function, ensuring that the sum of all weights is 1. This mechanism realizes dynamic feature screening, which greatly improves the efficiency and accuracy of the model processing long sequence data. The network structure of Attention is shown in Figure 3.

DOI: 10.54254/2755-2721/2026.34274

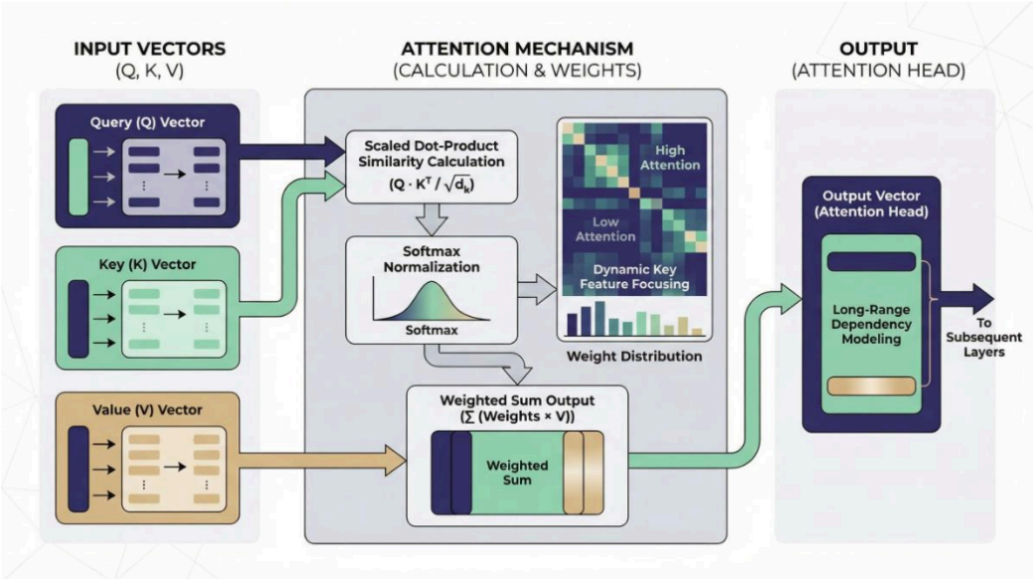


Figure 3. The network structure of Attention

3.4. RBMO-BiLSTM-Attention

RBMO-BiLSTM-Attention is a composite deep learning model that integrates optimization algorithm, sequence modeling and dynamic focusing, and collaboratively improves sequence prediction and classification performance through a three-layer architecture. The bottom layer takes the Red-billed Blue Magpie optimization algorithm as the core, which is responsible for optimizing the hyperparameters and network weights of the BiLSTM-Attention model. The RBMO algorithm encodes the learning rate, number of hidden units, Dropout rate and other parameters of the model into population individuals, and iteratively searches for the optimal parameter combination through foraging, attacking and food storage behaviors. The middle layer is a BiLSTM sequence encoder, which receives the preprocessed multidimensional timing or text data, and extracts the context features of forward history and reverse future in parallel by bidirectional LSTM. After two-way hidden state splicing, the sequence features rich in global dependencies are output, which solves the defect of incomplete information in one-way network. The top layer is the attention mechanism layer, which receives the BiLSTM output features and calculates the dynamic weight of each time step or feature dimension. Through weight allocation, the key features that have the greatest influence on the prediction target are highlighted, and the interference of noise and redundant information is suppressed. During model training, the RBMO algorithm continuously optimizes the model parameters and reduces the prediction loss function value. In the inference stage, the optimized BiLSTM-Attention model receives new data, and outputs accurate prediction results through bidirectional feature extraction and dynamic weight focusing. The three-layer structure collaboratively realizes intelligent parameter optimization, global context capture and key information focusing, and comprehensively improves the modeling accuracy and generalization ability of complex sequence tasks. The network structure of RBMO-BiLSTM-Attention is shown in Figure 4.

DOI: 10.54254/2755-2721/2026.34274

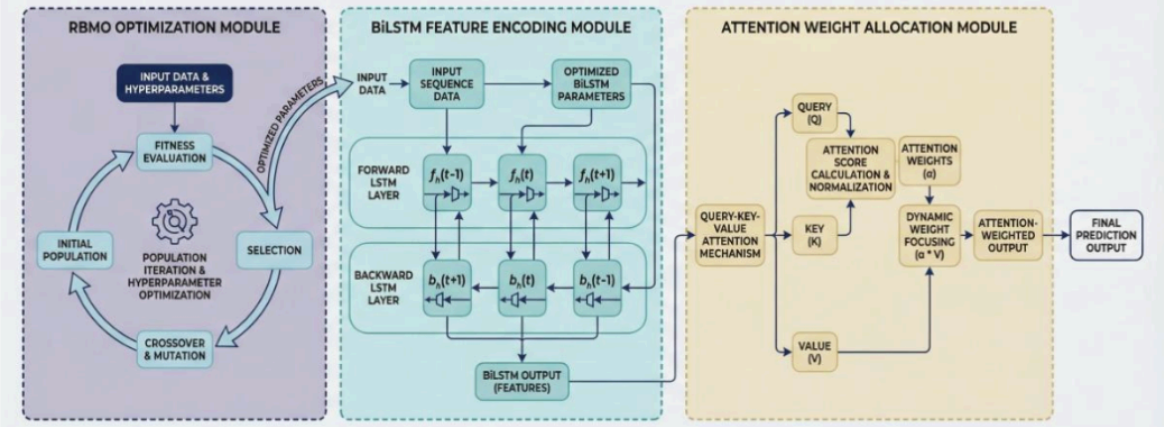


Figure 4. The network structure of RBMO-BiLSTM-Attention

4. Result

The optimization algorithm of this project sets the number of 12 search populations and the maximum number of 6 iterations. The lower limits of the three parameters to be optimized are 0.00001, 0.0001 and 10, and the upper limits are 0.01, 0.002 and 50, respectively. The maximum training times of neural network training are set to 500 times, the gradient threshold is set to 1, the learning rate is adjusted in segments with an attenuation factor of 0.2 after the 400th training, the discarded layer ratio is set to 0.01, and the proportion of training set to total samples in the dataset is 0.7.

Comparative experiments were performed using a variety of machine learning algorithms, and the results are shown in Table 1. Output the index comparison results of the RBMO-BiLSTM-Attention test set, as shown in Figure 5.

Table 1. The results of the comparative experiment

Model	Accuracy	Recall	Precision	F1	AUC
AdaBoost	0.927	0.927	0.927	0.926	0.962
BP neural network	0.765	0.765	0.768	0.766	0.891
Decision tree	0.954	0.954	0.954	0.954	0.937
CatBoost	0.956	0.956	0.957	0.956	0.969
Random forest	0.95	0.95	0.951	0.95	0.968
LightGBM	0.956	0.956	0.957	0.956	0.966
GBDT	0.95	0.95	0.951	0.95	0.971
XGBoost	0.96	0.96	0.96	0.96	0.974
ExtraTrees	0.946	0.946	0.947	0.946	0.973
Logistic regression	0.78	0.78	0.786	0.774	0.896
RBMO-BiLSTM-Attention	0.975	0.975	0.976	0.975	0.98

DOI: 10.54254/2755-2721/2026.34274

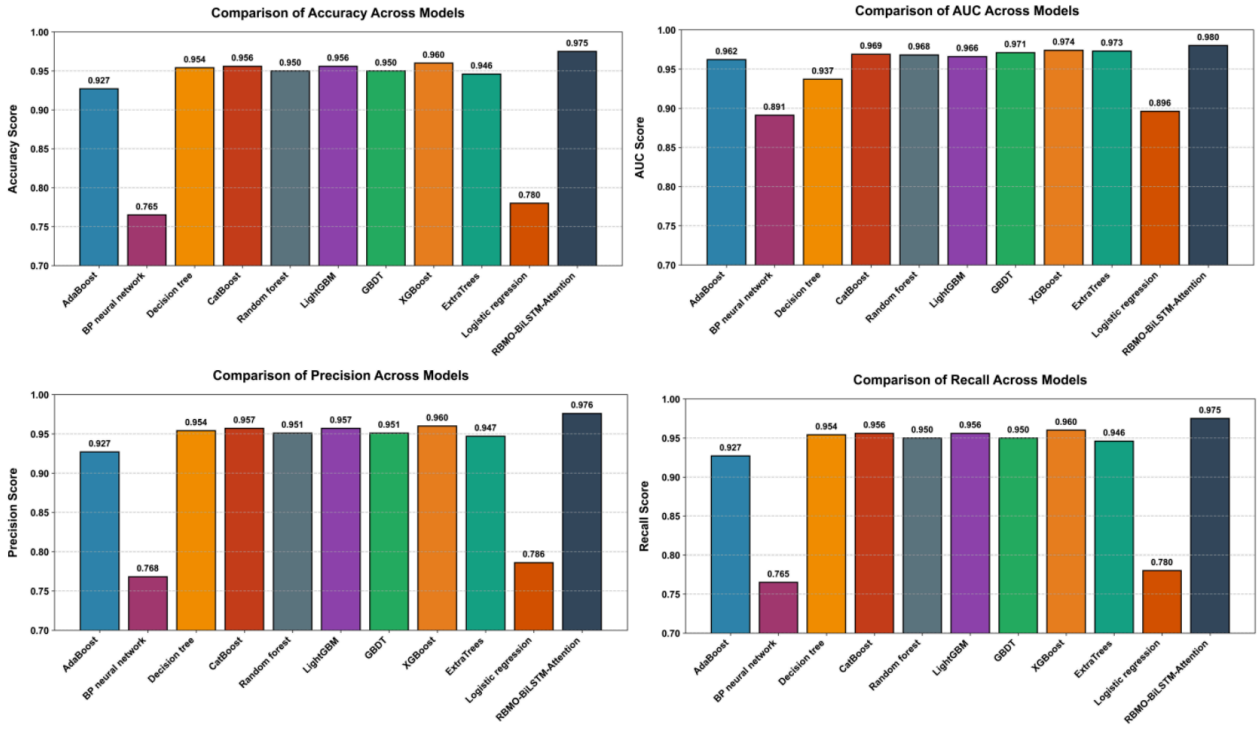


Figure 5. The comparison results of each index of each RBMO-BiLSTM-Attention test set

From the perspective of various evaluation indicators, the RBMO-BiLSTM-Attention algorithm proposed in this paper is significantly better than all comparative algorithms in terms of accuracy, recall, accuracy, F1 value and AUC value. Its accuracy rate reaches 98%, recall, accuracy, F1 value are both 98%, and AUC value also reaches 98%. The overall performance is the most outstanding.

5. Conclusion

Aiming at the problems of insufficient feature utilization and low classification accuracy in large language model illusion prediction, this paper designs a classification model integrating RBMO optimization, BiLSTM and attention mechanism, which effectively solves the defects of insufficient sequence feature extraction and inefficient parameter optimization in traditional methods. Experimental data verify that the evaluation indexes of the proposed algorithm are significantly better than AdaBoost, XGBoost, BP neural network and other comparison methods, and show stronger feature mining and classification discrimination ability. The model realizes the organic combination of feature learning, parameter optimization and classification decision-making, can accurately predict the probability of illusion occurrence, and make up for the shortcomings of existing pre-prediction technology. The research results provide a feasible solution for risk avoidance in the reasoning process of large models, expand the research path of safe and trusted deployment of large models, and can be popularized and applied in professional scenarios with high reliability requirements such as medical treatment and law. In the future, the scale of data sets and scenario types can be further expanded, and the cross-domain adaptability of models can be improved.

DOI: 10.54254/2755-2721/2026.34274

References

[1] Ashok, Dhananjay, and Jonathan May. "Language models can predict their own behavior." arXiv preprint arXiv: 2502.13329 (2025).
[2] Downey, Anthony. "Algorithmic predictions and pre-emptive violence: artificial intelligence and the future of unmanned aerial systems." Digital War 5.1 (2024): 123-133.
[3] Perez-Ferreiro, Pablo Miguel, et al. "Generating trustworthy explanations with a language model enriched by fuzzy rule-based systems." 2025 IEEE International Conference on Fuzzy Systems (FUZZ). IEEE, 2025.
[4] Li, Haoyang, et al. "A survey on large language model acceleration based on kv cache management." arXiv preprint arXiv: 2412.19442 (2024).
[5] Banerjee, Debodeep, et al. "Learning to guide human decision makers with vision-language models." arXiv preprint arXiv: 2403.16501 (2024).
[6] Xiao, Dannier. Machine learning-based hazard monitoring for autonomous driving: investigating multi-modality and generative language models. Diss. University of Warwick, 2024.
[7] Clarke, Charles LA, et al. "Report on the 2nd Search Futures Workshop at ECIR 2025." ACM SIGIR Forum. Vol. 59. No. 1. New York, NY, USA: ACM, 2025.
[8] Xu, Zhiyuan, et al. "Steering in the Shadows: Causal Amplification for Activation Space Attacks in Large Language Models." arXiv preprint arXiv: 2511.17194 (2025).
[9] Park, Changmin, et al. "Understanding physical properties of unseen deformable objects by leveraging large-language models and robot actions." International Journal of Advanced Robotic Systems 23.2 (2026): 17298806261430024.
[10] Xuesong, Li, et al. "The Legal Governance Logic of the Artificial Intelligence Open-Source Ecosystem: Evolutionary Path from Post-Event Accountability to Pre-Event Prevention from a Technological Perspective." Available at SSRN 5897262 (2025).