Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 14
TPCEE 2026

# Research on Theories and Methods of Interpretability in Large Language Models

Yang Xiang

School of Software Technology, DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, 116600, China

**Abstract.** In recent years, large language models have made significant progress in various aspects of natural language processing due to their powerful language understanding and generation capabilities. However, their internal decision-making processes are highly complex and typically "black-box" in nature, which limits their widespread application in fields such as healthcare, finance, and law. This paper primarily discusses the interpretability issue of large language models. First, we define the concept and scope of interpretability from three perspectives: model transparency, comprehensibility, and trustworthiness, while summarizing the challenges faced in interpretability research—such as massive parameter scale, strong emergent abilities, and complex attention mechanisms. Second, based on different approaches, existing interpretability methods are categorized into three types: mechanism-based methods (e.g., attention visualization, neuron activation analysis, circuit discovery), probe-based methods (e.g., linear probes, causal probes), and generation-based explanation methods (e.g., chain-of-thought prompting, self-explanation generation). We introduce the principles, advantages, and limitations of each category. Additionally, we discuss the relationship between interpretability and model performance, safety, and fairness, arguing that enhancing interpretability can effectively detect model biases, prevent hallucinatory outputs, and increase user confidence. Finally, addressing current research issues such as inconsistent evaluation criteria, difficulty in ensuring explanation fidelity, and limited cross-modal interpretability studies, we propose future research directions—including building causal reasoning frameworks, establishing unified standards, and exploring human-machine collaborative explanation—to advance the field of large language model interpretability.

**Keywords:** Large language models; Interpretability; Attention mechanism; Probing methods; Trustworthy artificial intelligence.

## 1. Introduction

With the advancement of deep learning technology, large language models have become one of the most disruptive innovations in the field of artificial intelligence. Since the release of the BERT model in 2018, large language models such as the GPT series, LLaMA, and Claude have made significant progress in both parameter scale and performance, reaching billions or even trillions of parameters by 2023. According to the Stanford University Artificial Intelligence Index Report, global training costs for large language models increased more than 300-fold between 2019 and 2024, while their performance in natural language understanding, text generation, and code writing improved by over 42%. However, as these models continue to grow in size, their internal mechanisms are becoming increasingly opaque—resembling a "black box"—which raises serious concerns in high-risk applications such as medical diagnosis, financial risk control, and judicial decision-making. In 2022, the European Union's Artificial Intelligence Act mandated that high-risk AI systems must be explainable, and the U.S. National Institute of Standards and Technology has also identified explainability as a key component of trustworthy AI. Therefore, research on the explainability of large language models is not merely a technical issue but also involves accountability, ethics, and legal considerations. It requires a comprehensive theoretical foundation and methodology to help people understand how these models operate internally, enhance transparency in their decision-making processes, and ensure they are used responsibly.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

## 2. Theoretical Foundations of Large Language Model Explainability

### 2.1. Definition and Classification of Explainability

The interpretability of large language models refers to the extent to which humans can understand and trace the reasons behind a model's decisions, encompassing three aspects: model transparency, prediction explainability, and result credibility. In terms of transparency, interpretability requires that the model's architecture, training data, and hyperparameters be accessible and inspectable. Regarding explainability, it means making the model's outputs comprehensible—clarifying how the model arrived at its conclusions. As for credibility, it ensures that explanations align [1] with the model's actual decision-making process. Academically, interpretability methods are generally categorized into two types: intrinsic interpretability, where the model itself is transparent (e.g., decision trees), and post-hoc interpretability, which uses external techniques to explain black-box models. For large language models, interpretability research also needs to consider local and global explanations: local explanation clarifies the reasoning behind individual predictions, while global explanation interprets the overall behavior of the entire model. These different levels of classification help enable comprehensive evaluation and optimization of interpretability in large language models.

### 2.2. The Black-Box Nature of Large Language Models

The black-box nature of large language models stems from their highly complex deep neural network architectures, particularly evident in the Transformer structure. Current mainstream models such as GPT-4 contain over 1.7 trillion parameters, with hundreds of layers of self-attention and feed-forward networks, through which each word undergoes thousands of nonlinear transformations. The sheer scale of parameters renders traditional methods of analyzing weights ineffective; knowledge within these models is stored in a distributed manner, making it difficult for any single parameter or neuron to correspond directly to a specific semantic concept. Moreover, large language models exhibit emergent capabilities—advanced functions such as reasoning and code generation suddenly appear once the model reaches a certain size, though the exact mechanisms behind these abilities remain unclear. Additionally, multi-head attention enables different attention heads to learn distinct linguistic patterns, which interconnect to form a complex network. As a result, analyzing only one aspect cannot fully capture the entire decision-making process, presenting a major challenge in the interpretability research of large language models.

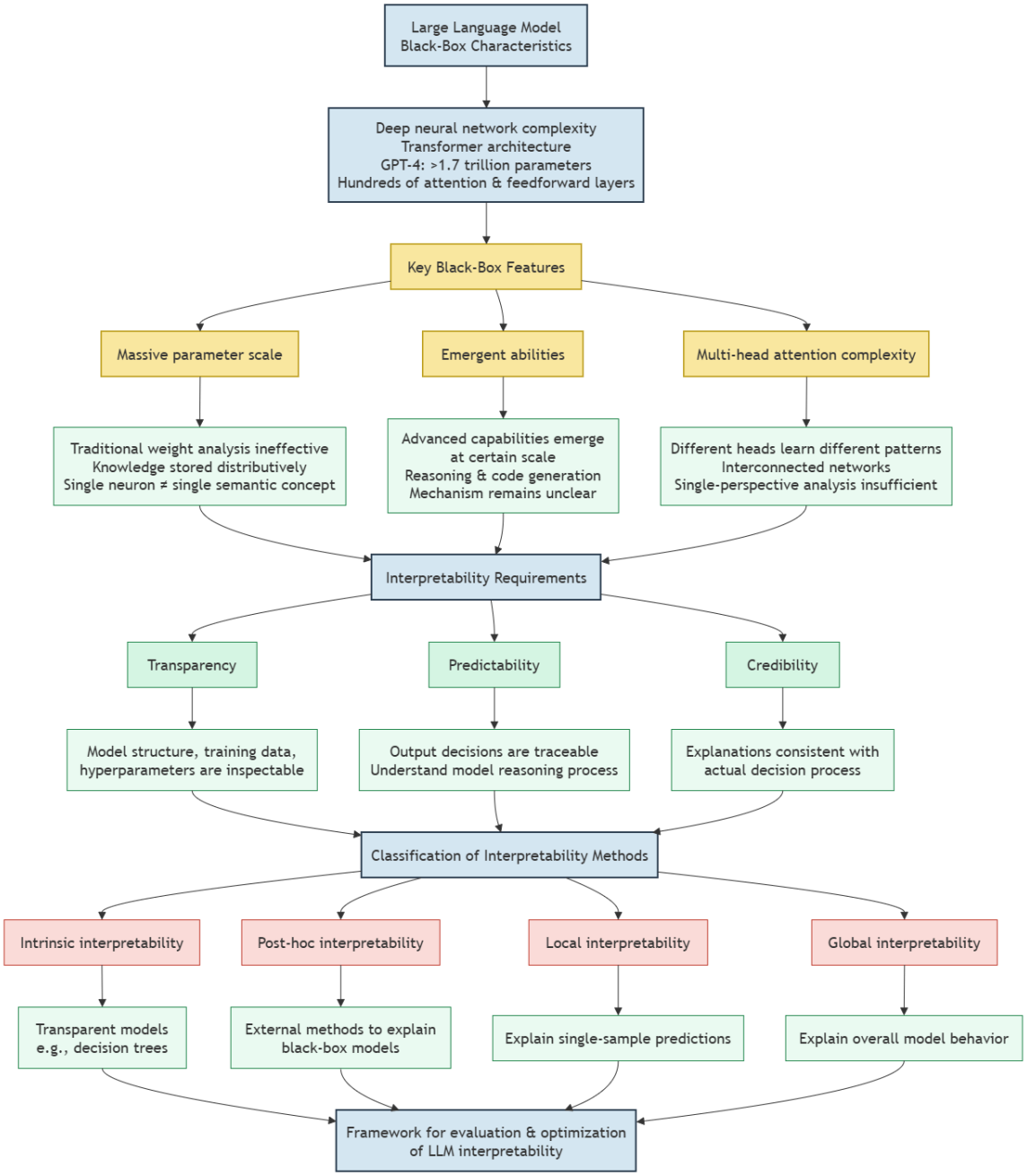


**Figure 1.** Mechanism Diagram of the Black-Box Characteristics of Large Language Models

### 2.3. The Necessity and Challenges of Explainability Research

The importance of research on the interpretability of large language models lies primarily in three aspects. First, there is a need for risk prevention: during generation, large language models may produce factual errors, harmful biases, or privacy leaks, and interpretability helps identify the root causes [2] of such issues. Second, regulatory compliance requires it: under Article 22 of the EU's GDPR, users have the "right to explanation," meaning they have the right to understand the reasons behind automated decisions. Third, there is significant scientific value: studying how large language models work can help optimize their architecture and training methods. However, this field faces considerable challenges. The explosive growth in model parameters leads to a dramatic increase in computational cost—just one forward pass of GPT-3 involves processing 175 billion parameters. The models' multi-task learning capabilities result in highly entangled internal knowledge representations, making it difficult to isolate neurons corresponding to specific functions. Capabilities such as context-based learning emerge nonlinearly, and no comprehensive theoretical framework currently exists to describe these phenomena. Moreover, there is no universally accepted metric to evaluate the accuracy

of different interpretation methods; varying approaches may yield inconsistent results, necessitating more systematic and multi-layered research frameworks from researchers.

### 2.4. Explainability Evaluation Metric System

Establishing reasonable and interpretable evaluation metrics is a crucial prerequisite for advancing this field. Current research primarily categorizes evaluation criteria into four aspects: faithfulness, completeness, simplicity, and usefulness. Faithfulness refers to the extent to which an explanation reflects the model's actual decision-making process, typically assessed through perturbation tests—by modifying input features deemed important by the explanation and observing whether the model's output changes as expected. Completeness requires that explanations include all factors influencing the outcome, without omitting any significant elements. Simplicity emphasizes conveying information to users with minimal details, avoiding excessive technic [3] ality. Usefulness aims to ensure that explanations help users understand model behavior and make informed decisions. For large-scale language models, evaluations must also consider cross-sample consistency and cross-task generalization—for instance, whether the model provides consistent explanations for similar questions or if the explanation method can be applied across different downstream tasks. In recent years, researchers have proposed automated evaluation tools such as the ERASER benchmark and LIME-stability, but a unified standard has yet to emerge.

## 3. Mainstream Approaches to Explainability in Large Language Models

### 3.1. Explanation Methods Based on Attention Mechanisms

Attention-based explanation methods represent the most intuitive technical approach in research on the interpretability of large language models. The core idea is to explicitly represent attention weights within Transformers as a basis for model decision-making. By visualizing the attention score matrices among tokens in multi-head self-attention layers, one can understand what content the model focuses on when generating a particular output. Studies have shown that different attention heads serve distinct functions: some capture syntactic dependencies such as subject-verb relationships, while others address semantic anaphora resolution. However, this method has significant limitations. Attention weights merely reflect how information is integrated, not necessarily the degree of causal influence; high attention weights do not always correspond to important components. In deep networks, the cumulative effects of attention patterns across layers are difficult to trace, and layer-by-layer analysis may overlook cross-layer information flow. Moreover, the redundancy inherent in multi-head attention leads to ambiguous interpretations, as different heads may learn similar representations. Recent work has attempted to improve explanation quality by using attention flow techniques [4]—tracking information flow across layers—and adjusting attention weights with gradient information. These enhancements have improved effectiveness to some extent, but computational costs remain prohibitively high for ultra-large-scale models.

### 3.2. Feature Attribution and Gradient Analysis Methods

Feature attribution methods are a way to analyze model interpretability by assessing the degree of influence that input features have on the final output, based on the concept of Shapley values from game theory. For large language models, each input token can be regarded as a feature, and its importance is measured by comparing the change in model output when that token is removed. In contrast, gradient-based methods use backpropagation to compute partial derivatives of the output with respect to the input as significance scores—examples include integrated gradients or layer-wise relevance propagation. These methods benefit from rigorous mathematical foundations and can provide fine-grained attribution at the pixel or token level. However, they also face challenges: in cases of gradient saturation—where activation functions lie in flat regions—the gradients approach zero, making attribution impossible. Additionally, for discrete text inputs, embeddings must first convert tokens into continuous vectors before gradients can be computed, introducing potential errors.

Moreover, computing exact Shapley values has an exponential time complexity relative to input length, making precise calculation impractical for long texts containing thousands of tokens. To address these issues, researchers have proposed techniques such as Monte Carlo sampling approximations and hierarchical attribution to reduce computational overhead, as well as adversarial attribution methods that use adversarial examples to validate the reliability of attribution results. These improvements have enabled feature attribution methods to achieve strong performance in tasks such as sentiment analysis and text classification.

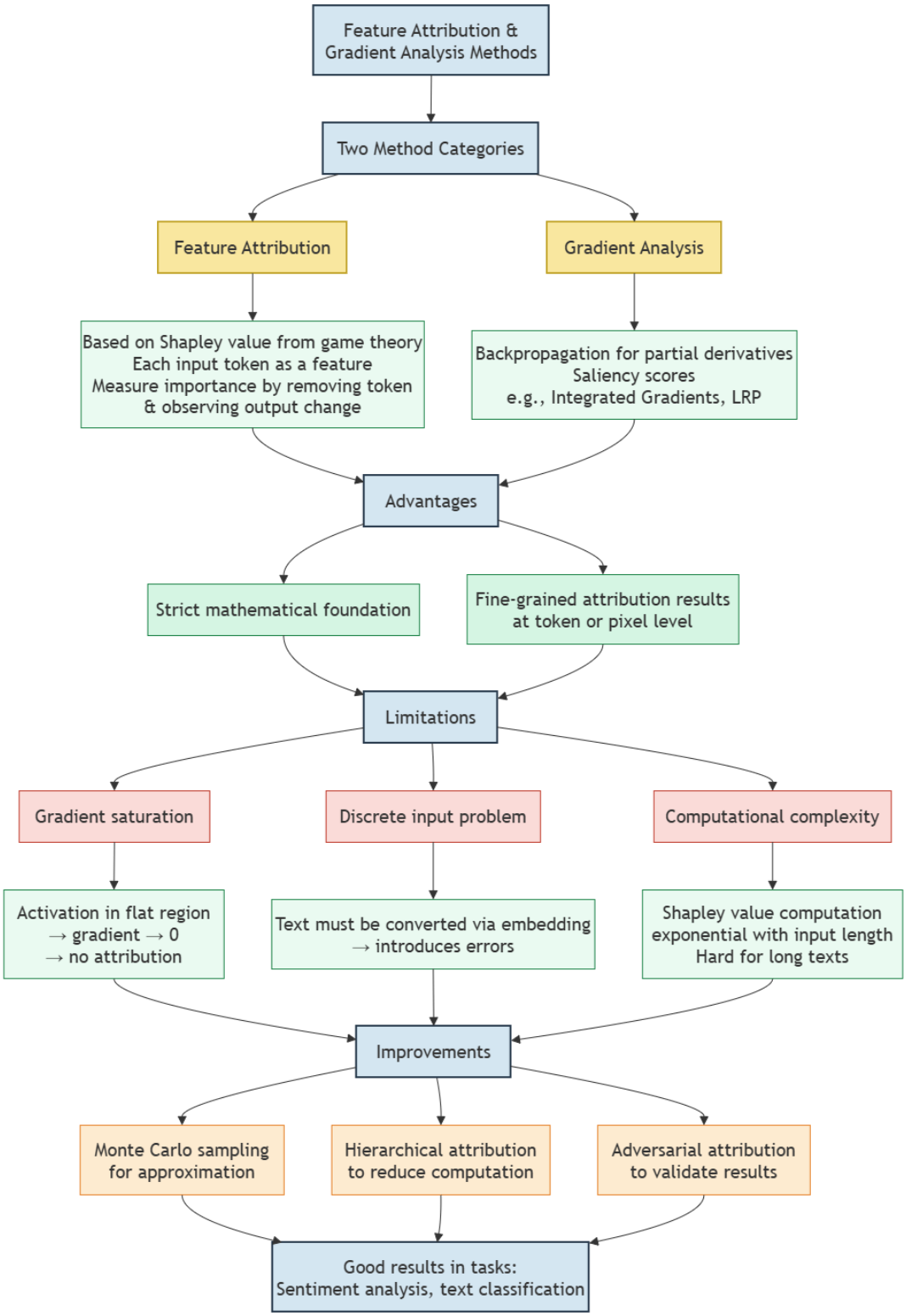


**Figure 2.** Schematic of the mechanism for feature attribution and gradient analysis methods

### 3.3. Detection Classifier Method

The method of probing classifiers involves using an auxiliary classifier to detect linguistic knowledge embedded in the hidden layer representations of large language models, thereby gaining insight into the model's internal representations. Specifically, the parameters of a pre-trained language model are fixed, and hidden state vectors from intermediate layers are extracted as input features. A simple linear classifier or multi-layer perceptron is then trained on these features to perform a linguistic task such as part-of-speech tagging [5], syntactic parsing, or semantic role labeling. If the probing classifier performs well, it indicates that the linguistic information has been learned by the model and stored within its representation space. Experiments have shown that intermediate layers of BERT models effectively capture syntactic tree structures, while deeper layers in GPT series models tend to represent semantics and world knowledge more prominently. The advantage of this approach is that it allows research without modifying the original model and enables large-scale exploration of where various linguistic phenomena are encoded. However, there are limitations: linear probes may underestimate the richness of representations, as non-linear patterns cannot be detected. Moreover, the learning capacity of the classifier might mask flaws in the representations, leading to overly optimistic results. More importantly, probing methods only confirm the presence of information but do not reveal how the model actually uses this information to make decisions; therefore, complementary approaches such as causal interventions are needed for a comprehensive understanding.

### 3.4. Causal Intervention and Counterfactual Explanation

Causal intervention methods establish causal relationships among variables by manipulating internal states or inputs within a model, offering a more rigorous scientific approach to studying the interpretability of large language models. Drawing on Pearl's framework for causal inference, these methods treat the model as a set of nodes in a causal graph and apply do-calculus interventions to observe their effects. In practice, two main approaches are used: one is neuron ablation experiments, which involve deliberately disabling certain neurons or attention heads to measure performance degradation and thereby identify critical computational units; the other is activation patching, where [6] activation values at a specific layer are replaced during inference to analyze output changes and locate stored knowledge. Counterfactual explanation methods generate minimally modified input samples that alter model outputs, revealing decision boundaries by comparing original and counterfactual instances. The advantage of these methods lies in establishing clear causal chains rather than merely inferring correlations, providing valuable guidance for model debugging and knowledge editing. Recent studies have successfully located factual knowledge in GPT models using causal tracing techniques, identifying the middle-to-late layers' MLP modules as playing a crucial role. However, challenges remain regarding intervention effectiveness—ensuring that manipulations affect only the target variable without introducing confounding effects—and the assumption of additive causality may not hold for highly complex deep models, necessitating improved intervention designs and methods for effect decomposition.

### 3.5. Concept Activation Vector Method

The concept activation vector (CAV) method provides large language models with an interpretable framework based on human-understandable concepts, overcoming limitations inherent in traditional neuron-level analyses. The core idea involves defining a directional vector in the model's representation space that corresponds to a higher-order semantic concept, then using this vector to measure the extent to which the concept is activated under a given input. Specifically, researchers first select a set of examples representing a particular concept—such as "formal tone" or "medical terminology"—and compute the average of their activation vectors at a specific hidden layer to obtain the CAV for that concept. They then use directional derivatives to assess how this concept influences the model's output. A key advantage of this approach lies in transforming distributed representations within neural networks into perceptible, conceptual terms, enabling even individuals without

technical expertise to understand how models make decisions. However, the method also faces challenges: the quality and representativeness of concept samples directly affect results, while the subjective nature of concepts can introduce bias. In high-dimensional representation spaces, different concepts may exhibit collinearity, leading to contradictory interpretations. Automatically identifying meaningful concepts and constructing a hierarchical conceptual ontology remains a significant challenge.

Interpretability of large language models remains a significant barrier to achieving trustworthy and responsible AI. This paper provides a comprehensive review and summary of interpretability in large language models from both theoretical and methodological perspectives, highlighting current research hotspots and future trends. Theoretically, various definitions of interpretability serve as criteria for evaluating and enhancing model transparency. However, due to their massive parameter sizes and emergent capabilities, large language models exhibit strong black-box characteristics, posing substantial challenges to traditional explanation methods. Methodologically, diverse approaches—from attention mechanisms and causal interventions to feature attribution and concept activation vectors—each have their strengths and limitations, complementing one another to form a multi-layered framework for interpretation [7]. Nevertheless, existing research still faces challenges such as inconsistent evaluation standards, difficulty in ensuring the faithfulness of explanations, and high computational costs, indicating a significant gap toward truly transparent and reliable AI systems.

Looking ahead, research on the interpretability of large language models must advance simultaneously in theoretical innovation and engineering practice. On one hand, efforts should be intensified to study the role of causal reasoning frameworks within neural networks, establishing a set of universal interpretability evaluation criteria and public datasets to facilitate collaboration among institutions in developing industry standards. On the other hand, human-machine collaborative explanation methods should be developed, integrating expert knowledge with automated analysis to create user-friendly interpretation tools. Particularly given the rapid advancement of multimodal large models, research on cross-modal information integration and interpretability has become an urgent priority. Only by continuously deepening our understanding of the internal mechanisms of these models can we fully harness their powerful capabilities while mitigating potential negative impacts, enabling artificial intelligence to truly benefit human society and building a safer, more reliable, and intelligent world.

## 4. Application and Limitations Analysis of Explainable Methods

### 4.1. Applications of Explainable Methods in Text Generation Tasks

In text generation tasks, interpretability methods are crucial for understanding how large language models generate text. Attention visualization techniques reveal the degree to which the model focuses on different positions in the input sequence when generating each token, enabling researchers to observe how information flows and identify key contextual elements the model relies on. Neuron activation analysis further uncovers the linguistic functions of specific neurons, such as their roles in semantic concepts or syntactic structures—for instance, certain neurons in story generation tasks maintain character consistency or narrative coherence. Chain-of-thought prompting encourages models to output intermediate reasoning steps, clarifying the logical connections between question understanding and answer generation, making it especially useful [8] in multi-step reasoning tasks like summarization and dialogue systems. Circuit discovery techniques analyze how models leverage distinct subnetworks to perform complex tasks from a modular perspective, helping identify primary computational pathways responsible for creative expression, fact retrieval, and style control. These interpretability methods not only enhance transparency in the text generation process but also provide practical approaches for improving model architecture, reducing hallucinations, and enhancing generation quality.

### 4.2. Applications of Explainable Methods in Sentiment Analysis

Sentiment analysis is a crucial issue in natural language processing and plays a significant role in evaluating the effectiveness of explainability methods. Linear probes employ simple classifiers on hidden layer representations to identify positions and dimensions that encode sentiment polarity, and studies show that large language models often exhibit relatively stable representations in intermediate layers. Attention weight analysis reveals how much the model focuses on sentiment words, negation words, and degree adverbs when judging the sentiment of a text, thereby enhancing understanding of fine-grained sentiment recognition processes. Causal probes use counterfactual interventions to demonstrate that certain neurons causally contribute to sentiment judgments, distinguishing between correlation and causation. Self-explanation generation enables models to provide natural language explanations such as "This sentence contains negative words like 'disappointed' and 'terrible,' so it is classified as negative," significantly improving the readability of predictions and boosting user trust in the system, which is beneficial for commercial applications.

### 4.3. Analysis of Limitations in Existing Methods

Although interpretability research has made significant progress, current methods still face many fundamental issues. First is the problem of fidelity in explanations—attention weights do not necessarily reflect the true reasons behind a model's predictions. Some studies suggest that attention maps may be misleading, and the actual reasoning process of models is far more complex than it appears on the surface. Second, there is no unified evaluation metric across different interpretability methods, making it difficult to compare and select among them; the same explanation may appear superior or inferior depending on the evaluation criteria used. Third [9], as model sizes grow into the trillions of parameters, traditional interpretability approaches become prohibitively expensive—methods such as probe training and neuron analysis require substantial time and computational resources, which are impractical for large-scale models. Fourth, emergent capabilities in large language models make many higher-order cognitive functions resistant to local explanations. For instance, models can perform reasoning and analogical tasks without being explicitly trained on them, yet the underlying mechanisms remain largely unknown. Finally, most existing research focuses on single-modal text tasks, while the effectiveness of current interpretability methods in handling cross-modal interactions and integration of information across multiple modalities in multimodal large models remains underexplored, limiting their practical applicability.

### 4.4. Trade-off between Explainability and Model Performance

Explainability and model performance are interdependent yet mutually constraining, particularly evident in the design and application of large language models. On one hand, enhancing explainability typically requires imposing certain constraints or simplifications on the model—such as enforcing sparse attention or reducing the number of layers—which may diminish its ability to learn complex linguistic patterns and lead to suboptimal performance on challenging tasks. On the other hand, achieving top performance often involves designing deeper, more complex architectures with extensive parameter interactions, which in turn increases model opacity. However, recent research suggests this trade-off is not absolute: by studying explainability to eliminate unnecessary parameters or identify harmful shortcut learning, it is possible to improve or at least maintain model performance while increasing transparency. In high-stakes applications such as financial risk control and medical diagnosis, slightly sacrificing accuracy to gain greater explainability is essential, as auditability and accountability of decisions hold values that cannot be measured in monetary terms.

## 5. Conclusions

This article provides a comprehensive overview of the background, research methodologies, and related applications of interpretability in large language models, serving as a summary of the field. Interpretability of large language models is not merely a technical issue but also crucial to the reliability of artificial intelligence systems and their acceptance [10] by society. A comparative

analysis of various approaches—mechanism-based methods, probe-based methods, and generation-based explanation methods—reveals that each has its own strengths and limitations. Traditional techniques such as attention visualization and neuron activation analysis can help improve understanding of models, yet their accuracy requires further investigation. Probe techniques are highly effective in identifying specific functional modules, but establishing causal relationships remains challenging. Generation-based methods like chain-of-thought prompting facilitate better communication between users and models, although the validity of generated explanations against real-world scenarios still needs verification. While interpretability research has achieved certain progress in areas such as text generation and sentiment analysis, challenges remain, including the lack of unified evaluation standards, high computational resource consumption, and difficulties in cross-modal transfer.

Looking ahead, research on the interpretability of large language models needs to move toward greater diversity and deeper insights. Developing an interpretability framework that incorporates causal reasoning can effectively distinguish between superficial correlations and true causal relationships. Establishing unified standards covering aspects such as faithfulness, completeness, and comprehensibility is essential foundational work for promoting healthy development in this field. Designing a human-AI collaborative approach to generating explanations—leveraging both automated model analysis and human expertise—can lead to more useful explanation systems. Furthermore, with the advancement of multimodal large models, cross-modal interpretability has emerged as a new [11] frontier. To transition large language models from being merely "usable" to truly "trustworthy," interpretability must remain a top priority. Only then can these models be genuinely applied across high-risk domains, making artificial intelligence more transparent, controllable, and responsible.

## References


[1] Yu, H., Wang, Y., Cheng, T., Yan, Y., Dawson, K. A., Li, S. F. Y., Zheng, Y., & Jin, Y. (2025). Empowering scientific discovery with explainable small domain-specific and large language models. Artificial Intelligence Review, 58(12).

[2] Ranaldi, L. (2025). Survey on the Role of Mechanistic Interpretability in Generative AI. Big Data and Cognitive Computing, 9(8).

[3] Salvi, M., Seoni, S., Campagner, A., Gertych, A., Acharya, U. R., Molinari, F., & Cabitza, F. (2025). Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. International Journal of Medical Informatics, 197.

[4] Martín, A. P., Criado, F. J. S., Mateo, C. G., Fernández, L. D., & Pérez, M. d. C. L. (2025). Context Is King: Large Language Models' Interpretability in Divergent Knowledge Scenarios. Applied Sciences, 15(3).

[5] Wang, J., Xu, Z., Yu, J., Bian, L., & Li, M. (2026). A Review on Explainability of Large Language Models in Recommender Systems. Computer Systems Application, 35(2), 1–22. (in Chinese)

[6] Zhu, Z., Wu, H., Xu, J., & Wu, Y. (2025). Explainability of Large Model Thinking Patterns: A Mind-Set Model Based on the Integration of Gestalt Theory and Speech-Action Theory. Open Education Research, 31(6), 29–40. (in Chinese)

[7] Cheng, Z., & Duan, Y. (2025). Research on an Explainable Interactive Fake Review Detection Model Integrating DeepSeek Reasoning Clue Features. Information Theory and Practice, 48(11), 20–27. (in Chinese)

[8] Li, J., & Shao, J. (2025). Research on Predicting Sentiment Characteristics of Financial News Based on Explainable Large Language Models. Intelligent Computer and Applications. (in Chinese)

[9] Zuo, D., Jiang, W., Zheng, H., & Bao, J. (2025). An Explainable Visual Question Answering Method for Defect Recognition. Computer Integrated Manufacturing Systems, 31(6), 2084–2097. (in Chinese)

[10] Lin, M., Dai, C., & Guo, T. (2024). A Method for Generating Explanations of Offensive Memes Based on Multimodal Large Language Models. Journal of Computer Research and Development, 61(5), 1206–1217. (in Chinese)

[11] Ju, T., Liu, G., Zhang, Z., & Zhang, R. (2024). A Review of Probe-based Explainable Methods in Natural Language Processing. Chinese Journal of Computers, 47(4), 733–758. (in Chinese)