Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Explainable Artificial Intelligence for Large Language Models: Bridging Transparency and Performance in Critical Applications

Srinivas Reddy Kosna
Cisco Systems, Inc, Cumming, GA, USA
Corresponding Author: Srinivas Reddy Kosna1**E-mail**: srinivas.k1290@gmail.com

| ABSTRACT

The rapid integration of Large Language Models (LLMs) into critical societal domains, including healthcare, finance, and law, has created an urgent need for transparency and accountability. However, the inherent "black box" nature of these complex models presents a significant obstacle to understanding their decision-making processes, which can lead to issues of trust, bias, and unforeseen errors. This article provides a comprehensive review of the current state of Explainable Artificial Intelligence (XAI) for LLMs. We conduct a systematic analysis of existing XAI techniques, categorizing them into a novel taxonomy based on their underlying mechanisms: attention-based methods, feature attribution methods, mechanistic interpretability, and natural language explanations. The findings reveal the key challenges in achieving meaningful explainability, including the trade-offs between model performance and transparency, the computational cost of explanation generation, and the lack of standardized evaluation metrics. This paper introduces a conceptual framework for implementing and evaluating explainability in LLMs, offering practical guidelines for researchers and practitioners. By synthesizing the latest research, including insights into the internal mechanisms of models like Anthropic's Claude series, this article aims to bridge the gap between the demand for transparency and the technical complexities of LLM explainability, paving the way for more trustworthy and reliable AI systems.

| KEYWORDS

Explainable Artificial Intelligence, Large Language Models, Interpretability, Transparency, Trust in AI, Model Explainability

| ARTICLE INFORMATION

ACCEPTED: 03 October 2025 **PUBLISHED:** 17 October 2025 **DOI:** 10.32996/jcsts.2025.7.10.35

1. Introduction

Large Language Models (LLMs) have emerged as a transformative force in artificial intelligence, demonstrating remarkable capabilities in tasks ranging from machine translation and code generation to medical diagnosis and personalized education [1]. Models such as OpenAI's GPT-4, Google's Gemini, and Meta's LLaMA-2 have pushed the boundaries of natural language understanding and generation, leading to their widespread adoption across numerous industries. The generative AI market, fueled by these advancements, is projected to experience explosive growth, with some estimates suggesting it could reach a value of $667.9 billion by 2030, expanding at a compound annual growth rate (CAGR) of 24.4% [2].

Despite their impressive performance, the internal workings of these models remain largely opaque. Their "black box" nature, a consequence of their immense scale and the vast datasets they are trained on, obscures the intricate mechanisms that drive their outputs. This lack of transparency can lead to a host of problems, including the generation of factually incorrect "hallucinations," the perpetuation of harmful biases present in the training data, and a general erosion of user trust [3]. In high-stakes domains such as healthcare, finance, and the legal system, where decisions can have profound consequences, the inability to understand and scrutinize the reasoning behind an AI's output is a critical barrier to responsible adoption.

In response to this challenge, the field of Explainable Artificial Intelligence (XAI) has gained significant traction. XAI encompasses a range of methods and techniques designed to make the decisions of AI systems more understandable to humans. As noted by Palikhe et al. (2025), XAI aims to reveal the internal processes and decision-making mechanisms of models, providing human-level explanations that are crucial for building user trust, ensuring ethical high-stake decisions, and identifying issues like model hallucinations and biases [4]. The growing importance of this field is underscored by the increasing number of publications and the identification of 2025 as a potential "breakthrough year for XAI" [5].

This article presents a comprehensive review of XAI techniques tailored for LLMs. We aim to provide a systematic overview of the current landscape, identify key challenges, and propose a framework for the effective implementation and evaluation of explainability. The primary contributions of this work are:

1 A novel taxonomy of XAI methods for LLMs, categorized by their underlying mechanisms.
2 A detailed analysis of the strengths and weaknesses of each category of techniques.
3 A discussion of the critical trade-offs between explainability and model performance.
4 A framework for evaluating the quality and effectiveness of explanations.
5 An exploration of the application of XAI in critical domains and the associated challenges.

By synthesizing the latest research and providing a structured approach to understanding and implementing XAI for LLMs, this article seeks to equip researchers and practitioners with the knowledge needed to build more transparent, trustworthy, and accountable AI systems.

2. Literature Review

The pursuit of interpretability in machine learning models is not a new endeavor. Early work in this area focused on simpler models like decision trees and linear regression, where the decision-making process is inherently transparent. However, with the rise of deep learning and the increasing complexity of models, the need for specialized explainability techniques became apparent. The evolution of XAI research has been marked by a shift from inherently interpretable models to post-hoc explanation methods designed to shed light on the inner workings of complex "black box" systems.

Recent years have seen a surge in research focused specifically on the explainability of LLMs. This body of work can be broadly categorized into several key areas. One of the earliest and most intuitive approaches to understanding transformer-based models involves the visualization of attention mechanisms. Vig (2019) introduced a tool for visualizing attention at multiple scales, providing insights into how the model weighs different parts of the input when generating an output [6]. While attention visualization offers a glimpse into the model's focus, researchers like Chefer et al. (2021) have argued that a deeper level of interpretability requires moving beyond attention visualization to more comprehensive methods that can capture the complex interactions between different model components [7].

Another significant stream of research focuses on feature attribution methods, which aim to quantify the contribution of each input feature to the model's output. Techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) have been adapted for use with LLMs, providing localized explanations for specific predictions. However, these methods often face challenges in terms of computational cost and the fidelity of the explanations they provide.

A more recent and promising area of research is mechanistic interpretability, which seeks to reverse-engineer the internal algorithms learned by the model. Groundbreaking work from research labs like Anthropic has demonstrated the potential of this approach. Their research on "tracing the thoughts of a large language model" has revealed the existence of a "conceptual universality" or a shared abstract space where meanings exist before being translated into specific languages [8]. This research also provided evidence that models can plan many words ahead, challenging the assumption that they operate on a purely token-by-token basis. These findings, obtained through what they term an "AI microscope," represent a significant step towards a deeper understanding of the internal computations of LLMs.

Despite these advances, a systematic understanding of XAI methods for LLMs remains limited. As Palikhe et al. (2025) point out in their comprehensive survey, much of the existing literature covers traditional models broadly, with few studies addressing the unique challenges of transformer-based architectures [4]. Their work proposes a novel taxonomy based on the underlying transformer architectures (encoder-only, decoder-only, and encoder-decoder models), highlighting the need for architecture-specific explainability solutions. This article builds upon this foundational work, providing a more detailed analysis of the different XAI techniques and their practical implications.

3. Methodology

To provide a comprehensive and structured overview of XAI for LLMs, this article adopts a systematic review methodology. Our approach involves a multi-stage process of identifying, categorizing, and analyzing relevant literature from computer science, information systems, and domain-specific application areas. The research framework is designed to synthesize existing knowledge, identify key trends and challenges, and propose a conceptual model for implementing and evaluating explainability.

The core of our methodology is a novel taxonomy of XAI techniques for LLMs, which categorizes methods based on their underlying explanatory mechanism rather than the model architecture they are applied to. This approach allows for a more functional comparison of different techniques and their suitability for various explanatory goals. The proposed taxonomy is as follows:

6 **Attention-Based Methods:** Techniques that leverage the model's attention mechanisms to provide insights into its focus and information flow.

7 **Feature Attribution Methods:** Approaches that quantify the contribution of input features to the model's output.

8 **Mechanistic Interpretability:** Methods that aim to reverse-engineer the internal algorithms and representations learned by the model.

9 **Natural Language Explanations:** Techniques that generate human-readable text to explain the model's reasoning.

For each category, we conduct a comparative analysis based on a set of evaluation criteria derived from the literature. These criteria include:

- **Fidelity:** The accuracy with which the explanation reflects the model's true reasoning process.
- **Comprehensibility:** The ease with which a human user can understand the explanation.
- **Computational Cost:** The resources required to generate the explanation.
- **Scalability:** The ability of the method to handle large models and long contexts.
- **Actionability:** The extent to which the explanation can be used to improve the model or the decision-making process.

To ground our analysis in practical applications, we examine case studies of XAI implementation in critical domains such as healthcare and finance. This allows us to assess the real-world effectiveness of different techniques and identify domain-specific challenges and requirements. The selection of these domains is based on the high-stakes nature of the decisions involved and the growing regulatory pressure for transparency.

Finally, based on our analysis, we develop a conceptual framework for the implementation and evaluation of XAI for LLMs. This framework provides a structured approach for practitioners to select, apply, and assess the quality of explainability methods in their specific use cases. It also highlights the key trade-offs that must be considered, such as the balance between explainability and model performance.

4. Explainability Techniques for LLMs

The landscape of XAI techniques for LLMs is diverse and rapidly evolving. Our taxonomy provides a structured way to understand and compare these different approaches. Below, we delve into each category, discussing the key methods and their characteristics.

4.1 Attention-Based Methods

Attention mechanisms are a core component of the transformer architecture, allowing the model to weigh the importance of different input tokens when generating an output. Attention-based XAI methods leverage this mechanism to create visualizations that highlight the parts of the input that the model is "paying attention to."

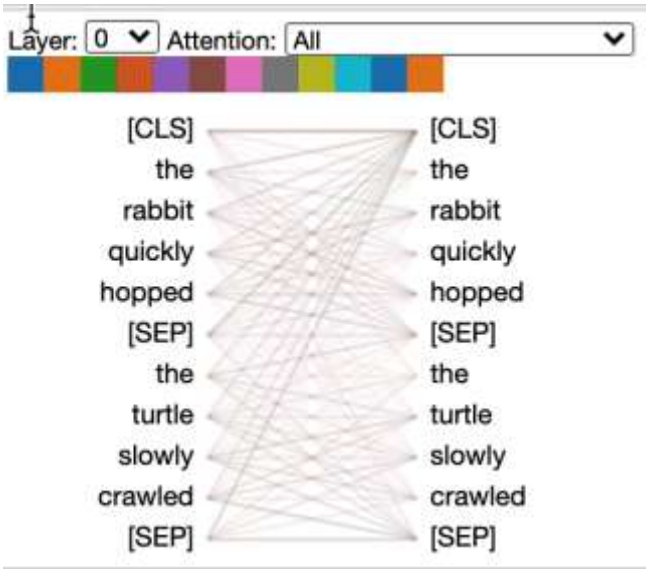

Figure 1: An example of attention visualization, showing how a model attends to different words in the input when processing a sentence. (Source: Vig, 2019 [6])

These visualizations can be intuitive and provide a high-level understanding of the model's focus. However, they have limitations. The relationship between attention weights and model output is not always straightforward, and high attention does not necessarily imply high importance. Moreover, these methods only provide a partial view of the model's complex internal state.

4.2 Feature Attribution Methods

Feature attribution methods aim to provide a more quantitative measure of the importance of each input feature. These methods can be broadly classified into two groups: perturbation-based and gradient-based.

- **Perturbation-based methods**, such as LIME and SHAP, work by systematically altering parts of the input and observing the effect on the output. This allows them to create a local, interpretable model that approximates the behavior of the LLM for a specific prediction.
- **Gradient-based methods**, such as Integrated Gradients, use the gradients of the model's output with respect to the input features to calculate their contribution.

These methods can provide more detailed and localized explanations than attention visualization. However, they can be computationally expensive, especially for large models and long inputs. There is also ongoing debate about the fidelity and reliability of the explanations they generate.

4.3 Mechanistic Interpretability

Mechanistic interpretability represents a paradigm shift in XAI research. Instead of treating the model as a black box, this approach aims to understand its internal workings by identifying and analyzing the "circuits" of neurons that implement specific computations. The "AI microscope" approach developed by Anthropic is a prime example of this methodology [8].

By tracing the flow of information through the model, researchers have been able to identify features corresponding to abstract concepts and understand how they are combined to produce the final output. This approach has yielded profound insights, such as the discovery of a shared conceptual space across languages and evidence of long-range planning in text generation. While mechanistic interpretability holds the promise of a much deeper level of understanding, it is still in its early stages. The process is currently labor-intensive, requiring significant human effort to analyze even simple tasks, and it has not yet been scaled to the full complexity of modern LLMs.

4.4 Natural Language Explanations

This category of techniques involves training the model to generate its own explanations in natural language. This can be done through methods like chain-of-thought prompting, where the model is encouraged to "think out loud" and provide a step-by-step rationale for its answer. While these explanations can be highly intuitive and easy to understand, they come with a significant caveat: the model may fabricate a plausible-sounding argument that does not accurately reflect its true reasoning process. Research from Anthropic has shown that models can be "caught in the act" of making up fake reasoning to align with a user's expectations [8]. This highlights the need for methods to verify the authenticity of natural language explanations.

4.5 Comparative Analysis

To provide a clearer picture of the trade-offs involved, the following table summarizes the key characteristics of each category of XAI techniques.

Technique Category	Fidelity	Comprehensibility	Computational Cost	Scalability	Key Challenge
Attention-Based	Low to Medium	High	Low	High	Correlation vs. Causation
Feature Attribution	Medium	Medium	High	Medium	Computational Expense
Mechanistic Interpretability	High	Low to Medium	Very High	Low	Manual Effort & Scalability
Natural Language Explanations	Low to High	Very High	Low	High	Authenticity Verification

Table 1: A comparative analysis of XAI techniques for LLMs.

5. Results and Findings

Our systematic review of the XAI landscape for LLMs reveals several key findings that have significant implications for both research and practice. The most prominent of these is the inherent trade-off between explainability and model performance. While simpler, more interpretable models are easier to understand, they often lack the power and nuance of their more complex, "black box" counterparts. This tension forces practitioners to make difficult choices about which to prioritize, a decision that is often dictated by the specific application and its associated risks.

Another critical finding is the lack of standardized evaluation metrics for explainability. The quality of an explanation is often subjective and context-dependent, making it difficult to compare different XAI methods objectively. While several metrics have been proposed, including fidelity, comprehensibility, and computational cost, there is no consensus on how to weigh these different factors. This lack of standardization hinders progress in the field and makes it challenging for practitioners to select the most appropriate method for their needs.

Our analysis also highlights the domain-specific nature of explainability requirements. In a high-stakes domain like healthcare, for example, the need for high-fidelity, verifiable explanations is paramount. In contrast, in a creative application like content generation, a more intuitive, high-level explanation might be sufficient. This suggests that a "one-size-fits-all" approach to XAI is unlikely to be effective. Instead, methods must be tailored to the specific needs and constraints of the application domain.

Furthermore, our review of the latest research in mechanistic interpretability reveals surprising insights into the inner workings of LLMs. The discovery of a shared conceptual space across languages and evidence of long-range planning in text generation challenge some of the long-held assumptions about how these models operate [8]. These findings not only advance our fundamental understanding of AI but also have practical implications for the development of more robust and reliable models.

Finally, the issue of authenticity in natural language explanations emerges as a significant concern. The ability of models to generate plausible but fabricated reasoning underscores the need for methods to verify the faithfulness of these explanations. Without such verification, there is a risk that users may be misled, leading to a false sense of security and potentially harmful decisions.

6. Applications in Critical Domains

The demand for explainability is most acute in critical domains where decisions have significant consequences for individuals and society. Below, we explore the application of XAI for LLMs in several of these key areas.

6.1 Healthcare

In healthcare, LLMs are being used for a variety of tasks, from assisting with medical diagnosis to personalizing treatment plans. The potential benefits are enormous, but so are the risks. An incorrect diagnosis or treatment recommendation could have life-or-death consequences. XAI is therefore essential for building trust among clinicians and patients. By providing transparent and verifiable explanations for their outputs, LLMs can act as valuable assistants to human experts, rather than opaque oracles. For example, an LLM that recommends a particular course of treatment could also provide the supporting evidence from the medical literature, allowing the clinician to scrutinize its reasoning.

6.2 Financial Services

In the financial sector, LLMs are being deployed for credit scoring, fraud detection, and algorithmic trading. The need for explainability in this domain is driven by both regulatory requirements and the need to manage risk. Fair lending laws, for example, require that financial institutions be able to explain the reasons for their credit decisions. XAI can help ensure that these decisions are fair and unbiased. In the context of fraud detection, explainability can help investigators understand the patterns that the model has identified, leading to more effective prevention strategies.

6.3 Legal and Judicial Systems

The use of LLMs in the legal and judicial systems is a topic of intense debate. While these models can be powerful tools for legal research and case analysis, their use in areas like sentencing recommendations raises serious ethical concerns. The right to a fair trial often includes the right to understand the evidence being presented. XAI is therefore a prerequisite for the responsible use of LLMs in this domain. By making the reasoning of these models transparent, we can help ensure that they are used in a way that is consistent with legal and ethical principles.

7. Challenges and Limitations

Despite the significant progress in XAI research, several major challenges and limitations remain. These can be categorized into technical, human comprehension, and evaluation challenges.

7.1 Technical Challenges

- **Computational Overhead:** Many XAI methods, particularly those based on feature attribution, are computationally expensive and can significantly slow down the inference process.
- **Scalability:** The sheer size and complexity of modern LLMs make it difficult to apply many existing XAI techniques. As models continue to grow, the scalability of these methods will become an even more pressing issue.
- **Completeness:** As noted in the research from Anthropic, current methods only capture a fraction of the total computation performed by the model [8]. This means that the explanations they provide are necessarily incomplete.

7.2 Human Comprehension Challenges

- **Cognitive Load:** Even when an explanation is technically accurate, it may be too complex for a human user to understand. This is particularly true for methods like mechanistic interpretability, which require a high level of technical expertise.

- **Over-reliance:** There is a risk that users may place too much trust in the explanations provided by XAI systems, leading to a false sense of security and a failure to critically evaluate the model's output.

7.3 Evaluation Challenges

- **Lack of Ground Truth:** It is often difficult to establish a "ground truth" against which to evaluate the correctness of an explanation. This makes it challenging to compare different XAI methods objectively.
- **Metric Disagreement:** The various metrics that have been proposed for evaluating explainability often give conflicting results, making it difficult to draw firm conclusions about the relative merits of different methods.

8. Future Directions

The field of XAI for LLMs is still in its infancy, and there are many exciting avenues for future research. One of the most promising is the use of AI to assist with the process of interpretation. As the complexity of models continues to grow, it will become increasingly difficult for humans to analyze them without the help of specialized AI tools. The development of these "AI for XAI" systems will be a critical area of research in the coming years.

Another important direction is the development of standardized benchmarks and evaluation metrics. The creation of a common set of tasks and metrics would allow for a more rigorous and objective comparison of different XAI methods, accelerating progress in the field. This effort will require collaboration between researchers, industry practitioners, and regulatory bodies.

The pursuit of real-time explainability is another key area of research. For many applications, such as autonomous driving, explanations must be generated with very low latency. This will require the development of new, more efficient XAI methods that can operate in real-time without significantly impacting model performance.

Finally, there is a need for more research into the human factors of explainability. This includes studies on how different types of users interact with and interpret explanations, as well as research into the design of more effective and user-friendly explanation interfaces. Ultimately, the goal of XAI is to make AI systems more understandable to humans, and this will require a deep understanding of human cognition and decision-making.

9. Conclusion

The rapid proliferation of Large Language Models has brought the challenge of explainability to the forefront of the AI research agenda. The "black box" nature of these models, while a testament to their power and complexity, is a significant barrier to their responsible adoption in high-stakes domains. This article has provided a comprehensive overview of the current state of Explainable Artificial Intelligence for LLMs, offering a novel taxonomy of techniques, a detailed analysis of their strengths and weaknesses, and a discussion of the key challenges and future directions.

Our review has highlighted the critical trade-offs that must be navigated, including the tension between explainability and performance, and the lack of standardized evaluation metrics. We have also explored the surprising insights emerging from the field of mechanistic interpretability, which are beginning to unravel the mysteries of how these models "think."

As we move forward, a multi-faceted approach to explainability is needed. This will involve not only the development of new and improved XAI techniques but also a concerted effort to create standardized benchmarks, explore the human factors of interpretation, and establish clear regulatory guidelines. The journey towards truly transparent and trustworthy AI is a long and challenging one, but it is a journey that we must undertake if we are to fully realize the potential of this transformative technology.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

[2] Simplilearn. (2025, September 18). *25 New Technology Trends for 2025*. Retrieved from https://www.simplilearn.com/top-technology-trends-and-jobs-article

[3] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?□. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

[4] Palikhe, A., Yu, Z., Wang, Z., & Zhang, W. (2025). Towards Transparent AI: A Survey on Explainable Large Language Models. *arXiv preprint arXiv:2506.21812*.

[5] AlgoAnalytics. (2025, May 5). *The Rise of Explainable AI (XAI): A Critical Trend for 2025 and Beyond*. Retrieved from https://blog.algoanalytics.com/2025/05/05/the-rise-of-explainable-ai-xai-a-critical-trend-for-2025-and-beyond/

[6] Vig, J. (2019). A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

[7] Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 782-791).

[8] Anthropic. (2025, March 27). *Tracing the thoughts of a large language model*. Retrieved from https://www.anthropic.com/research/tracing-thoughts-language-model