InSight Bulletin: A Multidisciplinary Interlink International Research Journal
Peer Reviewed International, Open Access Journal.
ISSN : 3065-7857 / Website: https://ibrj.us / Volume-3, Issue-1 / January - 2026

Original Article

Explainable AI (XAI) Framework for Large Language Models: A Transparent Reasoning Approach

Neeta Bonde
Department of Computer Science, Dr. D.Y. Patil Science and Computer Science College, Pune

Manuscript ID:
IBMIIRJ -2026-030107

Submitted: 06 Dec. 2025

Revised: 10 Dec. 2025

Accepted: 05 Jan. 2026

Published: 31 Jan. 2026

ISSN: 3065-7857

Volume-3

Issue-1

Pp. 32-36

January 2026

Correspondence Address:
Neeta Bonde
Department of Computer Science. Dr. D.Y. Patil Science and Computer Science College, Pune
Email: neetadpu@gmail.com

OPEN ACCESS

Quick Response Code:

Web. https://ibrj.us

DOI: 10.5281/zenodo.18918277

DOI Link:
https://doi.org/10.5281/zenodo.18918277

Abstract

Large Language Models (LLMs) are really important for lots of things we do with language on computers like creating text, helping us make decisions and figuring out answers. Even though they are good at what they do we do not really know how they come up with their answers. This is a problem because it is hard to trust something when we do not understand how it works. Large Language Models can be a problem when it comes to safety. They can also be biased. This is especially worrisome in areas, like education, healthcare and law where Large Language Models are used to make decisions. Large Language Models need to be more transparent so we can trust them and know they are being fair. The current methods we have for making Artificial Intelligence understandable, which is called Explainable AI do not work well with transformer-based models. This is because they cannot show us the steps the models take to reason things out or how they use information from a time ago or what they really know. This paper is about an idea for an Explainable AI (XAI) system that can help us see how Large Language Models think. The system has four parts: it can show us the path the model takes to make a decision, it can tell us where the model got its information from, it can analyse how each piece of information contributed to the decision and it can check if the decision is confident and safe. The new Explainable AI (XAI) system is designed to make Large Language Models more transparent. Together, these layers provide interpretable insights into model behaviour, improve reliability, and support ethical and regulatory compliance. The proposed framework aims to bridge the gap between high-performance language models and the growing need for trustworthy and explainable artificial intelligence.

Keywords: *Explainable AI, Large Language Models, Transformers, NLP, Reasoning Transparency, Interpretability, XAI, Model Accountability.*

Introduction

Large Language Models (LLMs) such as GPT-4/5, Llama 3, PaLM 2, and Claude have exhibited incredible language understanding and generation capabilities. Examples of their applications include summarization, translation, question answering, dialogue systems, and content generation. Although these models have achieved tremendous success in language understanding, they can be considered black boxes because they do not possess methods capable of interpreting their reasoning processes. The black box nature of these models hinders their effectiveness in critical domains, which demand high levels of transparency and trust (Kosna, 2025)(Herrera, 2025). Explainable AI (XAI) aims to render these AI models comprehensible by inferring interpretable explanations from them. However, the challenge of applying XAI solutions to large transformer models is a major research gap in this area. Current solutions such as attention analysis or post-hoc analysis of attribute explanations can achieve a better insight into model behaviour but not into the path of obtained knowledge in a multi-step model (Thogesan et al., 2025), (Fantozzi & Naldi, 2024).

Problem Statement

At present, a broad Explainable AI framework does not exist which can systematically understand the inferences in large language models on different tasks. State-of-the-art models used today in explainability do not analyse multiple-step inferences, contexts, or hidden inferential patterns in generative models. A good XAI framework will thus play a critical role in making inferences in large language models transparent and trustworthy. (Kosna, 2025; Herrera, 2025)

How to cite this article:
Bonde, N. (2026). Explainable AI (XAI) Framework for Large Language Models: A Transparent Reasoning Approach. InSight Bulletin: A Multidisciplinary Interlink International Research Journal, 3(1), 32–36. https://doi.org/10.5281/zenodo.18918277

Literature Review

Some research work has focused on the interpretability of machine learning models and/or NLP models, including: Transformers & Attention: The transformer model was proposed by Vaswani et al. in (Vaswani et al., 2017). The model used self-attention to enable parallel processing of contexts. Later studies proved that attention weights alone aren't a sufficient explanation for model predictions (Fantozzi & Naldi, 2024). Attention and Explain ability: Wiegreffe et al. have studied the shortcomings of attention in explanation and emphasized a need for additional tools in this area (Wiegreffe & Pinter, 2019). (Kosna, 2025; Herrera, 2025)

Xai Methods in Nlp

Research works such as LIME and Shapley Values provide model agnostic results but have a high computation cost in large models and lack a deep chain of reasoning (Ribeiro et al., 2020; Miller et al., 2024). LLM Interpretability: Some recent surveys have focused on the requirements for better interpretability methods targeted at LLMs, such as path-of-reasoning recovery and knowledge source identification. They highlight a major deficiency in state-of-the-art methods used for the explanation of LLMs. (Kosna, 2025; Herrera, 2025)

Research Gap

- While there has been significant progress in the areas of XAI research for traditional ML models, following are some of the remaining gaps:
- Lack of standardized frameworks for explaining LLM reasoning sequences.
- Lack of approaches to expose hidden knowledge representations along with multi‑ step logics.
- No integrated approach to combine token‑ level attributions with global reasoning explanations.
- Lack of confidence and safety metrics quantifying model uncertainty and the risk of hallucination in explanations.

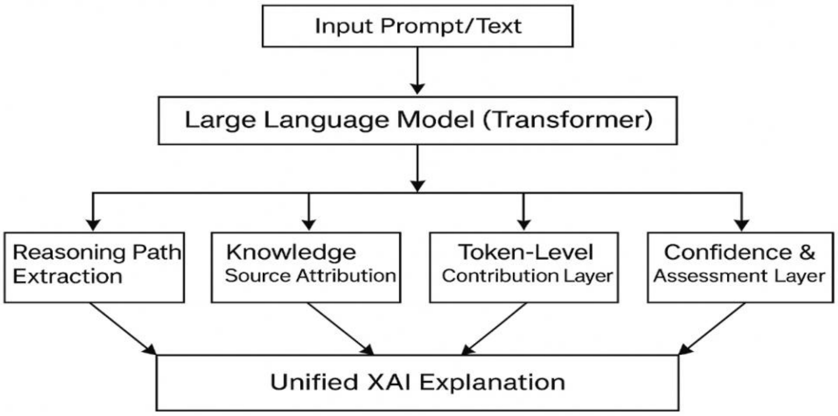
Objectives

The paper's objectives are:

- To suggest an XAI framework pertaining to transparent reasoning in a transformer-based LLM.
- To discover paths of multi-step reasoning and knowledge contribution.
- To facilitate contribution analysis at the token level for generated responses.
- To incorporate assessment of confidence and safety in LLM explanations.
- To enable trusted deployment of LLMs in critical contexts.

Proposed Xai Framework

The proposed Explainable AI (XAI) framework is designed to make the internal reasoning of Large Language Models (LLMs) transparent, interpretable, and verifiable. Unlike traditional post-hoc explanation methods that only highlight token importance; this framework introduces four complementary layers of interpretability. Together, they provide a multi-dimensional view of how an LLM processes information, retrieves knowledge, constructs reasoning, and generates final responses.


Input Prompt/Text
Large Language Model (Transformer)
Reasoning Path Extraction
Knowledge Source Attribution
Token-Level Contribution Layer
Confidence & Assessment Layer
Unified XAI Explanation

A. Reasoning Path Extraction Layer

"This level aims at reconstructing the steps of hidden reasoning performed by LLM.

LLMs have an inherent multi-step logic in them, which involves decomposing Task X into a series of smaller tasks based on given inputs, constructing an interim representation, searching for relevant information in memory, and filling in a structured output. Due to this being an invisible step, this level analyses "internal activations and hidden states to infer:" (Kosna, 2025; Herrera, 2025)

Logical sub-steps

- Hidden transitions among layers
- Internal chain-of-thought patterns
- Representational shifts in generation

This will allow a better understanding of how a given model arrived at a given part of an answer.

B. Knowledge Source Attribution Layer

It identifies, in this layer, what knowledge the model relied on.

LLMs derive information from: Training knowledge-in-world knowledge learnt before deployment. Contextual knowledge: input prompt, attached documents Implicit associations: patterns that emerged during pre-training The attribution module scores the influence of each knowledge source. This helps determine: Whether the model relied on factual information If the answer was drawn from hallucinated or irrelevant sources Which sentences/paragraphs in the input have shaped the output? This would be crucial for high-risk applications such as education, healthcare, and analysing case laws. (Kosna, 2025; Herrera, 2025)

C. Token-Level Contribution Layer

This layer quantifies the contribution of each token towards the final output. It uses scalable attribution techniques (such as gradient-based scoring, integrated gradients, attention rollouts) optimized for LLMs. It provides:

- Word-by-word contribution scores
- Emphasis on influential parts
- Comparative importance across layers

Decision-critical token visualization by It thus provides granular interpretability for text classification, summarization, reasoning, or question-answering tasks.

D. Confidence and Safety Layer

This is the part that checks if the model is giving us results. The validation layer looks at the model's output to see if it is really reliable. The validation layer is very important because it helps us figure out if we can trust the model's output. (Kosna, 2025; Herrera, 2025). It does the math. Figures things out:

- Uncertainty Score (probabilistic confidence)
- Hallucination Likelihood (risk of fabricated content)
- Bias Metrics (social, cultural, or contextual bias)
- Safety Violations (toxic content, harmful suggestions)

This layer ensures that the explanations also include safety insights, making the system trustworthy for real-world deployment.

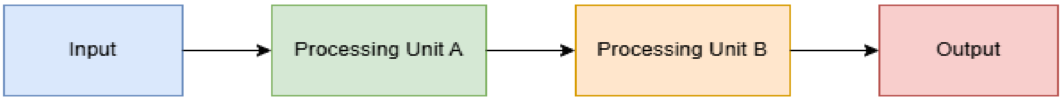
E. Unified XAI Explanation Output

All four layers are used to make an explanation report, which has the following things:

- Reasoning steps
- Knowledge sources
- Token-level influence
- Reliability scores

The unified output now serves as a clear and structured explanation that is understandable by humans, suitable for end-users, domain experts, or auditors.

Architecture Description



The following presents the architecture textually:

- Input Prompt or Text is fed into a transformer based LLM.
- Multi-layer explain ability modules extract the reasoning, knowledge source contributions, token importance, and safety metrics.
- The different layers are integrated by the XAI explanation output into a structured explanation to be presented to the end users.

That architecture supports the interpretation of diverse tasks that include factual generation, creative text, decision support applications, among others.

Methodology

The methodology comprises these steps:

- Input acquisition: A natural language input sentence or paragraph is given.
- LLM Execution: The responses and hidden representations are produced using a transformer model.
- Identification of paths of reasoning: Linking responses to layers by using logical paths.
- Knowledge Attribution: Calculate influence scores based on training data or context.
- Token Contribution Analysis: Ascertaining the role of tokens in output decisions.
- Confidence Scoring
- Uncertainty and Biases in Model Predictions
- Explanation Generation: Generate a multi-layer explanation report to end-users.

Expected Outcomes

- The proposed framework will:

- Improve transparency of LLM outputs.
- The detection of hallucinations and inconsistencies in reasoning.
- Providing interpretable contributions both at the token and reasoning levels.
- Develop systems that can be deployed more safely in health, law, and education.
- Support ethical auditing and AI accountability.

Applications

The proposed XAI framework for LLMs enables transparent and trustworthy AI behaviour in a wide variety of real-world applications. It facilitates interpretable reasoning, knowledge attribution, and confidence estimation and therefore can be responsibly deployed in everything from low-risk to high-stakes environments. (Kosna, 2025; Herrera, 2025)

A. Explainable Chatbots for Customer Service

Today, customer service chatbots serve a variety of industries: banking, e-commerce, telecommunications, and public services. But sometimes, bad or unclear responses may mislead the user or diminish satisfaction. The XAI framework proposed here provided the capability of justification in chatbot responses by showing reasoning paths along with token-level importance. This would help the organization audit decisions made through chatbots. This increases users' confidence and reduces total erroneous automated responses while permitting human agents to always double-check and even discard certain decisions if necessary. (Kosna, 2025; Herrera, 2025)

B. Expert Systems: Medical and Legal Applications

AI systems in healthcare and legal domains provide support to professionals by analysing patient records, medical literature, legal documents, and case laws. Mistakes or hallucinations in these domains bear severe consequences. The proposed framework furthers the decision-support systems by explaining how conclusions are generated, identifying the knowledge sources used, and providing confidence scores. This can meet ethical standards, and help doctors and legal experts validate the recommendations of AI in real-world application decision-making. (Kosna, 2025; Herrera, 2025)

C. Transparency Tools for Educational AI

AI systems are increasingly being deployed for automated tutoring, content generation, grading assistance, and personalized learning. The lack of explain ability can lead to biased content or incorrect explanations, or an unfair assessment. Integrating XAI into educational AI tools enables them to explain why specific answers or feedback were generated, trace knowledge coverage, and point out areas of uncertainty. Fairness will be advanced, learning outcomes will improve, and educators will be supported in their efforts to maintain academic integrity. (Kosna, 2025; Herrera, 2025)

D. Auditing Content to Ensure Safety and Fairness

Large Language Models currently generate millions of pages of content, ranging from summaries of news and social media posts to policy documents. The proposed XAI framework will support content auditing in the identification of biased language, detection of hallucinations, and the assessment of safety risks. Token-level explanations make auditors aware of which parts of the input influenced a potentially harmful output, thus motivating corrective measures that make ethical content generation possible. (Kosna, 2025; Herrera, 2025) E. Regulatory Compliance with Emerging AI Governance Frameworks Increasingly, in these times, the time for AI regulations such as the EU AI Act and Global AI Governance Policies has begun mandating transparency and explainability. The proposed XAI framework enables an organization to meet the regulatory standards by the interpretable explanation, audit trail, and the confidence of AI decisions. This will ensure accountability, traceability, and legal compliance and hence make the deployment of LLM viable in regulated industries. (Kosna, 2025; Herrera, 2025)

Challenges and Limitations

Key challenges include:

- A. Scalability to Billion-Parameter Models
- Detail in Explanation vs. Computational Complexity.
- Uncovering informative reasoning from hidden representations.
- Explanations not exposing the sensitive training data.
- Protecting Sensitive Training Data from Explanation Exposure.

Future Scope

Future work may include:

- Applying this framework to multimodal LLMs (text & image).
- Creating standards for judging quality in explanation.
- Including Explainability in Model Training Processes.
- Applying this framework to domain-specific generative tasks.

Conclusion

This paper presented a new framework for XAI, dedicated to explaining internal reasoning in transformer-based Large Language Models. By integrating multi-layer reasoning extraction, knowledge attribution, token-level contribution analysis, and confidence scoring in one framework, interpretable explanations can be assured, allowing better transparency, trust, and safety. This work contributes to trustworthy AI research and supports the responsible deployment of LLMs in mission-critical applications. (Kosna, 2025; Herrera, 2025)

Acknowledgment

The author would like to express sincere gratitude to the Department of Computer Science, Dr. D. Y. Patil Science and Computer Science College, Pune, for providing the academic environment and resources necessary for carrying out this research.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. (Kosna, 2025) S. R. Kosna, "Explainable Artificial Intelligence for Large Language Models: Bridging Transparency and Performance in Critical Applications," J. Comput. Sci. Technol. Stud., vol. 7, no. 10, pp. 326–333, 2025.
2. (Herrera, 2025) F. Herrera, "Making Sense of the Unsensible: Reflection, Survey, and Challenges for XAI in Large Language Models Toward Human‑ Centered AI," arXiv:2505.20305, 2025.
3. (Thogesan et al., 2025) T. Thogesan, A. Nugaliyadde, and K. W. Wong, "Integration of Explainable AI Techniques with Large Language Models for Enhanced Interpretability for Sentiment Analysis," arXiv:2503.11948, 2025.
4. (Vaswani et al., 2017) A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017, pp. 6000–6010.
5. (Wiegreffe & Pinter, 2019) S. Wiegreffe and Y. Pinter, "Attention Is Not Explanation," in Proc. ACL, 2019.
6. (Fantozzi & Naldi, 2024) P. Fantozzi and M. Naldi, "The Explainability of Transformers: Current Status and Directions," Computers, vol. 13, no. 4, 2024.
7. (Ribeiro et al., 2020) M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High‑ Precision Model‑ Agnostic Explanations," in Proc. AAAI, 2020.
8. (Miller et al., 2024) "Explainable AI for Text Classification: Lessons from a Comprehensive Evaluation of Post Hoc Methods," Cogn. Comput., Aug. 2024.
9. K. Zhou, W. Chen, and W. Sheng, "Explainable AI with fine‑ tuned large language models for sustainable cultural heritage management," Sci. Rep., vol. 15, art. no. 41370, 2025.
10. H. R. Lekshmi Ammal and A. K. Madasamy, "A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers," J. Big Data, vol. 12, art. no. 46, 2025.
11. B. Sebin, N. Taskin, and N. Mehdiyev, "Exploring the Intersection of Large Language Models (LLMs) and Explainable AI: A Systematic Literature Review," Proc. Inf. Res. Manag., 2024.
12. A. Zytek, S. Pido, and K. Veeramachaneni, "Explingo: Explaining AI Predictions using Large Language Models," arXiv:2412.05145, 2024.
13. W. Alrajhi, N. Alangari, S. Al‑ Ghamdi, and H. Al‑ Khalifa, "From Theory to Practice: A Hands‑ on Tutorial in Explainable NLP," COLING, 2025.
14. D. Venkatesh, A. Jaiswal, and G. Nanda, "Comparing human text classification performance and explainability with large language and machine learning models using eye‑ tracking," Sci. Rep., vol. 14, art. no. 14295, 2024.
15. "Explainable Artificial Intelligence (XAI): From Inherent Explainability to Large Language Models," arXiv:2501. 09967, 2025.
16. E. Cambria, L. Malandri, F. Mercorio, N. Nobani, and A. Seveso, "Survey of key challenges in LLM interpretability," arXiv, 2024.
17. "Towards Transparent AI: A Survey on Explainable Large Language Models," arXiv, Jun. 2025.
18. "Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era," arXiv:2403.08946, 2024.
19. N. Feldhus et al., "LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools," NAACL, 2024.
20. L.F.Villa‑Arenas et al., "Anchored Alignment for Self‑Explanations Enhancement," arXiv, 2024.