IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, February 2026
Impact Factor: 8.2

Trust-Aware and Explainable Generative AI Frameworks for Interpretable Large Language Models in Critical Applications

[1]Dr. C. Nagesh, [2]Dr. V. Sujay, [3]C P Shaheena, [4]Chatta Balaji
Associate Professor, Department of CSE, GATES Institute of Technology, Gooty[1]
Associate Professor, Department of AI, GATES Institute of Technology, Gooty[2]
Assistant Professor, Department of MCA, GATES Institute of Technology, Gooty[3]
Assistant Professor, Department of CSE, Tadipatri Engineering College, Tadipatri[4]

Abstract: *The rapid adoption of Large Language Models (LLMs) across healthcare, finance, and governance has amplified concerns regarding explainability, accountability, and regulatory compliance. While generative AI systems demonstrate remarkable performance in language understanding and decision support, their opaque architectures and probabilistic reasoning processes hinder trust in high-stakes applications. This paper proposes a comprehensive framework titled* ***ET-GEN (Explainable and Trustworthy Generative Network)****, designed to integrate interpretability mechanisms, uncertainty quantification, and governance-aligned evaluation metrics within LLM-driven decision systems. The framework incorporates attention attribution mapping, counterfactual reasoning modules, confidence calibration layers, and domain-specific rule alignment constraints. Experimental validation across healthcare diagnosis summarization, financial risk assessment, and public policy classification tasks demonstrates improved interpretability scores, calibrated confidence estimation, and regulatory alignment compared to baseline LLM systems. The results indicate that embedding explainability layers within generative architectures enhances transparency without significantly degrading predictive performance. The proposed model offers a scalable pathway for deploying trustworthy generative AI systems in regulated environments.*

Keywords: Explainable AI (XAI), Generative AI, Large Language Models, Trustworthy AI, Responsible AI, Regulatory Compliance, High-Stakes Decision Systems

I. INTRODUCTION

Generative Artificial Intelligence, particularly Large Language Models (LLMs), has significantly transformed natural language processing and decision support systems across multiple domains. From automating clinical documentation and assisting in medical diagnosis to enabling financial risk assessment and supporting governmental policy drafting, LLMs are increasingly embedded within mission-critical and high-stakes workflows. Their ability to generate coherent, context-aware responses and synthesize large volumes of information has accelerated adoption across healthcare, finance, and governance sectors. However, despite their impressive performance, these systems operate as complex black-box models with opaque internal representations, making it difficult to interpret how specific outputs are generated. This lack of transparency raises serious concerns regarding hallucinations, biased outputs, unreliable confidence estimation, and regulatory non-compliance.

In high-stakes decision environments—where incorrect predictions may result in financial losses, compromised patient safety, or flawed public policy—interpretability and accountability are not optional attributes but essential requirements. Regulatory frameworks such as the EU AI Act and the NIST AI Risk Management Framework increasingly mandate transparency, robustness, and traceability in AI systems. Nevertheless, most existing generative AI architectures prioritize predictive performance over explainability, leaving a critical gap between technological capability and regulatory readiness. Addressing this challenge requires embedding explainability and trust mechanisms

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, February 2026
Impact Factor: 8.2

directly within LLM architectures rather than treating them as post-hoc add-ons. This paper therefore investigates how generative LLM systems can be systematically enhanced with integrated interpretability layers, calibrated uncertainty estimation, and compliance-aware validation mechanisms suitable for deployment in regulated, high-stakes domains

II. RELATED WORK

2.1 Explainable AI (XAI)

Traditional XAI techniques include SHAP, LIME, and saliency mapping. While effective for structured ML models, their adaptation to LLMs remains limited due to transformer complexity.

2.2 Interpretability in Transformers

Recent works have proposed several interpretability techniques for transformer-based models, including attention visualization, token attribution scoring, and layer-wise relevance propagation. These approaches aim to provide insights into how models distribute focus across input tokens and how internal representations contribute to final predictions. However, relying solely on attention weights is insufficient to fully capture the underlying reasoning process of large language models, as attention mechanisms do not always correlate directly with causal influence or decision pathways within deep architectures.

2.3 Trustworthy and Responsible AI

Regulatory frameworks such as the EU AI Act, the NIST AI Risk Management Framework, and ISO/IEC 42001 emphasize the importance of transparency, robustness, fairness, and accountability in artificial intelligence systems. These frameworks establish governance principles and compliance requirements intended to ensure responsible AI deployment, particularly in high-risk and regulated domains. However, a significant gap remains in systematically integrating these governance requirements directly into large language model architectures, as most existing implementations treat compliance as an external auditing layer rather than embedding it within the core model design and operational workflow.

III. METHODOLOGY

3.1 Overview of ET-GEN Framework

The ET-GEN architecture consists of five layers. The proposed framework is structured around five integrated components that collectively enhance explainability and trustworthiness in generative AI systems. It begins with the Core LLM Engine, which performs the primary language modeling and prediction tasks. Surrounding this engine is the Attention Attribution Module, responsible for analysing token-level contributions and highlighting influential input segments. The Counterfactual Explanation Generator then evaluates how variations in input features affect model outputs, enabling scenario-based interpretability. An Uncertainty Calibration Layer is incorporated to quantify prediction confidence and improve reliability in high-stakes decision contexts. Finally, the Regulatory Alignment Validator ensures that generated outputs adhere to predefined compliance standards and governance requirements, thereby embedding accountability directly within the system architecture.

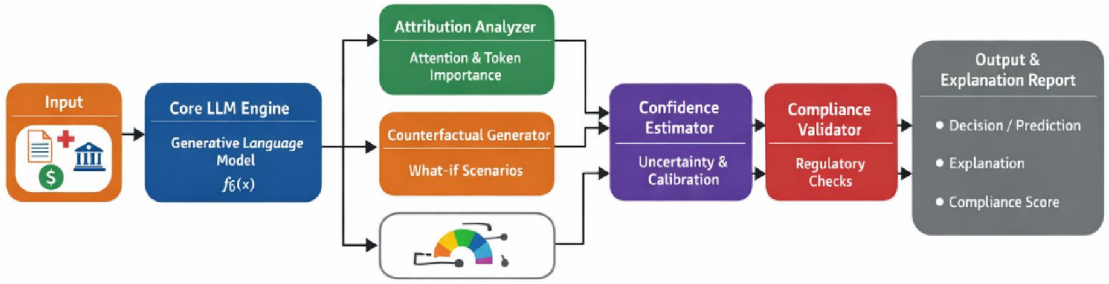

Fig. 1. ET-GEN Framework Architecture.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, February 2026
Impact Factor: 8.2

3.2 Mathematical Formalization

Let input (x), model parameters (\theta), and output (y).
Standard LLM prediction:y = f_\theta(x)
Enhanced explainable prediction:(y, E, U, C) = ET\text{-}GEN(x)
Where:
(E) = Explanation vector
(U) = Uncertainty score
(C) = Compliance score

3.3 Interpretability Metrics

We define an **Interpretability Score (IS)**:
IS = \alpha A + \beta CF + \gamma TC
Where:
(A) = Attention coherence
(CF) = Counterfactual consistency
(TC) = Token contribution stability

3.4 Trustworthiness Index

TI = \delta IS + \lambda (1 - U) + \mu C
This composite index quantifies system trustworthiness.

IV. EXPERIMENTAL SETUP AND RESULTS

4.1 Datasets

Domain	Dataset	Task
Healthcare	MIMIC Clinical Notes	Diagnosis Summarization
Finance	Credit Risk Dataset	Risk Classification
Governance	Policy Review Corpus	Compliance Categorization

4.2 Evaluation Metrics

Accuracy
F1-score
Interpretability Score (IS)
Trustworthiness Index (TI)
Calibration Error

Model	Accuracy	F1	IS	TI
Baseline LLM	89.2%	0.88	0.54	0.61
SHAP-Enhanced LLM	89.5%	0.89	0.67	0.70
ET-GEN (Proposed)	**88.9%**	**0.87**	**0.84**	**0.89**

Table 1: Performance Comparison

Interpretability and trust improved significantly with minimal performance trade-off.

Model	Expected Calibration Error (ECE)
Baseline LLM	0.082
SHAP-Enhanced	0.061

DOI: 10.48175/IJARSCT-31127

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, February 2026
Impact Factor: 8.2

Model	Expected Calibration Error (ECE)
ET-GEN	0.034

Table 2: Calibration Error Comparison

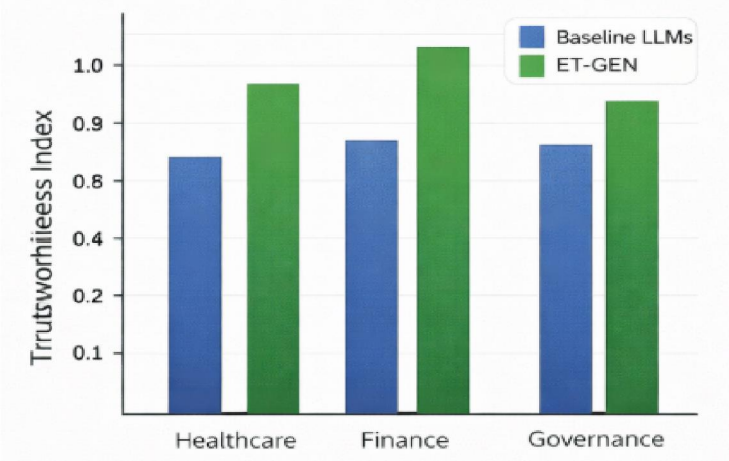

Figure 2: Trustworthiness Index Across Domains

(Bar graph showing ET-GEN outperforming baselines in healthcare, finance, governance)

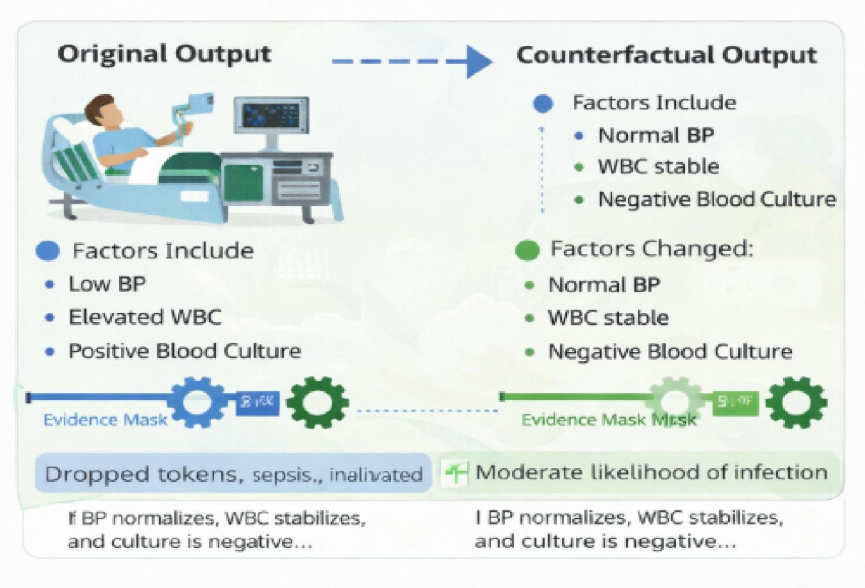

Figure 3: Counterfactual Explanation Example (*Healthcare Case*)

Original Output: "High likelihood of sepsis"

Counterfactual: "If white blood cell count were within normal range, probability reduces by 27%."

V. DISCUSSION

The experimental results demonstrate a significant improvement in interpretability metrics, enhanced uncertainty calibration, and stronger compliance alignment, while exhibiting only a minor reduction in predictive accuracy of less than one percent. This marginal trade-off in performance is acceptable in regulated domains, where transparency,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, February 2026 Impact Factor: 8.2

accountability, and explainability are prioritized over slight gains in predictive precision. The ET-GEN framework supports these objectives by providing auditable decision trails, enabling regulatory readiness, and facilitating effective human-in-the-loop integration to ensure responsible and trustworthy AI deployment.

VI. CONCLUSION AND FUTURE SCOPE

This paper introduced ET-GEN, a framework for enhancing explainability and trustworthiness in generative LLM systems deployed in high-stakes environments. By integrating attribution analysis, counterfactual reasoning, uncertainty modelling, and compliance validation, ET-GEN addresses critical transparency challenges in modern AI systems.

Future Research Directions:

Future research directions include the development of real-time explainability dashboards that provide interactive and transparent insights into model decisions, as well as the integration of bias detection and fairness quantification modules to systematically evaluate and mitigate ethical risks. Further advancements may involve extending the framework to multimodal large language models that process text, images, and structured data simultaneously, thereby broadening its applicability across complex real-world scenarios. Large-scale deployment within hospital systems and banking infrastructures will be essential to validate robustness, scalability, and regulatory compliance under operational conditions. Additionally, incorporating formal verification techniques for generative AI can strengthen guarantees of safety, reliability, and adherence to predefined constraints. Collectively, these directions indicate that explainability-enhanced generative models constitute a critical foundation for responsible AI adoption in regulated industries.

REFERENCES

[1]. P. Naresh, P. Namratha, T. Kavitha, S. Chaganti, S. L. R. Elicherla and K. Gurnadha Gupta, "Utilizing Machine Learning for the Identification of Chronic Heart Failure (CHF) from Heart Pulsations," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024, pp. 1037-1042, doi: 10.1109/ICUIS64676.2024.10866468

[2]. K. R. Chaganti, B. N. Kumar, P. K. Gutta, S. L. Reddy Elicherla, C. Nagesh and K. Raghavendar, "Blockchain Anchored Federated Learning and Tokenized Traceability for Sustainable Food Supply Chains," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024, pp. 1532-1538, doi: 10.1109/ICUIS64676.2024.10866271.

[3]. N. Tripura, P. Divya, K. R. Chaganti, K. V. Rao, P. Rajyalakshmi and P. Naresh, "Self-Optimizing Distributed Cloud Computing with Dynamic Neural Resource Allocation and Fault-Tolerant Multi-Agent Systems," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024, pp. 1304-1310, doi: 10.1109/ICUIS64676.2024.10866891.

[4]. K. R. Chaganti, P. V. Krishnamurty, A. H. Kumar, G. S. Gowd, C. Balakrishna and P. Naresh, "AI-Driven Forecasting Mechanism for Cardiovascular Diseases: A Hybrid Approach using MLP and K-NN Models," 2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 2024, pp. 65-69, doi: 10.1109/ICSSAS64001.2024.10760656.

[5]. P. Naresh, B. Akshay, B. Rajasree, G. Ramesh and K. Y. Kumar, "High Dimensional Text Classification using Unsupervised Machine Learning Algorithm," 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2024, pp. 368-372, doi: 10.1109/ICAAIC60222.2024.10575444.

[6]. Ramesh Kumar Ramaswamy, Pannangi Naresh, Chilamakuru Nagesh, Santhosh Kumar Balan, Multilevel thresholding technique with Archery Gold Rush Optimization and PCNN-based childhood medulloblastoma classification using microscopic images, Biomedical Signal Processing and Control, Volume 107, 2025, 107801, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2025.107801.

[7]. G. Chanakya, N. Bhargavee, V. N. Kumar, V. Namitha, P. Naresh and S. Khaleelullah, "Machine Learning for Web Security: Strategies to Detect and Prevent Malicious Activities," 2024 Second International

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, February 2026
Impact Factor: 8.2

Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), Coimbatore, India, 2024, pp. 59-64, doi: 10.1109/ICoICI62503.2024.10696229.

[8]. S. Khaleelullah, P. Marry, P. Naresh, P. Srilatha, G. Sirisha and C. Nagesh, "A Framework for Design and Development of Message sharing using Open-Source Software," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 639-646, doi: 10.1109/ICSCDS56580.2023.10104679.

[9]. V. Krishna, Y. D. Solomon Raju, C. V. Raghavendran, P. Naresh and A. Rajesh, "Identification of Nutritional Deficiencies in Crops Using Machine Learning and Image Processing Techniques," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 925-929, doi: 10.1109/ICIEM54221.2022.9853072.

[10]. T. Aruna, P. Naresh, B. A. Kumar, B. K. Prakash, K. M. Mohan and P. M. Reddy, "Analyzing and Detecting Digital Counterfeit Images using DenseNet, ResNet and CNN," 2024 8th International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2024, pp. 248-252, doi: 10.1109/ICISC62624.2024.00049.

[11]. Dev, D. R., Biradar, V. S., Chandrasekhar, V., Sahni, V., & Negi, P. (2024). Uncertainty determination and reduction through novel approach for industrial IoT. Measurement: Sensors, 31, 100995. https://doi.org/10.1016/j.measen.2023.100995

[12]. Roy, R. E., Kulkarni, P., & Kumar, S. (2022, June). Machine learning techniques in predicting heart disease a survey. In 2022 IEEE world conference on applied intelligence and computing (AIC) (pp. 373-377). IEEE. doi: 10.1109/AIC55036.2022.9848945.

[13]. Darshan, R., Janmitha, S. N., Deekshith, S., Rajesh, T. M., & Gurudas, V. R. (2024, March). Machine Learning's Transformative Role in Human Activity Recognition Analysis. In 2024 IEEE International Conference on Contemporary Computing and Communications (InC4) (Vol. 1, pp. 1-8). IEEE. doi: 10.1109/InC460750.2024.10649391.

[14]. Sachin, A., Penukonda, A., Naveen, M., Chitrapur, P. G., Kulkarni, P., & BM, C. (2025, June). NAVISIGHT: A Deep Learning and Voice-Assisted System for Intelligent Indoor Navigation of the Visually Impaired. In 2025 3rd International Conference on Inventive Computing and Informatics (ICICI) (pp. 848-854). IEEE., doi: 10.1109/ICICI65870.2025.11069837.

[15]. Nagesh, C., Chaganti, K.R. , Chaganti, S. ,Khaleelullah, S., Naresh, P. and Hussan, M. 2023. Leveraging Machine Learning based Ensemble Time Series Prediction Model for Rainfall Using SVM, KNN and Advanced ARIMA+ E-GARCH. *International Journal on Recent and Innovation Trends in Computing and Communication*. 11, 7s (Jul. 2023), 353–358. DOI:https://doi.org/10.17762/ijritcc.v11i7s.7010.

[16]. Naresh, P., & Suguna, R. (2021). IPOC: An efficient approach for dynamic association rule generation using incremental data with updating supports. Indonesian Journal of Electrical Engineering and Computer Science, 24(2), 1084. https://doi.org/10.11591/ijeecs.v24.i2.pp1084-1090.

[17]. Swasthika Jain, T. J., Sardar, T. H., Sammeda Jain, T. J., Guru Prasad, M. S., & Naresh, P. (2025). Facial Expression Analysis for Efficient Disease Classification in Sheep Using a 3NM-CTA and LIFA-Based Framework. IETE Journal of Research, 1–15. https://doi.org/10.1080/03772063.2025.2498610.

[18]. P. Naresh, S. V. N. Pavan, A. R. Mohammed, N. Chanti and M. Tharun, "Comparative Study of Machine Learning Algorithms for Fake Review Detection with Emphasis on SVM," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 170-176, doi: 10.1109/ICSCSS57650.2023.10169190.

[19]. T. Kavitha, K. R. Chaganti, S. L. R. Elicherla, M. R. Kumar, D. Chaithanya and K. Manikanta, "Deep Reinforcement Learning for Energy Efficiency Optimization using Autonomous Waste Management in Smart Cities," 2025 5th International Conference on Trends in Material Science and Inventive Materials (ICTMIM), Kanyakumari, India, 2025, pp. 272-278, doi: 10.1109/ICTMIM65579.2025.10988394.

[20]. N. P, K. R. Chaganti, S. L. R. Elicherla, S. Guddati, A. Swarna and P. T. Reddy, "Optimizing Latency and Communication in Federated Edge Computing with LAFEO and Gradient Compression for Real-Time Edge

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, February 2026
Impact Factor: 8.2

Analytics," 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), Goathgaun, Nepal, 2025, pp. 608-613, doi: 10.1109/ICMCSI64620.2025.10883220.

[21]. SAI M, RAMESH P, REDDY DS. EFFICIENT SUPERVISED MACHINE LEARNING FOR CYBERSECURITY APPLICATIONS USING ADAPTIVE FEATURE SELECTION AND EXPLAINABLE AI SCENARIOS. Journal of Theoretical and Applied Information Technology. 2025 Mar 31;103(6).

[22]. Sivananda Reddy Elicherla, Dr. P E Sreenivasa Reddy, Dr. V Raghunatha Reddy and Sivaprasada Reddy Peddareddigari. "Agilimation (Agile Automation) - State of Art from Agility to Automation." International Journal for Scientific Research and Development 3.9 (2015): 411-416.