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

DOI: 10.48175/IJARSCT-31127