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: 7.67

Explainable and Trustworthy Generative AI: A Framework for Interpretable Large Language Models in High-Stakes Decision Systems

[1]Dr. Syeda Farhath Begum and [2]Dr. Farheen Sultana
[1]Associate Professor, Dept of CSE, Nawab Shah Alam Khan College of Engineering and Technology, Hyderabad
[2]Associate Professor, Dept of IT, Nawab Shah Alam Khan College of Engineering and Technology, Hyderabad

Abstract: *The rapid adoption of large language models (LLMs) in high-stakes decision systems such as healthcare, finance, law, and public governance has raised critical concerns regarding transparency, reliability, and trustworthiness. While generative AI models demonstrate remarkable performance, their black-box nature limits interpretability and poses significant risks when decisions directly impact human lives. This paper proposes a comprehensive framework for Explainable and Trustworthy Generative AI, aimed at enhancing the interpretability, accountability, and robustness of large language models deployed in high-stakes environments.The proposed framework integrates intrinsic interpretability mechanisms, post-hoc explanation techniques, and uncertainty-aware reasoning to provide transparent model behavior at both global and local decision levels. Trustworthiness is further strengthened through bias detection, fairness auditing, robustness evaluation, and human-in-the-loop validation. Additionally, the framework incorporates governance-oriented components, including ethical compliance, auditability, and regulatory alignment, to support responsible AI deployment.Experimental evaluation across multiple high-stakes use cases demonstrates that the proposed approach significantly improves decision transparency and user trust while maintaining competitive predictive and generative performance. The results highlight the potential of interpretable generative AI systems to bridge the gap between model capability and real-world accountability. This work contributes toward the development of reliable, explainable, and ethically aligned large language models suitable for critical decision-making applications.*

Keywords: Explainable Artificial Intelligence (XAI), Trustworthy AI, Generative AI, Large Language Models (LLMs), Interpretability, High-Stakes Decision Systems, Model Transparency, Ethical AI

DOI: 10.48175/IJARSCT-31102