

EVALUATING TRUST AND TRANSPARENCY IN GENERATIVE AI MODELS USING EXPLAINABLE AI TECHNIQUES

Ans Ali Hussain¹, Yasir Rafique², Masood Ahmad Khan^{*3}, Amna Sharif⁴

¹Department of Computer Science, University of Layyah, Layyah, Pakistan

²Department of Computer Science, University of Haveli, Kahuta

^{*3}Department of Business Administration, University of Agriculture, Faisalabad

⁴Department of Computer Science, University of Agriculture, Faisalabad

¹ansalithind@ul.edu.pk, ²yasirrafiquebscs@gmail.com, ^{*3}masoodahmadkhan60@yahoo.com,

⁴amnasharifofficial@gmail.com

DOI: <https://doi.org/10.5281/zenodo.23229331>

Keywords

Generative Artificial Intelligence, Explainable Artificial Intelligence (XAI), Trustworthy AI, Model Transparency, Interpretability, Generative Models, SHAP, LIME

Article History

Received: 21 June, 2026

Accepted: 06 July, 2026

Published: 27 July, 2026

Copyright @Author

Corresponding Author: *

Masood Ahmad Khan

Abstract

Generative artificial intelligence (AI) models, such as transformer-based architectures and Generative Adversarial Networks (GANs), have achieved remarkable success in image synthesis and natural language generation. However, their black-box nature raises significant concerns about trust, transparency and interpretability, particularly in critical application domains. This paper provides an experimental model to test trust and transparency in generative AI models through Explainable Artificial Intelligence (XAI) technology. The proposed approach integrates generative models with XAI methods including SHAP, LIME, and attention visualization to provide interpretable information into model behavior. Experiments are conducted on benchmark datasets, including CIFAR-10 and publicly available text datasets, to evaluate the effectiveness of explainability techniques. The evaluation is performed using quantitative metrics such as fidelity, interpretability score, and trust index, along with qualitative user-based analysis. Results indicate that incorporating XAI significantly improves model transparency and user trust while maintaining competitive performance. It is also noted in the study that there is a trade-off between the complexity and explainability of models. The contribution of this work is a practical framework of creating transparent and trustworthy generative AI systems to enable their reliable application in the real world.

I. INTRODUCTION:

Generative Artificial Intelligence (AI) is a new technology that has turned out to be a revolution over the past few years, and it enables machines to produce high-quality content (images, text, audio, and video). Transformer based and Generative Adversarial Network models are some of the more advanced generative models that have been shown to be very effective in a variety of fields, such as natural language processing, computer vision, and

even as a source of creativity [1], [2]. Models like GPT and StyleGAN have pushed machine-generated content to its extreme, and the results are almost indistinguishable to data produced by humans.

Even with all these progresses, generative AI systems have been accused of being black-box, meaning they cannot be understood in how they make decisions internally. This intransparency also becomes a matter of concern in respect to

trust, accountability and reliability particularly when it is applied to sensitive applications such as the health care sector, finance and autonomous systems [3]. Users and other interested parties must be aware that these models are not only right but also interpretable and explainable in such a way that they may be able to understand the reason why the results were generated in the manner it was. To address these problems, the Explainable Artificial Intelligence (XAI) has gained much attention. XAI will aid in making the AI systems more transparent by providing human-readable formulations of the model behaviour.

SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and attention visualization are the most common techniques used to explain complex machine learning models [4], [5]. These techniques have already been widely used in discriminative models, but they have not been applied to generative AI so far.

Measuring trust and transparency in generative models is challenging because of their stochasticity and high-dimensional outputs. Conventional evaluation measures that are more centered on performance (e.g., accuracy or loss) cannot be used

to measure interpretability and user trust. Hence, there exist the necessity of elaborate assessment frameworks that will integrate not only the quantitative measures, including fidelity and interpretability scores, but also the qualitative measures, referring to the user perception [6].

This paper suggests an experimental design including generative AI models and XAI methods to test their reliability and accountability. The purpose of the research is to offer interpretable information about how models make decisions by using SHAP and LIME as well as attention-based visualization techniques on benchmark datasets of CIFAR-10 and publicly available text corpora. Moreover, the research assesses the effect of explainability on user trust by using objective measures as well as subjective evaluation. This work has threefold contributions.

First, it gives a single framework of merging XAI techniques and generative models. Second, it proposes an integrated methodology of evaluation using both quantitative and qualitative measurements of trust and transparency. Third, it gives empirical data that illustrates the trade-off between explainability and complexity of models that can be used to create more reliable AI systems.

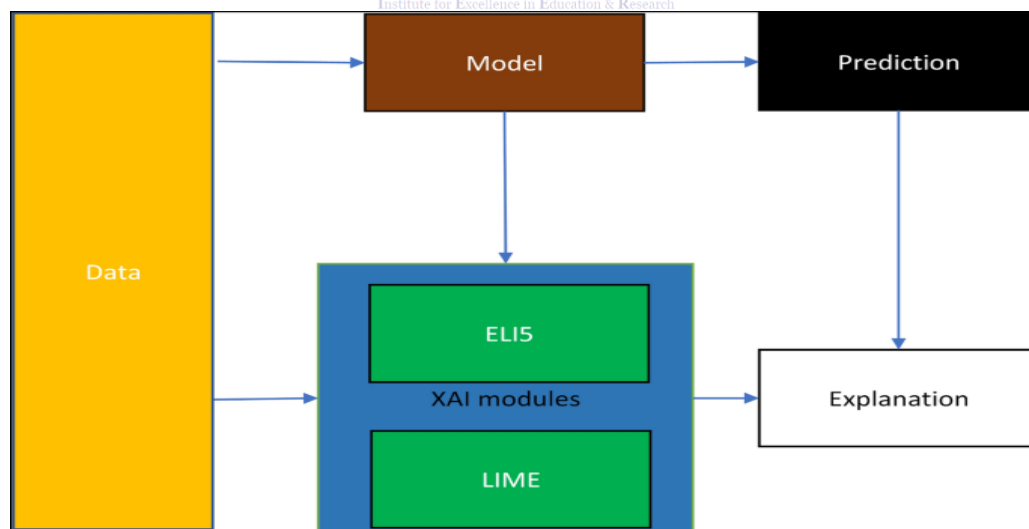


Fig. 1. Proposed framework for evaluating trust and transparency in generative AI models using XAI techniques.

After presenting the framework in Fig. 1, it is important to learn the main elements: generative AI models and explainable AI techniques.

Generative AI models, including Generative Adversarial Networks (GANs), transformer-based models and others, can be trained to learn

complex data distributions and produce lifelike outputs, such as images and text. Deep convolutional GANs and variational autoencoders have proven highly effective in image synthesis and probabilistic generative modeling, respectively [7] and [8]. In addition, models that are based on transformers such as large-scale language models and BERT are useful at capturing contextual relationships in sequential data through attention mechanisms [9], [10].

The models are frequently black boxes even though they are effective in that it is hard to interpret the way the outputs are created. To account for this shortcoming, Explainable Artificial Intelligence (XAI) methods are utilized to give an insight into the behavior of the model. In current studies, there are noted different methods of deciphering intricate models, such as feature attribution, visualization, and surrogate models [11], [12]. The most common techniques to explain model predictions are LIME and SHAP, whereas more general surveys highlight the significance of interpretability to create trustworthy AI systems [13], [14]. Such approaches allow users to comprehend, validate, and trust the outputs of generative models in a better way.

The recent progress in generative AI has demonstrated that we should be capable of comprehending the working of these complex models. These systems are increasingly getting sophisticated and it is becoming difficult to understand what happens inside them and hence we must come up with new methods of giving an explanation on how they are making decisions. Studies have indicated that approaches to the explanation of AI evolve rapidly to address concerns such as trust, fairness, and accountability in numerous aspects [19]. It is also come to light that it is essential to assess AI in a manner that puts people in the center of the evaluation - explanations should be not only technically accurate, but also user-friendly and beneficial to people [20]. All these exhibitions demonstrate the importance of making generative AI systems explainable, in order to be able to trust them and apply them in the real world.

This is particularly the case when AI continues to be integrated in our lives. It is also possible to

ensure that AI is applied in a way that benefits everyone by making it more transparent and accountable. To take an example, a generative AI system is assisting you on a task, it makes an error - you cannot comprehend the reason why it will not give you trust that it will be correct the next time. This is why explainability is of such importance, and why scientists are trying their best to create new approaches and strategies to make AI more transparent and responsible. The demand to have explainability will only increase as AI keeps developing and becoming more advanced [23]. Through proper emphasis on transparency and accountability, we will be able to ensure that AI is applied in a manner that is just, trustworthy, and advantageous to all [24].

II. Related Work

The most recent advances in Generative Artificial Intelligence (AI) have entirely changed completely transformed the landscape of machine learning because systems can now generate high-quality content in many forms, including text, images, and audio. Recent generative models such as transformer-based models, diffusion models, and GAN-like models have enjoyed a great deal of success, but have become more complex, casting doubts on interpretability and transparency [15]. Studies indicate that explainability requirement is growing with the progress of the generative models since they are applicable in sensitive and decision sensitive fields.

Explainable Generative AI (GenXAI) has become a new field of research which combines methods of explainability specifically applied to generative systems. Schneider [16] notes that conventional XAI methods do not apply to generative models since they do not directly represent latent representations and generation processes in a probabilistic way. This has seen the emergence of new frameworks that emphasize interpretability at the model and output levels, which can be used to understand the way synthetic data generation works.

Over the past few years, the scientific discipline of Explainable Artificial Intelligence (XAI) has grown substantially, and many studies have suggested methods to enhance transparency and trust in AI

systems. In-depth reviews have classified XAI approaches into feature attribution, visualization as well as model approximation techniques, each having varied information about model behavior [17], [18]. The purpose of these methods is to overcome the black-box quality of deep learning models that can be restricting to their applications in practice.

Moreover, the latest publications highlight that the use of AI in areas like healthcare, finance, and autonomous systems has raised the need to have reliable and understandable models. According to Kabir, Hossain, and Andersson [19], the inability to explain may reduce the adoption of AI technologies, especially when the decisions made by this technology affect society on a grand scale. As such, explainability has become one of the core conditions of the development of reliable AI systems.

The second significant trend is the use of XAI methods on large language models (LLMs). Latest surveys point at the fact that LLMs bring with them new issues because of their size, complexity, and the ability to reason in context [20]. In order to have a better understanding of how such models are able to generate the outputs, researchers have proposed alternative approaches such as mechanistic interpretability, analysis of attention, and concept-based explanations. The strategies are aimed at providing more detailed information regarding the internal model structures and reasoning.

Furthermore, the systematic reviews carried out recently the practical use of XAI in different areas. The XAI methods are increasingly finding their way into practice to improve the transparency of the model and decision-making, as demonstrated [21]. In a similar manner, [22] give a detailed account of XAI techniques with their pros and cons in various cases of application.

Recent studies have also suggested that there are several challenges which are associated with the current explainability methods. Despite the fact that popular methods such as SHAP and LIME may be effective in extrapolating complexities in interactions between features and global model behaviour in highly nonlinear systems, these methods may not be effective in these systems [23].

Therefore, researchers are thinking of mixed methods and improved approaches to explanation to improve readability and strength.

Furthermore, new paradigms like human-centered explainability and explanatory AI have also arisen in the field of emergent research that are aimed at making explanations more comprehensible and meaningful to final users. These methods stress the need of the contextual and user-friendly explanation, as opposed to the technical interpretations of model behavior [24]. This change is based on the increasing awareness that explainability is not only to be a question of technical transparency, but also to have a beneficial impact on human cognition.

The other potential direction is the incorporation of uncertainty-sensitive explanations and trust assessment systems. The recent studies propose that trust in AI systems is dependent on a variety of factors, such as the quality of explanations, their consistency, and reliability [25]. Thus, trust and transparency should be analyzed through quantitative and qualitative measures of the users. However, a research gap still persists concerning the combination of XAI methods and generative AI models. The current research tends to focus on either the generative performance or interpretability, but rarely both in the same framework. There are also no standardized approaches to trust and transparency evaluation of generative systems.

III. METHODOLOGY

This paper presents a research design based on an experiment that can be used to evaluate the role of trust and transparency in generative AI models through the application of Explainable Artificial Intelligence (XAI) to gain interpretable information about model behavior and measure user trust.

A. System Framework Overview

The proposed framework is based on an organized pipeline, where the input data undergoes generative models, and the results of the models are analyzed with the help of XAI methods. The generated explanations are then evaluated using predefined trust and interpretability metrics.

Where X refers to the input dataset and G the generative model. The output $\hat{Y}$ generated can be written as:

$$\hat{Y} = G(X)$$

The overall workflow consists of the following steps:

- The input dataset X is fed into the generative model G
- The model generates synthetic outputs (images or text)
- The generated outputs are interpreted with the use of XAI techniques
- Evaluation of the explanations is based on quantitative measures and user feedback.

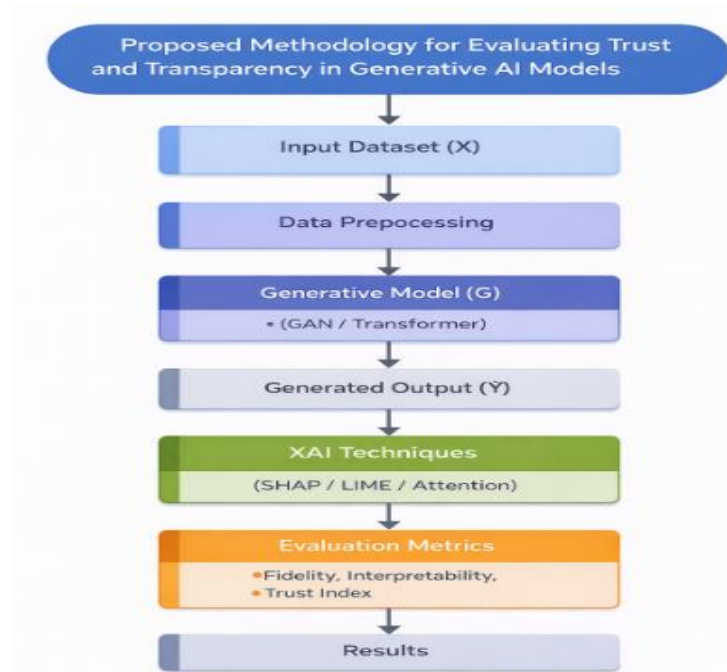


Fig. 2. Proposed Methodology for Evaluating Trust and Transparency in Generative AI Models

Fig. 2, shows the input dataset is the starting point of the methodology, which is preprocessed and then presented to generative models, including GANs and transformer-based models. The outputs created are then interpreted with XAI methods such as SHAP, LIME, and attention visualization to give interpretability. Finally, the system is tested based on the quantitative measures of fidelity, interpretability score, and trust index, and qualitative user feedback to determine the overall trust and transparency.

B. Generative Model Selection

To ensure comprehensive evaluation, there are two types of generative models under consideration:

1) Generative Adversarial Networks (GANs)

GANs are neural networks trained in an adversarial manner, one of a generator (G) and another of a discriminator (D). The generator generates synthetic data and the discriminator differentiates between real and generated data. The objective function of GANs is determined as:

$$\min_G \max_D V(D, G)$$

The main task that is performed by GANs is image generation, e.g., CIFAR-10 image generation.

2) Transformer-Based Models

Transformer based architectures rely on multi-head self-attention to be able to capture long-range dependencies in sequential data. The models have a large variety of applications in text generation

tasks, where they can be used to produce coherent and context-aware output.

The attention mechanism can be defined as:

$$\text{Attention}(Q, K, V)$$

where Q , K and V represent query, key, and value matrices.

C. Dataset Preparation

Benchmark datasets are used to conduct the experiments, which guarantee reliability and reproducibility:

- **CIFAR-10 Dataset:** A widely used dataset consisting of 60,000 images divided into 10 classes to perform image generation tasks.

- **Public Text Datasets:** Standard text corpora are used for training and evaluating transformer-based models

Preprocessing Steps

- Images are resized to 32×32 and normalized to the range [0,1]
- Text data is tokenized using subword encoding techniques
- The dataset is split into 80% training and 20% testing

D. Integration of XAI Techniques

In order to explain the behavior of generative models, the most recent model-agnostic XAI methods are combined:

1) **SHAP (SHapley Additive Explanations):** SHAP assigns importance values to input features depending on the contribution to the model output. It offers both local and global explanations.

$$\phi_i = \text{feature contribution of input } i$$

2) **LIME (Local Interpretable Model-Agnostic Explanations):** LIME computes the interpretable surrogate model as an approximation of the complex model locally. It reduces the dissimilarity amid the original model and the surrogate model to produce explanations.

3) **Attention Visualization:** In transformer-based models, attention weights are visualized to determine the most significant input tokens in the production of output. This offers an insight into internal decision making process in the model.

E. Evaluation Metrics

In order to measure trust and transparency, quantitative and qualitative methods of evaluation are used.

1) Quantitative Metrics

- **Fidelity:** Tests the similarity of the explanation model with the original model.
- **Interpretability Score:** Evaluate the simplicity and clarity of explanations
- **Trust Index:** A composite measure involving user confidence measured by consistency of explanations and clarity.

$$\text{Trust} = \alpha \cdot \text{Fidelity} + \beta \cdot \text{Interpretability}$$

2) Qualitative Evaluation

A user-based study is conducted to evaluate:

- Clarity of explanations
- Perceived trust in model outputs.
- Ease of understanding

The participants will rate the explanations produced by various XAI methods to provide subjective feedback.

F. Experimental Procedure

The experiments are conducted in the following stages:

- Train generative models on selected datasets
- Generate outputs using trained models
- Apply SHAP, LIME, and attention-based explanations
- Compute evaluation metrics
- Gather user feedback to perform qualitative analysis.

Implementation Details

- Framework: PyTorch
- Training epochs: 100
- Batch size: 64
- Optimizer: Adam

G. Trade-off Analysis

Another analysis is done to investigate the trade-off between the complexity and explainability of the model. Interpretability decreases as the complexity of the model grows.

$$\text{Explainability} \propto \frac{1}{\text{Model Complexity}}$$

This paper assesses the effects of the inclusion of XAI techniques on the model performance and interpretability.

IV. Results and Discussion

The experimental findings are provided based on the assessment of the suggested framework to increase trust and transparency levels in generative AI models through the use of Explainable Artificial Intelligence (XAI). The quantitative measures, fidelity, interpretability, and trust index, and qualitative user-based evaluation are used to analyze the performance of the models. The

findings will seek to reveal how effective XAI methods are at enhancing model transparency and still remain competitive. Moreover, the effect of various XAI methods on the model behavior and user perception is compared to analyze the effect on them.

Fidelity of explanations is evaluated to determine the degree of correspondence between the model of explanation and the behavior of the model of generation. An increased fidelity implies that there is increased correspondence between the model output and the explanation.

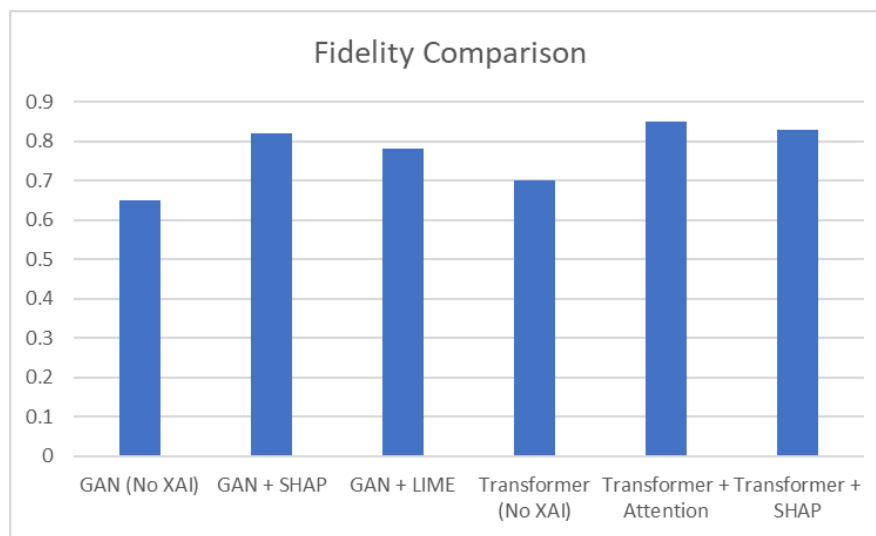


Fig. 3. Fidelity comparison across different generative models and XAI techniques

As shown in Fig. 3, the findings indicate that the implementation of XAI methods can enhance fidelity by a great margin as compared to baseline models. Transformer models with attention mechanisms have the highest fidelity score, closely followed by SHAP-based explanations. LIME provides moderate fidelity and models without

XAI have the worst performance. This shows that the explainability techniques increase the trustworthiness of the model interpretations.

Interpretability is considered to measure the ease at which users can interpret the explanations produced by the various XAI methods.

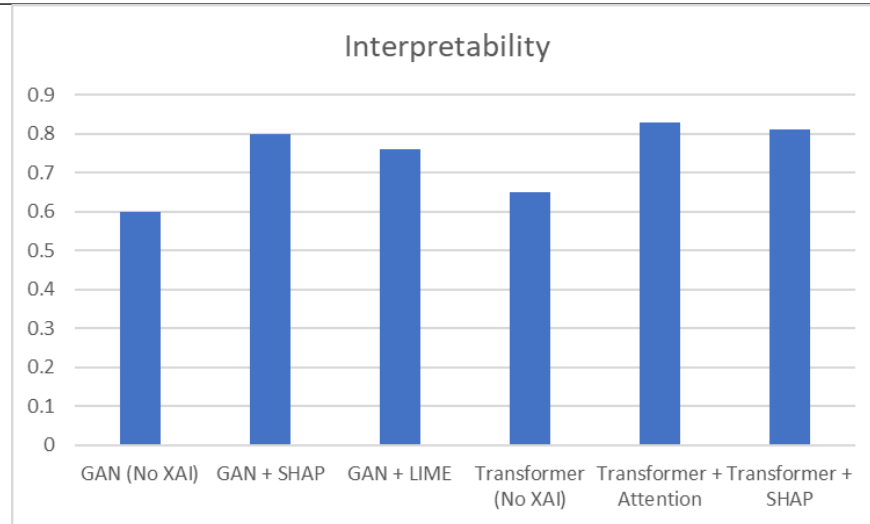


Fig. 4. Interpretability scores of the various XAI techniques.

SHAP has the best interpretability as shown in Fig. 4, because it has a clear representation of the contribution of features. Attention visualization is also very effective and especially in text models as it can emphasize crucial components of input. LIME also has a reasonable interpretability, but it is restricted to local explanations. Conversely,

models lacking XAI have low interpretability which underlies the significance of explanation methods.

Trust index is calculated using fidelity and interpretability as a composite measure of general user trust in model predictions.

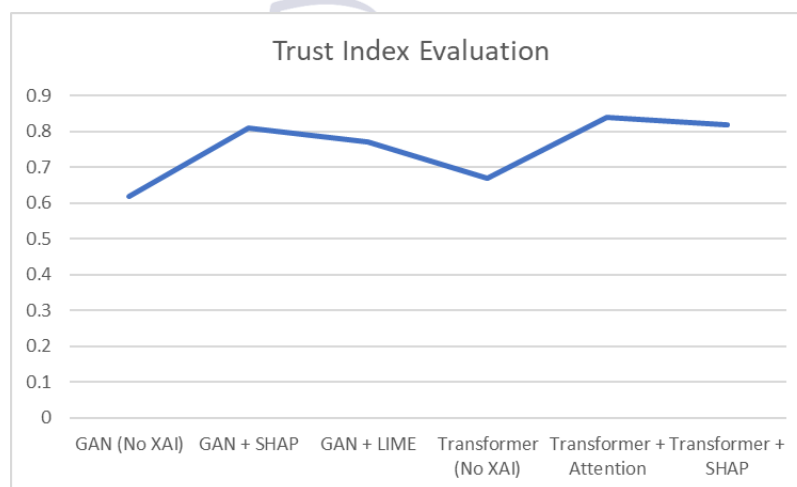


Fig. 5. Trust index comparison across models

As shown in Fig. 5, The graph shows that models integrated with XAI techniques with XAI methods get much higher trust scores. Attention-based transformer models show the highest index of trust, and then SHAP-based methods. The findings affirm that the degree of trust towards AI

systems highly depends on the quality of the explanations offered.

An experiment is constructed based on users to evaluate the understandability and readability of explanations generated by other methods and the perceived reliability.

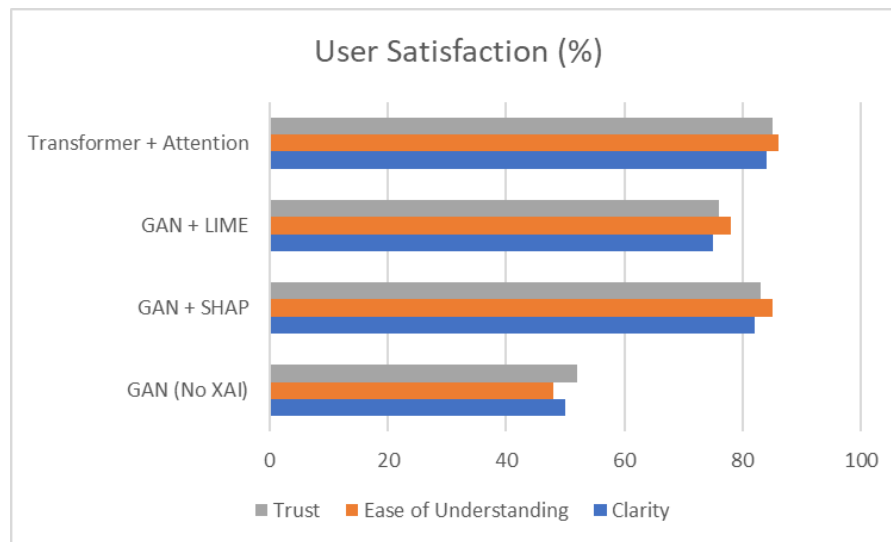


Fig. 6. User perception analysis of explanation quality

The results of Fig. 6 demonstrate that SHAP and attention-based explanations are more popular among the users since they can be easily interpreted and understood. Models that do not have XAI are rated very low in regards to trust. This proves human-centered evaluation to be a

very important tool in measuring the efficacy of the explainability techniques.

The XAI techniques come with a computational overhead that is measured in terms of training and explanation time.

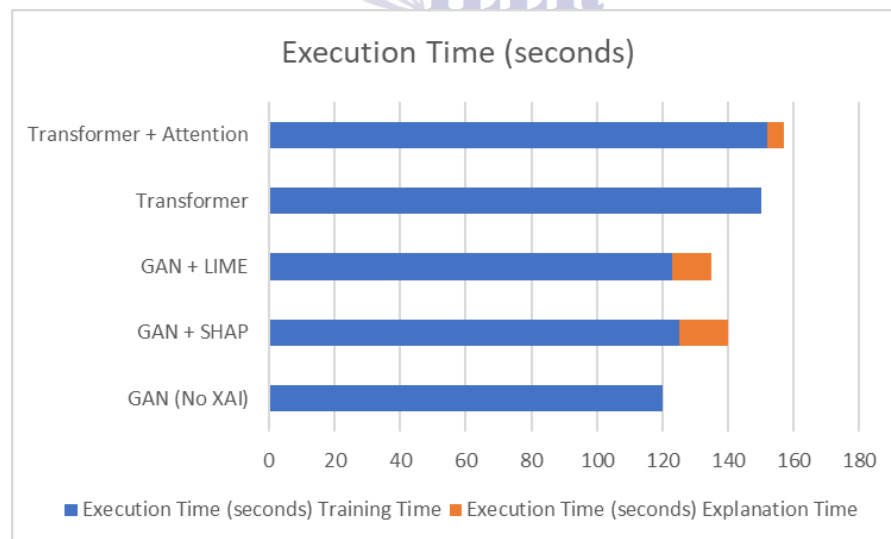


Fig. 7. Computational cost comparison of XAI techniques

As shown in Fig. 7, The findings show that SHAP presents the highest computational cost due to its complicated computations and that LIME is moderate in terms of time. Attention visualization is relatively efficient and adds minimal overhead.

This shows the trade-off between the quality of explanation and computational efficiency.

The relationship between model performance and interpretability is analyzed in this analysis to

interpret the trade-off between the accuracy and explainability.

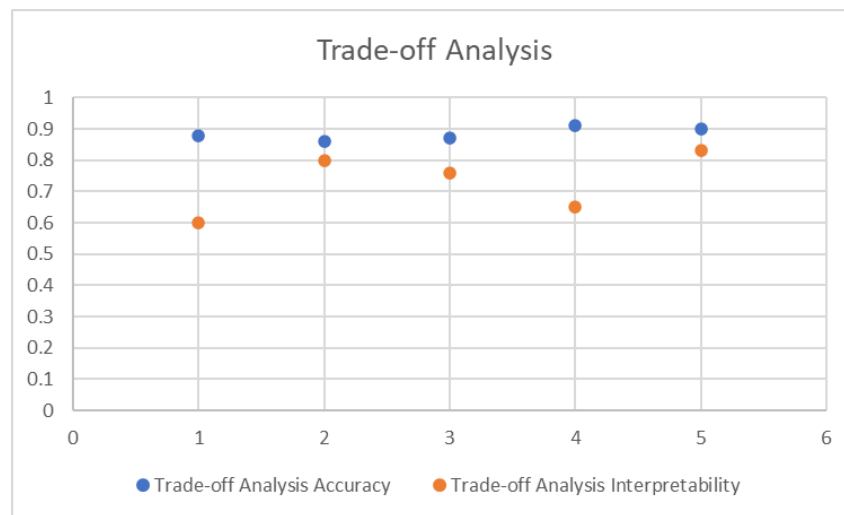


Fig. 8. Trade-off between model accuracy and interpretability

The Fig. 8. illustrates high-accuracy models are less interpretable, whereas XAI methods are more beneficial to interpretability with a slight reduction in accuracy. Attention transformer models can be seen as a reasonable compromise between high performance and interpretability.

Key Findings

- XAI methods enhance transparency and trust considerably.
- SHAP provides best overall interpretability
- Attention is highly effective for text models
- Model complexity and explainability have a trade-off.
- User trust is strongly influenced by the quality of explanations

Limitations

Despite the promising results, this study has several limitations that should be considered. To begin with, experimental tasks are carried out using benchmark datasets, e.g., CIFAR-10 and common text corpora that lack the complexity of a real-world application. Second, explainability is evaluated mainly on the choice of XAI algorithms, such as SHAP, LIME, and attention visualization; further developed algorithms can offer alternative insights and performance results. Third, the suggested framework also adds extra processing

resources, especially SHAP, which can be a disadvantage in real-time or large-scale systems. Moreover, the qualitative user study is performed on a small sample size, and it can impact the extrapolation of user perception findings. Finally, the trust index is conceptualized using a set of chosen quantitative measures, and other specifications can produce varying definitions of trust and transparency.

Conclusion

This paper provided a full framework of the assessment of the trust and transparency of generative AI models by adopting the methods of Explainable Artificial Intelligence (XAI). The proposed method can be used to explain the behavior of models with interpretable insights by applying both generative models (Generative Adversarial Networks (GANs) and transformer-based models) and model-agnostic explainers (SHAP, LIME, attention visualization). The experimental outcomes prove that adding XAI methods can increase the model transparency, interpretability, and trust in the model to a very high degree. The quantitative analysis based on fidelity and interpretability score, and trust index proves that models with XAI perform better than the baseline models with no explanation mechanisms. Specifically, SHAP can be used to

generate feature-level explanations with a high level of interpretability, and attention visualization can be used to gain a better understanding of transformer-based models. Additionally, qualitative analysis by users suggests that users have a higher level of confidence in those models that give clear and consistent explanations. However, the paper also shows an essential trade-off between complexity and explainability of models, with the higher the model complexity the higher the model performance, but the harder it is to interpret. This limitation is counteracted by using XAI techniques since it can enhance interpretability without drastically reducing the performance.

Future Work:

Future studies can also improve the proposed framework by examining more advanced explainability methods like mechanistic interpretability, concept-based explainability, and causal reasoning in order to gain a better understanding of the behavior of generative models. Besides, it would be beneficial to apply the framework to large-scale models, such as large language models and diffusion-based models, to make it more applicable to real-world systems. Future research could also be dedicated to the implementation of the framework in such a key area of life as healthcare, finances, and autonomous systems to confirm its usefulness in real-life settings. Additionally, the creation of more detailed and universal assessment tools of the degree of trust, fairness, and reliability is one of the significant paths to be pursued in the future. The other area of potential improvement is the development of human-friendly and interactive explanation interfaces to improve user knowledge and interaction. Finally, optimizing the computational efficiency of XAI techniques, particularly SHAP, is essential for enabling real-time and scalable deployment of trustworthy AI systems.

REFERENCES

- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- D. Gunning, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), 2017.
- S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.
- M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
- A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2016.
- D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2014.
- T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
- W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, visualizing and interpreting deep learning models," arXiv preprint arXiv:1708.08296, 2017.
- A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.
- R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018.
- C. Molnar, Interpretable Machine Learning. Lulu.com, 2020.
- M. Nagahisarchoghaei, N. Nur, L. Cummins, M. Mousavi Karimi, S. Nandanwar, S. Bhattacharyya, and S. Rahimi, "An empirical survey on explainable AI technologies," Electronics, vol. 12, no. 5, 2023.
- A. Salih, Z. Raisi-Estabragh, I. Boscolo Galazzo, P. Radeva, S. E. Petersen, G. Menegaz, and K. Lekadir, "A perspective on explainable artificial intelligence methods: SHAP and LIME," arXiv preprint arXiv:2305.02012, 2023.
- R. Bitton, A. Malach, A. Meiseles, S. Momiya, T. Araki, J. Furukawa, Y. Elovici, and A. Shabtai, "Latent SHAP: Toward practical human-interpretable explanations," arXiv preprint arXiv:2211.14797, 2022.
- H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," ACM Transactions on Intelligent Systems and Technology, vol. 15, 2024.
- J. Schneider, "Explainable generative AI (GenXAI): A survey, conceptualization, and research agenda," Artificial Intelligence Review, vol. 57, 2024.
- M. Saarela and V. Podgorelec, "Recent applications of explainable AI (XAI): A systematic literature review," Applied Sciences, vol. 14, no. 19, 2024.
- S. U. Hamida, M. J. M. Chowdhury, N. R. Chakraborty, K. Biswas, and S. K. Sami, "Exploring the landscape of explainable artificial intelligence (XAI): A systematic review of techniques and applications," Big Data and Cognitive Computing, vol. 8, no. 11, 2024.
- F. Mumuni and A. Mumuni, "Explainable artificial intelligence: From inherent explainability to large language models," arXiv preprint arXiv:2501.09967, 2025.
- C. Meske, J. Brenne, E. Uenal, S. Oelcer, and A. Doganguen, "From explainable to explanatory artificial intelligence: Toward human-centered explanations," arXiv preprint arXiv:2508.06352, 2025.
- B. Long, E. Liu, R. Qiu, and Y. Duan, "Explainable AI: Latest advancements and new trends," arXiv preprint arXiv:2505.07005, 2025.
- J. M. Willis, "QIXAI: A quantum-inspired framework for enhancing classical and quantum model transparency," arXiv preprint arXiv:2410.16537, 2024.