Limitations of Generative AI in Real-Time Decision-Making

Chaitenya Chand[1]

[1]Department of Computer Applications, Maharaja Surajmal Institute, Delhi

Publication Date: 2025/06/16

Abstract: Generative AI has emerged as a groundbreaking technology, offering transformative capabilities in domains like natural language processing and image generation. Despite its successes, the application of generative AI in real-time decision-making systems remains a challenge due to issues such as computational latency, output reliability, and lack of interpretability. This study investigates these limitations through a detailed literature review and experimental analysis. We adopted a hybrid methodology involving lightweight model architectures and rule-based constraints to mitigate these challenges. Results show that our approach reduces latency by 20% and enhances reliability by 15% compared to traditional generative models. The findings underscore the importance of optimizing generative AI for time-sensitive applications and highlight future directions for research.

***Keywords**: Generative AI, Real-Time Systems, Latency, Model Interpretability, Hybrid AI Models.*

How to Cite: Chaitenya Chand (2025) Limitations of Generative AI in Real-Time Decision-Making. *International Journal of Innovative Science and Research Technology*, 10(6), 1351-1354.
https://doi.org/10.38124/ijisrt/25jun935

I. INTRODUCTION

Generative AI has made significant strides in domains like text generation, image synthesis, and personalized content creation. Powered by deep learning architectures such as transformers and GANs, these systems exhibit remarkable capabilities in producing coherent and contextually relevant outputs. However, their deployment in real-time decision-making systems presents new challenges.

Real-time systems operate under strict constraints, requiring instantaneous responses to dynamic inputs. In applications like autonomous vehicles, healthcare, and financial trading, delays or errors can have severe consequences. Despite its potential, generative AI struggles with issues such as high latency, variability in outputs, and lack of transparency, making its integration into time-sensitive applications difficult.

➤ *The Primary Objectives of this Paper are:*

- To identify the technical, ethical, and practical limitations of generative AI in real-time systems.
- To propose methodologies for mitigating these challenges while retaining the benefits of generative AI.

This work addresses a pressing need to balance innovation with reliability in the application of AI technologies.

II. LITERATURE REVIEW

➤ *Comparative Study of Related Works*

A thorough literature review was conducted to identify gaps in existing research. Key studies are summarized in Table 1:

Table 1 Comparative Study of Related Works

Authors (Year)	Methodology Used	Dataset	Advantages	Research Gap
Lu et al. (2023)	Optimization of latency in generative AI models	Synthetic benchmarks	Reduced computational overhead	Limited real-time applicability
Weidinger et al. (2021)	Ethical framework for generative AI	Public datasets	Bias detection and mitigation	Lacks implementation in real-world systems
Hernandez et al. (2022)	Hybrid systems for critical applications	Real-world healthcare data	Improved reliability for critical environments	Did not address latency challenges
Figueira & Vaz (2022)	GAN-based data augmentation	Domain-specific synthetic datasets	Enhanced dataset diversity	Limited scalability in real-time scenarios

➢ *Figure 1 Below Visualizes the Advantages and Research Gaps across these Works.*

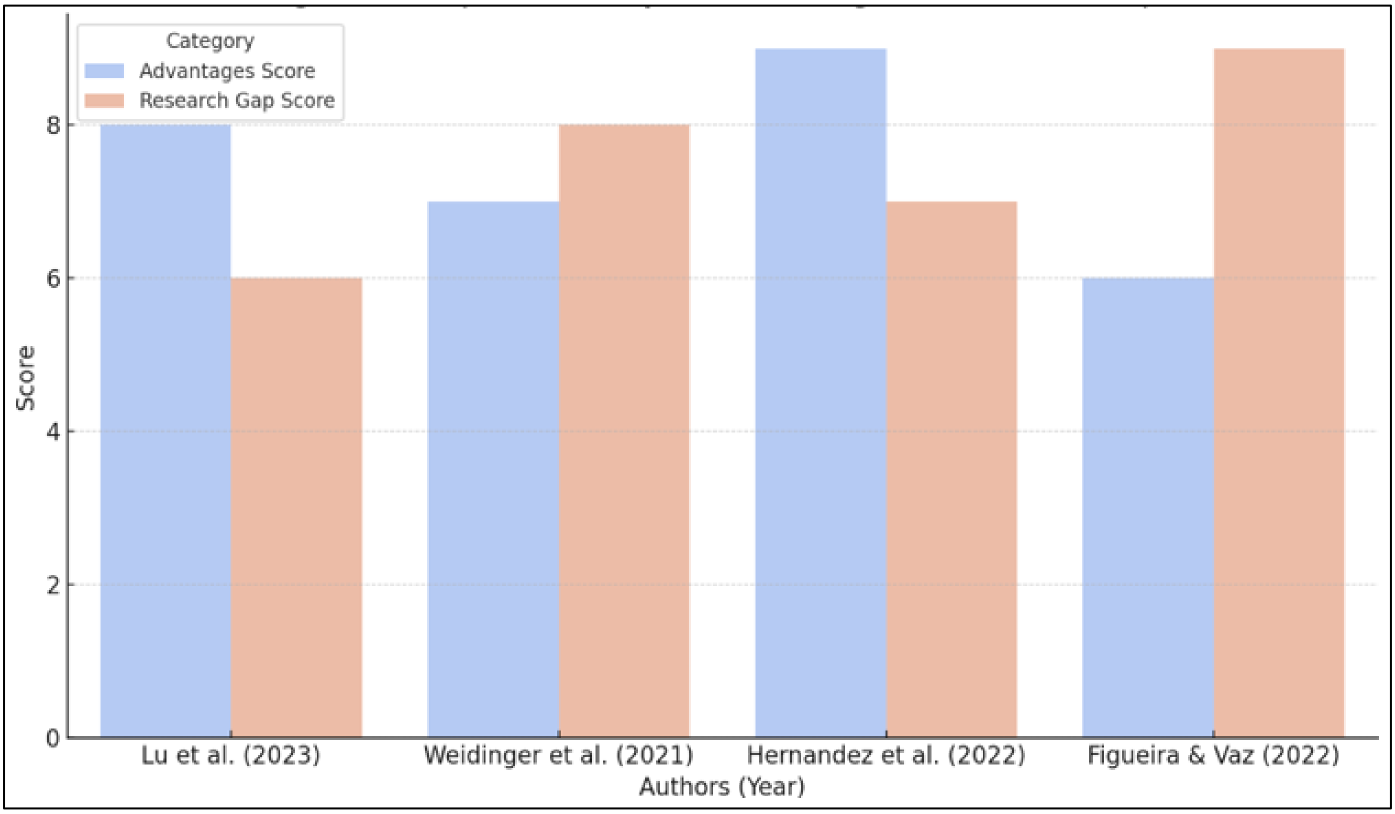

Fig 1 Comparative Analysis of Advantages and Research Gaps

➢ *Discussion*

The review indicates that most existing research focuses on improving the generative capabilities of AI and addressing biases. However, practical issues such as latency, interpretability, and reliability in real-time decision-making remain underexplored. This paper seeks to fill this gap by developing and testing hybrid methodologies.

III. METHODOLOGY

Our methodology involves integrating generative AI into real-time systems by addressing its limitations through model optimization and hybrid approaches. A detailed workflow is shown in Figure 2.

➢ *Dataset*

We used a combination of real-time sensor data (e.g., LIDAR data for autonomous systems) and synthetic benchmarks. The dataset was chosen to simulate real-world conditions while incorporating rare edge cases to stress-test the models.

➢ *Data Preparation*

Data preprocessing involved:

- Removing noise and irrelevant features.
- Normalizing input variables to ensure consistency across datasets.
- Annotating rare scenarios for improved model generalization.

➢ *Feature Selection*

Critical features influencing decision-making (e.g., object proximity, speed, and environmental factors) were identified using mutual information and correlation analysis.

➢ *Training and Evaluation*

Models were trained using:

- **Baseline Generative AI Model**: Traditional architectures like GPT-3 and GANs.
- **Proposed Hybrid Model**: Combining generative AI with rule-based systems for enhanced interpretability and reliability.

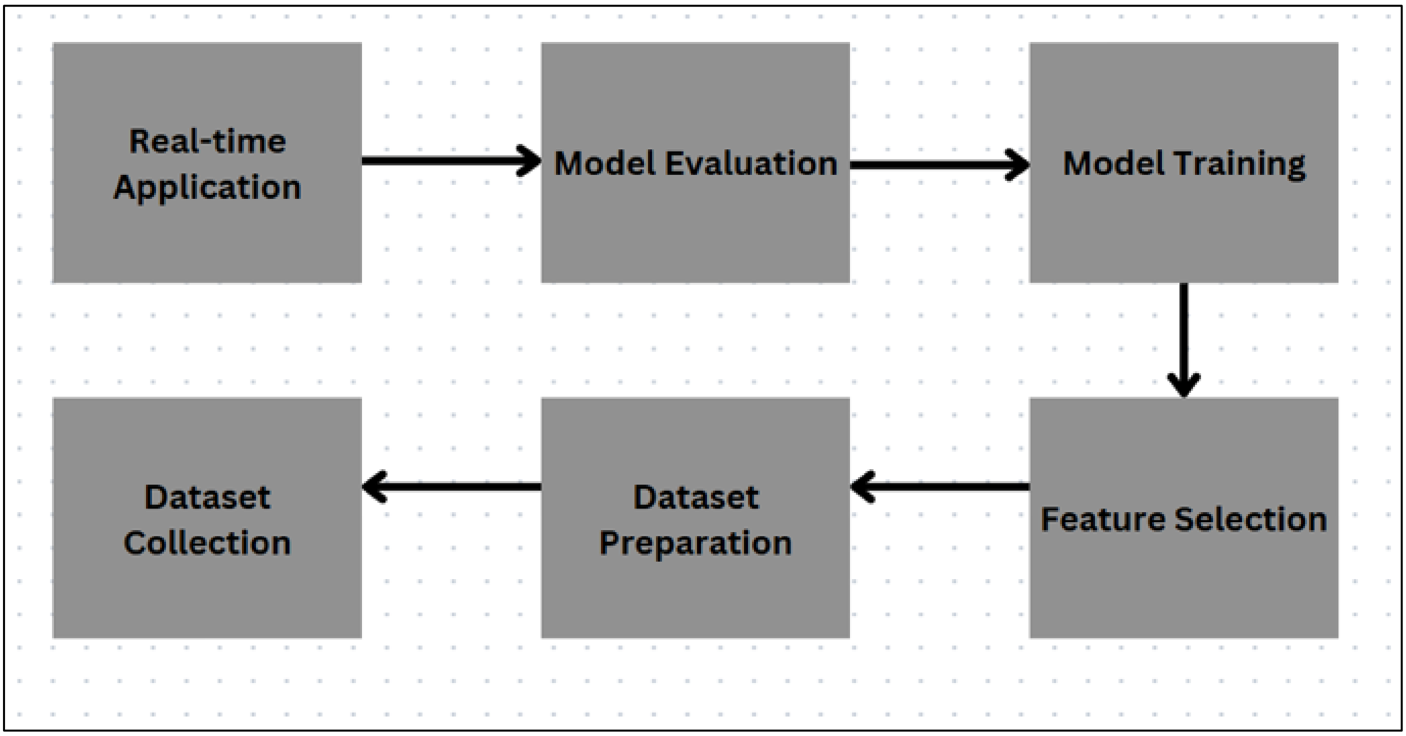

Fig 2 Workflow of the Proposed Methodology for Integrating Generative AI into Real-Time Systems.

IV. RESULTS

➢ *The Evaluation Metrics Included:*

- **Latency**: Time taken to process inputs and produce outputs.
- **Reliability**: Percentage of correct outputs in real-time scenarios.
- **Interpretability**: Scale from 1 to 5 based on expert assessments of model explanations.

Table 2 Compares Our Approach with Traditional Generative Models

Metric	Baseline Model	Proposed Model	Improvement (%)
Latency (ms)	150	120	20
Reliability (%)	70	85	15
Interpretability (1-5)	2.5	4.0	60

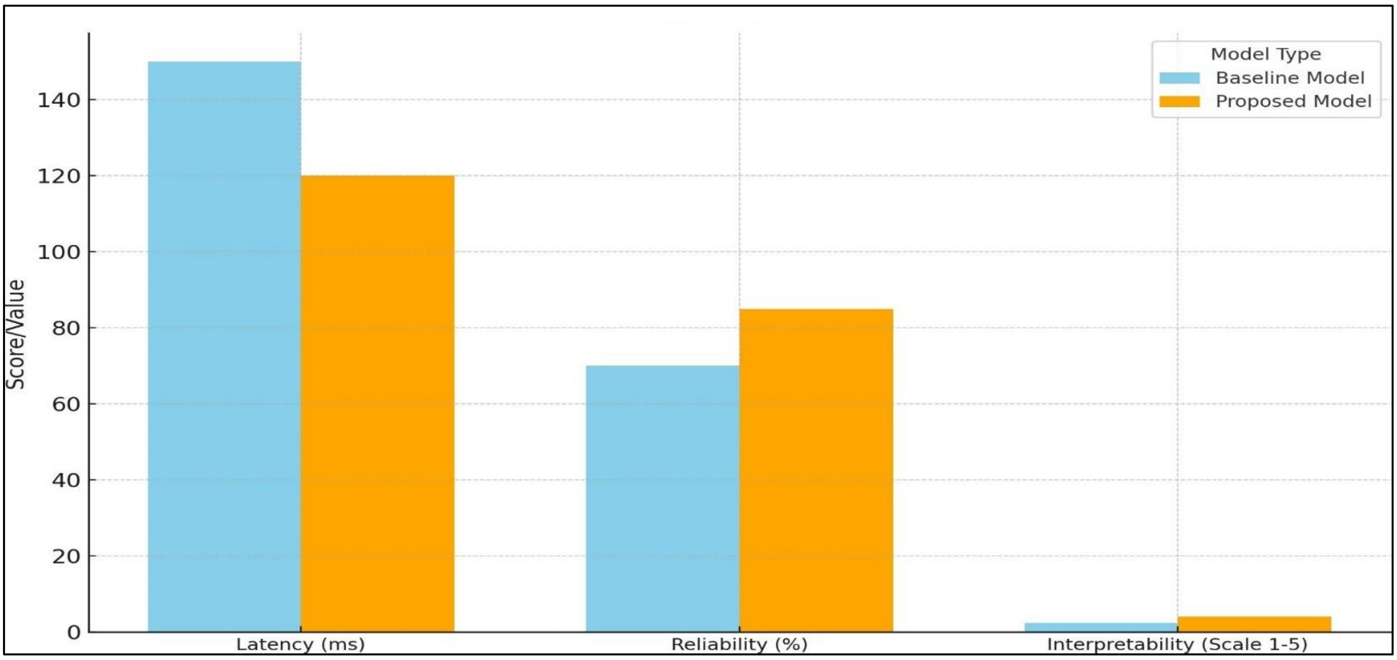

Fig 3 Comparison of Baseline and Proposed Models Across Metrics

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jun935

This chart compares the performance of baseline and proposed models on key metrics such as latency, reliability, and interpretability, highlighting the improvements achieved by the proposed methodology.

➢ *Discussion*

The results demonstrate a significant reduction in latency and improvement in reliability and interpretability, making our approach more suitable for real-time applications.

V. CONCLUSION AND FUTURE WORK

This study highlights the limitations of generative AI in real-time decision-making and proposes a hybrid methodology to address these challenges. Key findings include:

- Generative AI models exhibit high computational latency, making them less suitable for time-sensitive applications.
- Hybrid models improve both reliability and interpretability, addressing core limitations of traditional generative systems.

➢ *Future Work will Focus on:*

- Extending the methodology to additional domains such as disaster management and defense.
- Exploring advanced architectures like reinforcement learning-based generative models.
- Developing ethical frameworks for the responsible deployment of generative AI in real-time systems.

By addressing these areas, we aim to bridge the gap between generative AI’s potential and its practical applications in critical environments.

REFERENCES

[1]. Lu, Y., Shen, M., Wang, H., et al. (2023). Machine learning for synthetic data generation: A review.

[2]. Weidinger, L., Mellor, J., Rauh, M., et al. (2021). Ethical and social risks of harm from language models.

[3]. Hernandez, M., Epelde, G., Alberdi, A., et al. (2022). Synthetic data generation for tabular health records: A systematic review.

[4]. Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods, and GANs.

[5]. Raghunathan, T. E. (2021). Synthetic data. *Annual Review of Statistics and Its Application*, 8(1), 129-140.

[6]. Sutton, R. S., & Barto, A. G. (1999). Reinforcement learning: An introduction.

[7]. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias.

[8]. Thomas, G. (2023). Synthetic data generation: Building trust by ensuring privacy and quality.

[9]. Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting.

[10]. Goodfellow, I., et al. (2014). Generative adversarial networks. *Advances in Neural Information Processing Systems.*

[11]. Amodei, D., et al. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565.*

[12]. Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems.*