INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

Generative AI and Privacy-Preserving Big Data Analytic in Cloud Environments with AI Agents

Akanksha Shukla, Dr. Rohit Kumar

Haridwar University, Roorkee

DOI: https://doi.org/10.51583/IJLTEMAS.2025.1411000124

Received: 08 December 2025; Accepted: 15 December 2025; Published: 25 December 2025

ABSTRACT

While generative artificial intelligence (GenAI) technologies are revolutionising content production, they also pose serious privacy and data security issues. The potential of privacy violations, biases, and cyberattacks rises as these models process large datasets, many of which contain sensitive or private data. These issues are examined in this book, especially in important fields like cybersecurity, healthcare, and finance. The potential for GenAI models to reproduce or infer sensitive data from training datasets is a major problem that raises ethical and intellectual property issues. Data protection techniques like encryption, tokenisation, and anonymisation are crucial to reducing these dangers. This study assesses the efficacy of these techniques by looking at how they affect the functional performance and privacy risk reduction of GenAI systems. It evaluates the impact of tokenisation and anonymisation on a state-of-the-art large language model (LLM) through experimental analysis. Empirical results offer insights into the trade-offs between protecting model performance and data privacy using open-source tools such as Microsoft Presidio. The goal of the research is to help create safe and morally sound GenAI applications, making sure that advancements in AI are in line with data security guidelines while preserving accuracy and efficiency in practical applications.

Keywords: Privacy-Preserving, Big Data,AI-Driven,Cloud and Techniques.

INTRODUCTION

GenAI has exploded in the previous two years, with exponential growth. It can be used to create realistic, imaginative text, graphics, and other types of data, including music. This suggests that there will be substantial uses for this advancement in GenAI across a range of sectors, such as marketing, healthcare, finance, and entertainment. But such quick expansion brings up important issues about information security and data privacy. First, in order to train effectively, generative AI requires a vast amount of data [1]. Additionally, when sensitive personal data is included in the data, this naturally raises the danger of leakage. Any contact with the GenAI system could add to a dataset that contains personally identifiable information; therefore, if appropriate anonymisation or data protection is not in place, that dataset would become vulnerable. The lack of openness in data collecting, storage, and utilisation continues to be one of the primary issues. Most of the time, end users are unaware of the maximum amount of data that can be used, particularly when that data is shared or processed by an outside service provider. Because these outside contractors could not adhere to the same stringent privacy regulations as internal services and might utilise the data more frequently, outsourcing can raise the security risk. User data, for instance, could be utilised for reasons other than data. Apart from raising concerns about who owns and controls personal data when it is sent to these platforms, this seriously infringes on a person's right to privacy. Unintentionally disclosing intellectual property is the other significant risk [2].

Sensitive corporate data may be accidentally accessed or shared as a result of the model's training process absorbing proprietary or secret information that is supplied to it by individuals and businesses. Because so many GenAI platforms store data in cloud environments, there is an increased chance that private data could be stolen, intercepted, or used in other ways by cybercriminals. The fact that AI models are black-boxes makes it difficult to understand or even track decisions or internal data processing, which increases these dangers [4].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

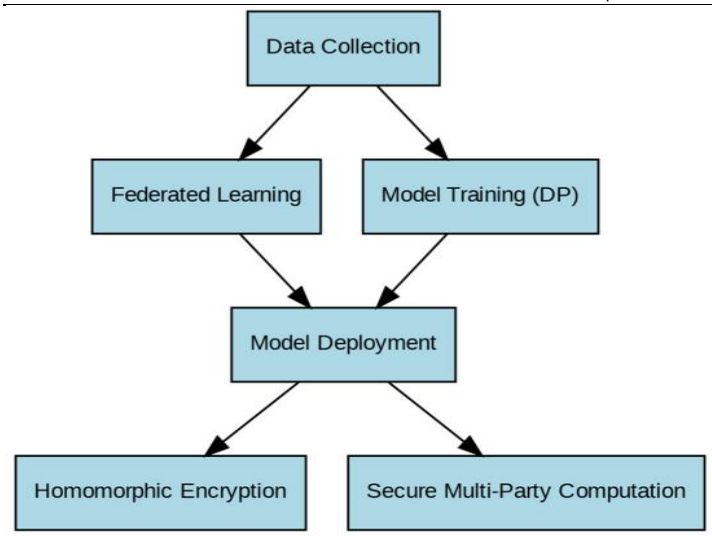

Figure 1. Proposed Privacy-Privacy workflow in AI [3]

In addition to creating accountability issues, this opaqueness makes it extremely difficult to ensure adherence to privacy rules and regulations, such as the GDPR. Given these grave worries, it goes without saying that GenAI platforms will seriously jeopardise user privacy, data security, and intellectual property issues if they are not protected, with far-reaching consequences for both people and organisations.The following are some of the report's key goals: Examining the threat landscape related to GenAI with an emphasis on the dangers to intellectual property, privacy, and security.

1. To investigate several approaches to risk reduction and data security that can be used for the responsible development and application of GenAI.

2. It explains a particular data tokenisation project, including its implementation, outcomes, and constraints, in order to thoroughly examine how data tokenisation is one potential way to improve data privacy in the context of GenAI.

The rest of the following section are, in Section II literature survey related this research has been explained. In Section III, proposed methodology has been elaborated. In Section IV, results has been showed and discussed with conventional work. Finally, in Section V, conclude the proposed work.

LITERATURE SURVEY

In order to improve Cyber-Physical Systems (CPS) in the medical domain, [5] provides a comprehensive analysis of deep learning in conjunction with image categorisation. The authors highlight the vital role that deep learning plays in picture classification while deftly navigating the complexities of secure medical environments. By addressing the intersection of technology and healthcare, the research contributes to the evolving field of safe systems in an important area.

In [6], examine the potential for future wireless communication with a focus on 6G. Their study demonstrates the advantages of integrating blockchain and artificial intelligence, offering insights into how these two technologies could cooperate to enhance security and privacy in the emerging 6G landscape. By providing both a theoretical foundation and real-world applications, this study significantly adds to the conversation about the security issues with growing wireless networks.

In [7] offers a comprehensive examination of critical security for Internet of Things (IoT) networks, encompassing blockchain, AI, and conventional methods. The paper is a priceless resource for academics and business people alike by analysing the numerous security concerns associated with IoT. To broaden the discussion and offer a thorough road map for understanding and addressing security concerns in the quickly evolving IoT network environment, a number of security paradigms are included.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

An in-depth analysis of the convergence of distributed ledger technology (DLT) with artificial intelligence (AI) can be found in [8]. Their paper provides a current evaluation of the state of this convergence, highlighting its primary challenges and outlining potential directions. This survey is a helpful resource for understanding the evolving landscape at the nexus of DLT and AI. It was published in IEEE Access. Both academics and business professionals can benefit from its insights.

Reliable and privacy-preserving federated deep learning is emphasised in [9], which contributes to the corpus of work in the context of the Industrial Internet of Things (IIoT). Their work addresses the critical need for robust security and privacy protections in IIoT designs. In addition to including federated deep learning, the proposed method prioritises privacy and trust preservation, taking into account the particular requirements of the industrial context.

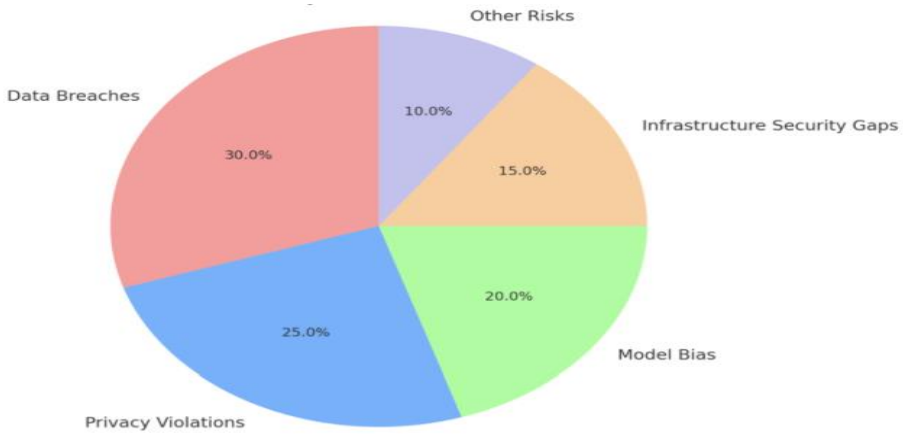

Figure 2. Security Risk Distribution on Generative AI Platforms [10]

Table 1. Comparative Analysis related to Cloud based AI gent techniques for privacy and Security [11]

References	Year	Title	Focus	Contribution
[12]	2024	Privacy and Security Implications of Cloud-Based AI Services: A Survey	Cloud-Based AI Services	Provides a comprehensive survey of privacy and security risks in cloud-based AI services, introducing a taxonomy to categorize these risks and discussing defenses for both model providers and consumers.
[13]	2023	Towards Confidential Computing: A Secure Cloud Architecture for Big Data Analytics and AI	Secure Cloud Architecture	Proposes a secure cloud architecture that ensures data, logic, and computation remain secure during transit, use, and at rest, addressing concerns in biomedical research and other sensitive fields.
[14]	2023	An Overview of AI and Blockchain Integration for Privacy-Preserving	AI and Blockchain Integration	Explores the integration of AI and blockchain technologies to enhance privacy, discussing applications in data encryption, de-identification, and multi-tier distributed ledgers.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

[15]	2024	Privacy-Preserving Data in IoT-based Cloud Systems: A Comprehensive Survey with AI Integration	IoT-based Cloud Systems	Provides a comprehensive survey of privacy issues in IoT and cloud systems, highlighting the role of AI in dynamic anonymization and secure data sharing.
[16]	2024	Generative AI Model Privacy: A Survey	Generative AI Privacy	Surveys privacy concerns specific to generative AI models, discussing potential risks and mitigation strategies to protect sensitive information.
[17]	2023	Systematic Survey: Secure and Privacy-Preserving Big Data Analytics in Cloud	Big Data Analytics in Cloud	Analyzes various security and privacy solutions for big data analytics in cloud environments, focusing on secure access control, data storage, and private learning.
[18]	2024	Security and Privacy on Generative Data in AIGC: A Survey	Generative Data Security	Discusses security and privacy challenges associated with generative data in AI-generated content (AIGC), providing insights into current solutions and future directions.

These models serve as the foundation for training machine learning models, enhancing datasets, and protecting data privacy since they tackle the issues of data imbalance, privacy, and scarcity.

PROPOSED METHODOLOGY

This suggested approach evaluates and selects a single anonymisation technique to be used for extensive LLMs and assesses how effectively it performs in a variety of scenarios involving different input types and personally identifiable information (PIIs). An detailed literature review and exploratory research that concentrate on the current state of anonymisation techniques and tools are the first steps in the process. The top-ranked open-source anonymizers are determined by consulting a variety of sources, including technical papers, industry publications, and scholarly studies.

The changes of community's recognition, integration potential, generative AI support, and tool creators' trustworthiness. The strengths and limitations of each of these tools will be compared with this selection criterion. The instrument with the best ratio of strengths to shortcomings will be chosen for additional examination.

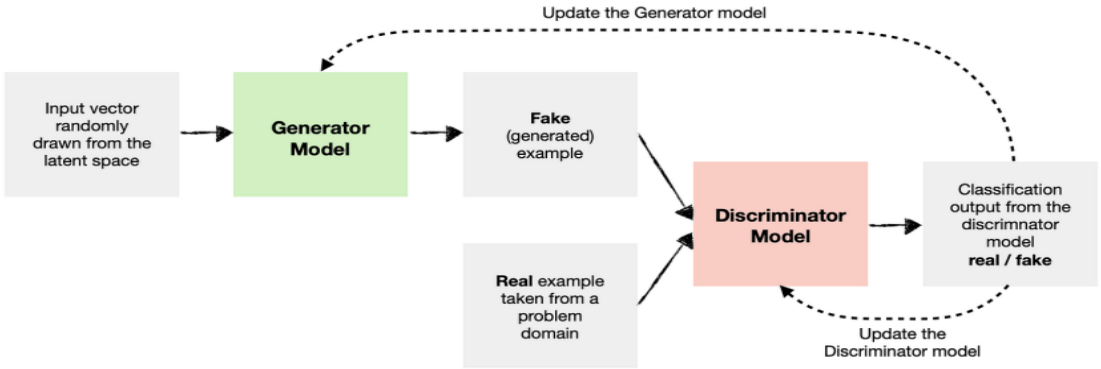

Figure 3. Proposed GAN workflow

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

After selecting an anonymisation tool, the next step is to establish an experimental scenario that will be used to evaluate the tool's functionality. This would entail defining the kinds of data that need to be anonymised and creating the experiment's architecture using a range of PII and additional input formats. The LangChain framework or a comparable tool will be used to achieve anonymisation, and performance measurements will include processing speed, anonymisation correctness, and effect on LLM comprehension.

The ROUGE/BLEU [19] anonymisation quality ratings, LLM-based evaluations of the model's comprehension for anonymised versus non-anonymized input, and human evaluations to offer qualitative insights into the tool's efficacy will be the evaluation metrics. These findings will be examined in a comparative analysis of the anonymisation tool's diverse performances across multiple scenarios, highlighting its versatility in handling different input types and PII types.

The last step will include an evaluation of the effectiveness of the selected anonymisation tool as well as recommendations for how to make it better or different. Additionally, it will discuss potential directions for future study as well as practical application based on findings. A more structured method of choosing a system for anonymisation ensures that enough data on how it operates in specific contexts is obtained.

Transparency and Data Minimisation

By collecting and processing just the necessary data, data minimisation plays a crucial role in lowering the likelihood of a breach. At the very least, handling less amounts of data implies a lower chance of a massive data breach, which is very troublesome when it comes to sensitive or personally identifiable information. Conversely, transparency refers to informing consumers about the potential benefits of the information gathered about them, so they are fully aware of how AI processes data. The degree of openness will foster trust, which will motivate users to provide informed consent in order to fulfil their ethical duties [20].

Protection of Data

The foundation of data protection will be adaptive character AI security solutions, which will evolve over time in tandem with new threats and technical advancements. As a result, the process of identifying and screening enabling technologies must be especially cautious when it comes to the tools, libraries, and frameworks that are crucial to the development and use of AI. With a few notable exceptions, open-source technologies have become more prevalent in the development of AI systems [21].

Verification

Following the screening of the enabling technologies, application and infrastructure security will be the main focus. The majority of AI systems function in extremely intricate ecological settings, where infrastructure flaws potentially jeopardise their reliability. Effective security methods, such as MFA, data encryption, and RBAC, will be necessary to safeguard the AI systems themselves [22].

Ongoing Observation

Organisations must also keep an eye on AI-specific risks, which are distinct from conventional cybersecurity issues, in addition to safeguarding their infrastructure. Adversarial input, for example, is a sort of attack that targets AI systems exclusively. By making small changes to the input data, malicious actors might alter the appearance of the AI output. Data poisoning, in which training datasets are tainted to produce inaccurate AI predictions and behaviours, is the other major hazard [23].

Handling Vulnerabilities

The institutionalisation of policies pertaining to vulnerability management will be the other key focus of AI security. This relates to routine risk assessments and vulnerability scanning for potential attempts to take advantage of system flaws. Here is where a company may guarantee weaknesses by continuing to take a proactive stance in identifying threats. These are vulnerabilities that can be exploited before they are exploited

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

if they are found and repaired. In addition, prompt incident response may be crucial to minimising security vulnerabilities as soon as feasible [25].

As a result, the following strategies are suggested for the fundamentals of GenAI: risk management, transparency, security, and data minimisation. Every company should have a comprehensive AI security plan that anticipates how technologies will always make it stronger and more secure, both at the data and infrastructure levels, which are the foundation of AI systems. In this regard, ongoing vulnerability monitoring and management, along with ethical standards, will help the company significantly lower the risks connected with AI while creating more safe, trustworthy, and socially responsible systems.

RESULTS AND DISCUSSION

Randomisation, the process of adding noise to data, is frequently done via a probability distribution. Randomisation is applied in sentiment analysis and surveys. Randomisation does not require knowledge of other entries in the data. It can be applied to the stages of data collecting and preparation. There is no anonymization-related overhead with randomisation. However, because of time complexity and data utility, randomisation is not practical for large datasets, as demonstrated by our experiment, which is described below [26].

More Mappers and Reducers were used as the amount of data increased. There was a considerable difference between the results before and after randomisation. Randomisation has little effect on a small number of outlier records, which are vulnerable to adversarial assault. When it comes to attribute sharing, randomisation might not be the best way to protect privacy because data utility is not valued when privacy is sacrificed.

Table 3: After utilising age and zip code anonymization [27]

Sr.no.	Zip	Disease	Age
1	262	Cardiac problem	1
2	362	Cardiac problem	3
3	414	Cardiac problem	2
4	536	Skin allergy	>50
5	458	Cardiac problem	>50

Table 3: T closeness privacy protection method [28]

Sr.no.	Age Record	Zip Record	Medical Record	Salary Record
1	5	263	Flu	5463
2	10	363	Cardiac problem	6352
3	15	424	Skin allergy	7246
4	>50	537	Cancer	8157
5	>60	459	Cardiac problem	9463
6	>70	378	Skin allergy	4681

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

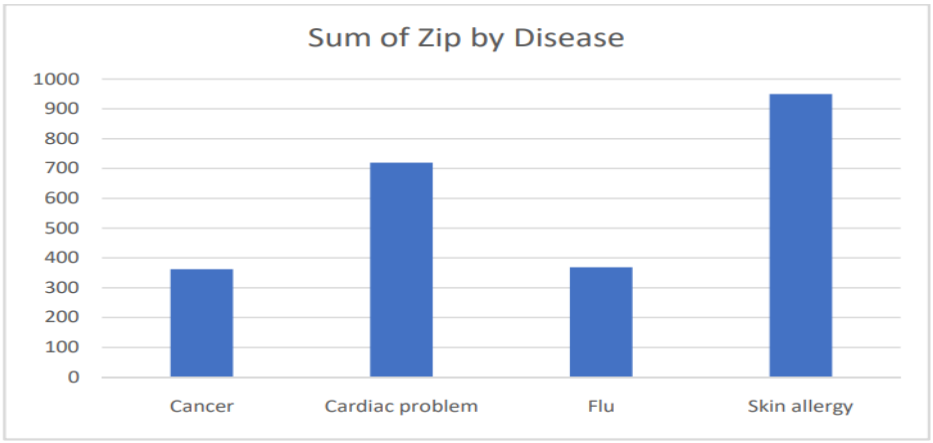

Figure 4. T closeness privacy protection method

The dataset provided contains information on a wide range of individuals who are identified by their serial numbers (Sno) and reside in different zip codes (Zip). Age, declared income, and any related medical conditions are used to identify each individual. The wage data is one significant feature of this dataset that adds a fresh viewpoint to the investigation. A range of ages are represented by patients with serial numbers 1 through 6, with an emphasis on those who are over 50 (referred to as ">50"). Interestingly, individuals in this age group have been diagnosed with a variety of diseases, including cancer, heart problems, and the flu.

The dataset illustrates the potential relationship between the prevalence of several illnesses, age, and income. Patients with heart problems report salaries ranging from 5463 to 9463, indicating a range of income levels within this health category. Similarly, there is a variety in the reported salaries of those with skin allergies or cancer diagnoses. This dataset offers an opportunity to investigate the relationships among age, socioeconomic position, and the likelihood of specific health issues. Healthcare professionals and policymakers may find it crucial to understand these links in order to develop targeted interventions and healthcare policies that take into account the complex nature of health disparities within this group.

CONCLUSION

Conclusively, this proposed effort highlights the complex interplay between the urgent need for strong data protection measures and the transformational promise of generative AI. The study illustrates the potential and difficulties presented by this quickly developing technology by following the development of GenAI throughout time and examining its present capabilities, constraints, and security threats. Although GenAI has enormous advantages in terms of automation and content creation, it also comes with serious concerns, such as false information, privacy violations, and cyberthreats. Therefore, in order to guarantee ethical use and minimise potential harm, GenAI must be developed and regulated responsibly. In order to create a future where GenAI can flourish while maintaining data security and privacy, more research and proactive governance will be necessary.

REFERENCES

1. Chen Y, et al. Blockchain-based medical records secure storage and medical service framework. J Med Syst. 2019;43:1–9.
2. Mayer AH, da Costa CA, Righi RDR. Electronic health records in a blockchain: a systematic review. Health Inf J. 2020;26(2):1273–88.
3. Ghadi YY, et al. The role of blockchain to secure internet of medical things. Sci Rep. 2024;14(1):18422.
4. Ghadi YY, Shah SFA, Mazhar T, Shahzad T, Ouahada K, Hamam H. Enhancing patient healthcare with mobile edge computing and 5G: challenges and solutions for secure online health tools. J Cloud Comput. 2024;13(1):93.
5. Saranya R, Murugan A. A systematic review of enabling blockchain in healthcare system: Analysis, current status, challenges and future direction. Mater Today Proc. 2023;80:3010–5.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

6. Andrew J, et al. Blockchain for healthcare systems: architecture, security challenges, trends and future directions. J Netw Comput Appl. 2023;215: 103633.
7. Sujan MA, Looking at the safety of ai from a systems perspective: two healthcare examples, in safety in the digital age: sociotechnical perspectives on algorithms and machine learning. 2023, Springer Nature Switzerland Cham. p. 79–90. .
8. Wehkamp K, Krawczak M, Schreiber S. The quality and utility of artifcial intelligence in patient care. Dtsch Arztebl Int. 2023;120(27–28):463.
9. Mondal H, Mondal S, Singla RK, Artifcial Intelligence in Rural Health in Developing Countries, in Artifcial Intelligence in Medical Virology. 2023, Springer. p. 37-48
10. Zuhair V, et al. Exploring the impact of artifcial intelligence on global health and enhancing healthcare in developing nations. J Primary Care Commun Health. 2024;15:21501319241245850.
11. Poalelungi DG, et al. Advancing patient care: how artifcial intelligence is transforming healthcare. J Personal Med. 2023;13(8):1214.
12. Taherdoost H, Machine learning algorithms: features and applications, in Encyclopedia of Data Science and Machine Learning. 2023, IGI Global. 938–960.
13. Mirjalili S, Gandomi AH. Comprehensive metaheuristics: algorithms and applications. Amsterdam: Elsevier; 2023.
14. Worden K, et al. Artifcial neural networks, in machine learning in modeling and simulation: methods and applications. Berlin: Springer; 2023. p. 85–119.
15. Kasneci E, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning Individual Dif. 2023;103: 102274.
16. Chang Y, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol. 2024;15(3):1–45.
17. Chaka C. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of fve AI content detection tools. J Appl Learning Teaching. 2023;6(2):12.
18. Wu X, Duan R, Ni J. Unveiling security, privacy, and ethical concerns of ChatGPT. J Inf Intell. 2024;2(2):102–15.
19. Ma S, et al. “Are you sure?” Understanding the efects of human self-confdence calibration in ai-assisted decision making. in proceedings of the CHI conference on human factors in computing systems. 2024.
20. Masood I, et al. A blockchain-based system for patient data privacy and security. Multimedia Tools Applications. 2024;83(21):60443–67.
21. Salah M, Al Halbusi H, Abdelfattah F, May the force of text data analysis be with you: Unleashing the power of generative AI for social psychology research. Comput Hum Behav Artif Hum, 2023: 100006.
22. Al-Hawawreh M, Aljuhani A, Jararweh Y. Chatgpt for cybersecurity: practical applications, challenges, and future directions. Clust Comput. 2023;26(6):3421–36.
23. Yao Y, et al., A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confdence Computing, 2024: p. 100211.
24. Singh K, Chatterjee S, Mariani M, Applications of generative AI and future organizational performance: The mediating role of explorative and exploitative innovation and the moderating role of ethical dilemmas and environmental dynamism. 2024. 133: 103021.
25. Tokayev K-J. Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns. Int J Soc Anal. 2023;8(9):17–33.
26. Usman M, Qamar U. Secure electronic medical records storage and sharing using blockchain technology. Proc Comput Sci. 2020;174:321–7.
27. Vanathi J, G. SriPradha. BreakTheChain: A Proposed AI powered Mobile Application Framework to handle COVID-19 Pandemic. Alochana Chakra Journal. 9: 108–114.
28. Yan L, et al. Practical and ethical challenges of large language models in education: A systematic scoping review. Br J Edu Technol. 2024;55(1):90–112.