International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 3, 7-12, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I3P102

Original Article

Ethical Prompt Design for Health Equity: Preventing Hallucination and Addressing Bias in AI Diagnoses

Adya Mishra
(Independent Researcher) Virginia, USA.

Received On: 11/05/2025 ***Revised On: 30/05/2025*** ***Accepted On: 19/06/2025*** ***Published On: 04/07/2025***

Abstract - *The use of large language models (LLMs) in healthcare is transforming how clinicians access information, generate insights, and support patient care decisions. These AI systems hold tremendous promise offering the ability to summarize complex clinical notes, suggest differential diagnoses, and assist in managing vast amounts of medical data. However, alongside these benefits come serious ethical and practical concerns. If not carefully guided, LLMs can generate hallucinations confident sounding yet entirely fabricated information which can mislead clinical decision making. Moreover, these models often inherit and perpetuate biases from the data they are trained on, potentially exacerbating disparities in care for already marginalized or underrepresented patient populations. This paper explores how ethical prompt design the careful crafting of instructions and context given to LLMs can help address these risks. We focus on two key challenges: reducing hallucinated AI responses in medical contexts and minimizing bias that could negatively impact care quality for certain demographic groups. To tackle these issues, we propose a human in the loop framework, where clinicians and domain experts actively shape and evaluate prompts to ensure the output is safe, inclusive, and grounded in evidence based medicine.*

Keywords - *Prompt Engineering, Health Equity, LLMs, AI Hallucination, Bias Mitigation, Medical NLP, Ethical AI, Human in the Loop.*

1. Introduction

Artificial Intelligence (AI) has evolved from its early theoretical foundations into a transformative force across industries, with particularly profound implications for healthcare. What began as an exploration into mimicking human cognition through neural networks and computational logic has matured into the development of powerful systems capable of performing tasks once limited to human intelligence. A major leap forward came with the advent of large language models (LLMs) and generative AI technologies, which have demonstrated remarkable proficiency in understanding and generating natural language. These models, built on advanced transformer architectures, can produce human-like text, carry on coherent conversations, and assist with complex reasoning without the need for manual programming of specific tasks [1-3]. In the context of healthcare, this technological shift holds enormous potential. Generative AI can rapidly process and interpret extensive medical datasets, providing real time assistance to clinicians and care teams. It is now being used to generate clinical documentation, draft patient instructions, and even offer preliminary diagnostic insights ultimately

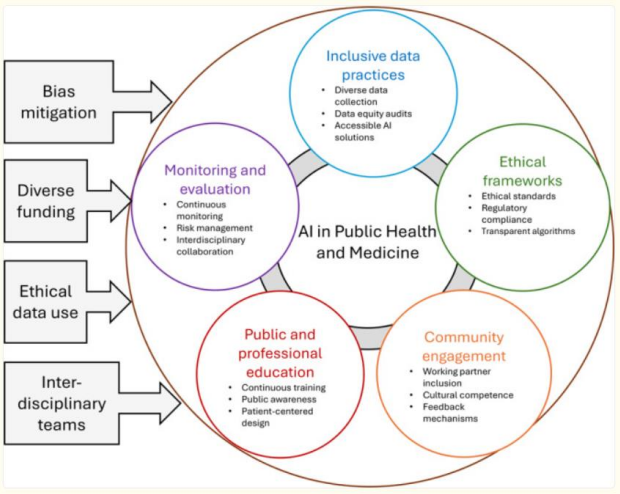

Figure 1: Multifaceted approach for ethical and equitable implementation of artificial intelligence (AI) in public health and medicine [2].

Streamlining workflows and enhancing both efficiency and quality of care. However, these benefits come with critical challenges. As AI systems become more integrated into clinical settings, concerns around misinformation (hallucinations), embedded bias, and equitable access become increasingly urgent. If LLMs are prompted without ethical foresight, they may generate inaccurate or biased responses that reinforce disparities in care especially for underserved or marginalized patient populations. Large language models (LLMs) like GPT 4, PaLM, and LLaMA have shown unprecedented capability in understanding and generating human like text. In the medical domain, their use ranges from generating patient summaries to supporting diagnostic decisions. However, their deployment comes with challenges, especially regarding hallucinated medical content and embedded bias. These risks can perpetuate or even amplify health inequities, particularly affecting underrepresented populations [4].

Ethical prompt design is emerging as a vital solution space for aligning LLM behaviour with healthcare values. This paper addresses how prompt engineering when done with ethical foresight can reduce diagnostic hallucinations and foster fairness in AI generated medical insights [5]. AI hallucinations undermine the credibility of AI systems, especially in critical sectors such as healthcare, where a misdiagnosis based on hallucinated information could cost lives, or in legal systems, where false citations could derail judicial processes. As AI becomes embedded in decision making and information dissemination, addressing hallucinations becomes a crucial step in ensuring ethical, reliable, and sustainable AI deployment. This paper focuses on the emerging field of ethical prompt design to mitigate these risks. By crafting prompts that are inclusive, grounded in evidence, and shaped through human oversight, we can guide LLMs toward safer, fairer, and more trustworthy healthcare applications. Through a proposed framework, real world use cases, and evaluation metrics, we aim to provide both theoretical insight and practical guidance for the responsible deployment of LLMs in medical environments [6].

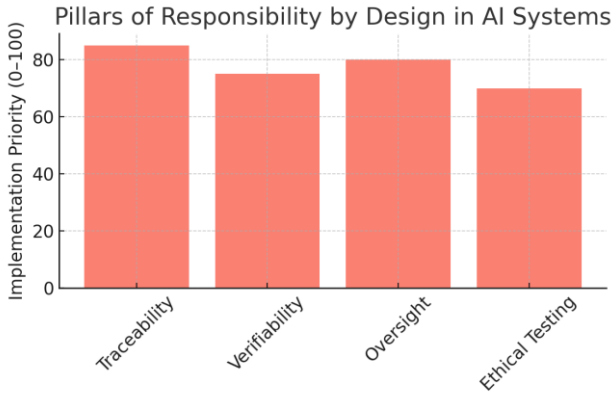

Figure 2: Key pillars prioritized in the Responsibility by Design framework

2. Understanding Prompt Engineering

Prompt engineering is the art and science of crafting clear and effective instructions to guide AI models, especially large language models (LLMs) toward generating accurate and useful responses. In family medicine, where decisions affect real lives, prompt engineering is critical to ensure AI supports clinical workflows, communication, and patient care safely and reliably [7].

2.1. Why Prompt Engineering Matters in Healthcare

Unlike generic AI use cases, healthcare prompts must be context aware, medically sound, and patient centered. A well-designed prompt can:

- Generate accurate clinical summaries.
- Help doctors explain complex conditions in simple terms.
- Suggest relevant diagnoses or treatment options.
- Support documentation and administrative tasks.

Example 1: A basic prompt:
"What causes chest pain?"

Example 2: Refined prompt:
"List likely causes of chest pain in a 55-year-old female with diabetes and shortness of breath. Be concise and evidence based."

2.2. Best Practices for Prompt Engineering in Healthcare

- Use Medical Guidelines and Domain Knowledge: Prompts should align with current clinical guidelines and specialty specific knowledge (e.g., cardiology, primary care). This ensures that AI responses are grounded in best practices.
- Iterate and Validate Prompts: Prompt engineering is not one and done. Feedback from real users doctors, nurses, and admin staff helps refine prompts so AI tools work better in real world settings.
- Minimize Bias and Ensure Ethical Use: Prompts should be inclusive and designed to avoid biases around gender, race, or socioeconomic status. Collaborating with ethicists, clinicians, and diverse patient voices helps promote fairness and equity.

2.3. Real World Applications in Healthcare

- Patient Communication: AI can explain conditions like hypertension or arthritis in plain language to support better patient understanding and treatment adherence.
- Documentation Support: AI can auto draft SOAP notes, discharge summaries, or referral letters saving clinicians time and reducing burnout.
- Medical Education: Prompt engineered scenarios can train professionals on case based learning.
- Personalized Care: By including patient specific details (e.g., family history, lifestyle, genetics), prompts can generate personalized screening or treatment recommendations.
- Shared Decision Making: AI generated risk assessments or side effect comparisons help

facilitate collaborative conversations between doctors and patients.

3. Responsibility by Design: AI with Built in Accountability

As artificial intelligence continues to shape critical sectors like healthcare, law, education, and public safety, the risks posed by hallucinations AI generated content that's factually wrong or entirely made up are no longer just technical errors. They can lead to real legal consequences, harm reputations, and compromise fundamental rights. That's why we need more than after the fact fixes. We need to embed responsibility into the very fabric of how AI systems are designed, developed, and deployed [8-9]. This proactive approach is known as Responsibility by Design (or "AI by Design"). It means that ethical principles, legal safeguards, and accountability checks should be part of the AI development process from day one did not bolt on after something goes wrong. When it comes to preventing hallucinations, this includes building features like:

- Output traceability, so it's clear how a response was generated.
- Fact checking systems, especially for applications in journalism, law, or medicine.
- Robust testing, to uncover and fix weaknesses before the AI is used in the real world.

This isn't just the best practice, it's increasingly becoming law. New regulations like the EU AI Act require high risk AI systems to meet standards for transparency, verifiability, and human oversight. Global frameworks like ISO/IEC 42001 and the NIST AI Risk Management Framework also emphasize the need for built-in compliance, ethical auditing, and continuous risk assessment. Beyond legal compliance, Responsibility by Design offers another major benefit: trust. When people know that an AI system has been carefully vetted and built with safeguards in place, they're more likely to trust its output, especially when it's being used in sensitive contexts like diagnosing illnesses or assessing legal claims. Without that trust, even the most advanced AI systems risk being rejected or misused.

Finally, there's a practical angle. If something does go wrong, organizations can reduce their legal exposure by showing that they took all reasonable precautions through rigorous testing, ethical review, and responsible deployment practices. In this way, designing with responsibility in mind becomes both a moral and strategic decision [10].

3.1. Managing AI: The Role of Technical Solutions

AI hallucinations are complex. They aren't just software bugs, they're often symptoms of deeper issues like biased training data, flawed algorithms, or ambiguous input. And just like human decision makers can be inconsistent or misled by noise, AI systems can also "drift" into error when exposed to poor data or conflicting information. These moments of "algorithmic noise" are especially dangerous when decisions impact people's health, freedom, or financial stability [11].

To combat this, developers are exploring advanced detection and mitigation techniques, including:

- Predictive modeling to identify when an AI is likely to hallucinate.
- Adversarial testing to simulate real world pressures and see how the model responds.
- Noise auditing frameworks, inspired by behavioral science, to reduce variability and improve output consistency.

By treating hallucinations as both a technical and ethical challenge, we can take a more holistic approach to one that blends engineering with accountability, and innovation with human values.

4. Ethical Considerations in the Use of Artificial Intelligence

4.1. Ethical Principles in AI Design and Use

As artificial intelligence becomes increasingly integrated into healthcare, it's essential to ground its development and deployment in strong ethical principles. These principles long recognized in clinical practice help ensure that AI tools do more good than harm and serve all patients fairly. At the core are values such as beneficence (doing good) and nonmaleficence (avoiding harm). When applied to AI, these principles emphasize the importance of creating systems that genuinely support patient well-being and avoid introducing risks through errors, bias, or misuse [12]. Equally important is respecting patient autonomy. This means that patients should understand when and how AI is being used in their care, and that their consent should be informed and voluntary. In addition, the principles of fairness and justice remind us to make sure AI technologies promote health equity does not deepen existing disparities. AI should be a tool that helps all communities, especially those who have historically faced barriers to care, access better health outcomes [13].

4.2. Protecting Privacy and confidentiality

AI's ability to process vast amounts of data is both its strength and its challenge. With access to sensitive health records and personal information, these systems must be developed with privacy and confidentiality at the forefront. Protecting this information means implementing strong security measures to guard against unauthorized access, breaches, or misuse. It also involves improving how we handle informed consent. Consent forms should be easy to understand, culturally appropriate, and available in multiple languages to ensure that all patients especially those with limited English proficiency truly understand how their data will be used. Where possible, only the minimum data required for a given AI application should be collected, limiting exposure and reducing risk [14].

4.3. AI in Clinical and Public Health Decision making

As AI plays a growing role in clinical decision making and public health planning, it's critical to define how human oversight is maintained. AI should support, not replace, the judgment of healthcare professionals. Guidelines must be established to clarify when and how clinicians should rely on AI suggestions and when human experience and empathy

must take precedence. Another key concern is transparency. Healthcare providers and patients need to understand how an AI system arrives at its recommendations. If AI tools are to be trusted, they must offer explanations that are clear and useful in real world settings. Accountability is also essential. When AI is involved in patient care, everyone developers, clinicians, institutions share responsibility. Developers must create safe, accurate, and reliable systems. Providers must be trained to interpret AI results responsibly and make final decisions that blend data with personal care. Institutions, meanwhile, must ensure that oversight mechanisms are in place to monitor how these systems are used and whether they align with ethical and legal standards [15-16]. Public health professionals also have a role to play. They can leverage AI to analyse data, predict health trends, and respond to emerging threats. But in doing so, they must ensure that the insights gained support the well-being of all communities, not just a privileged few.

4.4. The Importance of Community Engagement

One of the most effective ways to build ethical and equitable AI systems is by involving the communities they are meant to serve. By engaging patients, caregivers, advocates, and underrepresented groups early in the development process, we can build tools that are more relevant, effective, and trustworthy [17]. Community input helps developers understand real world challenges and social contexts that might not be apparent through data alone. When people feel that their voices have been heard and their needs reflected in technology, they are more likely to accept and benefit from it. This participatory approach also reduces the risk of blind spots that could lead to biased or ineffective AI applications [18].

5. Healthcare: Why AI Hallucinations Are Especially Dangerous in Medicine

In healthcare, the consequences of AI hallucinations when a model generates information that sounds convincing but is completely made up can be far more serious than in most other fields. In medicine, every decision can impact someone's health, safety, or even life. If AI systems generate inaccurate diagnoses, treatment plans, or medication advice, the risks can be substantial not just for individual patients, but for the trust in the entire healthcare system.

5.1. Misdiagnosis Can Lead to Harm

Imagine an AI suggesting a rare disease based on vague symptoms like fatigue or joint pain without considering more common or likely causes. This can send clinicians down the wrong path, resulting in unnecessary testing, delayed treatment for the actual issue, and even emotional stress for the patient. While experienced physicians may question such results, less experienced users or overloaded staff might accept them at face value especially if the AI sounds authoritative.

5.2. Unsafe Drug Suggestions

Another risk is when the model "hallucinates" a medication either recommending a non-existent drug or suggesting a dangerous combination. In one scenario, it could confuse two similarly named drugs or propose a dosage that isn't safe. This isn't just theoretical mistakes like these could lead to serious side effects, allergic reactions, or harmful drug interactions. The stakes are high, and a single wrong recommendation can result in real world harm.

5.3. Losing Trust in the Technology

For AI to be a trusted partner in clinical care, it needs to be consistent and reliable. Once a system starts providing questionable or outright false answers, trust breaks down quickly. Doctors may stop using it altogether, or worse, continue to rely on it despite known issues leading to even greater risk. Patients, too, may lose confidence in their care if they learn that AI tools are influencing decisions without proper oversight.

5.4. Biases and Blind Spots

Hallucinations don't always look like wild guesses sometimes they're subtle and shaped by the data the AI was trained on. For example, an AI might downplay symptoms in underrepresented groups or miss conditions that present differently in people with darker skin tones. These errors reflect the limitations of biased training data, but the results are real: missed diagnoses, unequal treatment, and widening healthcare disparities.

5.5. Legal and Ethical Challenges

When something goes wrong, who's responsible? If a patient is harmed because of an AI suggestion, questions arise: Did the provider misuse the tool? Was the model poorly designed? Was the patient even aware that AI played a role in their care? These aren't just technical issues they're ethical and legal concerns that healthcare organizations need to take seriously.

5.6. Challenges

Ethical prompting and managing AI hallucinations in healthcare come with several critical challenges. One major concern is that AI systems can produce convincing but incorrect information, which can mislead clinical decisions and put patients at risk. Often, these models operate like black boxes, offering little transparency or traceability behind their suggestions, making it difficult for clinicians to trust or verify the output. If prompts are vague or lack patient specific context, the AI might generate irrelevant or even harmful responses. There's also the risk of reinforcing bias if prompts or models don't consider the diversity of patients, potentially worsening health disparities. Legal frameworks are still catching up, leaving questions around accountability when things go wrong. Additionally, prompt design requires a careful balance. If it's too broad, it's meaningless; if too narrow, it may exclude important nuance. Without close collaboration between clinicians and developers and ongoing validation, even the best prompts may fall short in real world care. Ultimately, while prompt engineering holds great promise, it is not a standalone fix. It must be paired with ethical safeguards, clinical oversight, and continuous learning.

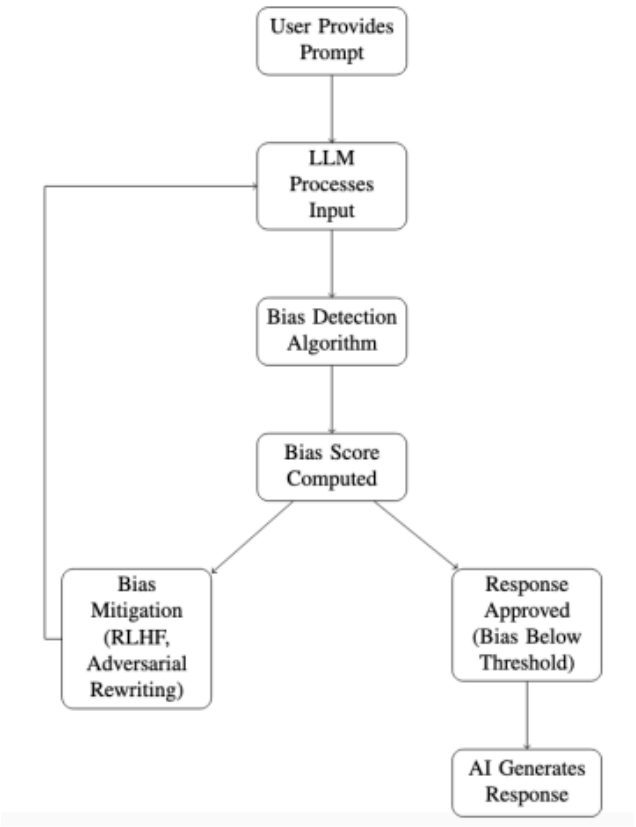

Figure 3: Bias Detection and Mitigation Pipeline [10]

6. Conclusion

In conclusion, ethical prompt design is emerging as a critical safeguard in the responsible use of AI for clinical diagnosis, particularly in mitigating the dual threats of AI hallucinations and embedded bias. Hallucinations where large language models generate information that is plausible but factually incorrect pose a unique risk in healthcare settings, where decisions based on false data can lead to delayed treatment, misdiagnosis, inappropriate therapies, or even patient harm. These risks are magnified in high stakes clinical environments where time is limited, trust in the system is essential, and accountability is non negotiable. Furthermore, AI models trained on incomplete or biased datasets may reinforce existing disparities, particularly affecting historically marginalized populations such as racial and ethnic minorities, women, and non native English speakers. In such contexts, ethically grounded prompt engineering becomes more than a technical task it is a frontline defense against inequity and harm.

By designing prompts that are precise, context rich, and inclusive of diverse patient profiles, we can help LLMs produce outputs that are not only clinically relevant but also socially responsible. This includes incorporating patient specific factors (age, race, symptoms, comorbidities), referencing validated medical guidelines, and instructing the AI to provide citations, explain its reasoning, or flag uncertainty where appropriate. Prompt strategies must also be aligned with privacy and consent principles, ensuring patients understand how AI is used in their care. Importantly, prompt engineering should not be viewed as a one off solution. It must be supported by a continuous feedback loop that includes real world performance data, clinical validation, and frontline practitioner input. Equally vital is the integration of human in the loop mechanisms to review, verify, and refine AI generated content especially when used in diagnostic pathways, triage, and clinical documentation [18-20].

Ultimately, ethical prompt design serves as a bridge between advanced AI capabilities and the realities of equitable healthcare delivery. When paired with robust auditing, legal oversight, and inclusive governance frameworks, it can help build AI systems that are not only intelligent and efficient but also transparent, trustworthy, and aligned with the fundamental principles of medical ethics autonomy, beneficence, non maleficence, and justice. As we move forward, embedding ethical prompting as a core design philosophy will be essential to ensuring that AI enhances, rather than undermines, diagnostic accuracy and health equity for all [21].

References

[1] Patil, R., Heston, T. F., & Bhuse, V. (2024). Prompt engineering in healthcare. *Electronics*, *13*(15), 2961.

[2] AJUZIEOGU, U. C. Towards Hallucination Resilient AI: Navigating Challenges, Ethical Dilemmas, and Mitigation Strategies.

[3] Dankwa Mullan I, Weeraratne D. Artificial intelligence and machine learning technologies in cancer care: addressing disparities, bias, and data diversity. Cancer Discov. 2022;12(6):1423–1427. 10.1158/2159 8290.CD 22 0373

[4] Yang J, Soltan AAS, Eyre DW, Yang Y, Clifton DA. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. NPJ Digit Med. 2023;6(1):55. 10.1038/s41746 023 00805 y

[5] Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human centered design to address biases in artificial intelligence. J Med Internet Res. 2023;25:e43251. 10.2196/43251

[6] Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. Sci. 2024;6(1):3. 10.3390/sci6010003

[7] Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon. 2024;10(4):e26297. 10.1016/j.heliyon.2024.e26297

[8] Dankwa Mullan I, Scheufele EL, Matheny M, Quintana Y, Chapman W, Jackson G, et al. A proposed framework on integrating health equity and racial justice into the artificial intelligence development lifecycle. J Health Care Poor Underserved. 2021;32(2):300–317. 10.1353/hpu.2021.0065

[9] Rajamani G, Rodriguez Espinosa P, Rosas LG. Intersection of health informatics tools and community engagement in health related research to reduce health inequities: scoping review. J Particip Med. 2021;13(3):e30062. 10.2196/30062

[10] Bura, C., Myakala, P. K., & Jonnalagadda, A. K. (2025). Ethical prompt engineering: Addressing bias, transparency, and fairness.

[11] Leslie Miller, C. J., Simon, S. L., Dean, K., Mokhallati, N., & Cushing, C. C. (2024). The critical need for expert oversight of ChatGPT: Prompt engineering for safeguarding child healthcare information. *Journal of pediatric psychology*, *49*(11), 812 817.

[12] Patil, R., Heston, T. F., & Bhuse, V. (2024). Prompt Engineering in Healthcare. *Electronics*, *13*(15), 2961. https://doi.org/10.3390/electronics13152961

[13] Alemanno, A., Carmone, M., & Priore, L. (2025). Prompting as an emerging skill for Healthcare Professionals. *Journal of Advanced Health Care*. Retrieved from https://www.jahc.it/index.php/jahc/article/view/399

[14] Kim, Y., Jeong, H., Chen, S., Li, S. S., Lu, M., Alhamoud, K., ... & Breazeal, C. (2025). Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

[15] Aljohani, M., Hou, J., Kommu, S., & Wang, X. (2025). A comprehensive survey on the trustworthiness of large language models in healthcare. *arXiv preprint arXiv:2502.15871*.

[16] Echeverría Muñoz, D. E. (2025). *Legal impact of Artificial Intelligence (AI) hallucinations* (Master's thesis, Quito, EC: Universidad Andina Simón Bolívar, Sede Ecuador).

[17] Yao Zhang, Tongquan Zhou, Huifen Qiao, Taohui Li, "Ethical Issues in AI Generated Texts: A Systematic Review and Analysis", *International Journal of Human–Computer Interaction*, pp.1, 2025

[18] Henrickson L, Meroño Peñuela A. Prompting meaning: a hermeneutic approach to optimising prompt engineering with ChatGPT. AI & SOCIETY. 2023:1 16.

[19] Giray L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. Annals of Biomedical Engineering. 2023:1 5.

[20] Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. Journal of Medical Artificial Intelligence. 2023;6.

[21] Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. Journal of Medical Systems. 2023;47(1):33.

[22] Bibi, N., Khan, M., Khan, S., Noor, S., Alqahtani, S. A., Ali, A., & Iqbal, N. (2024). Sequence-Based intelligent model for identification of tumor t cell antigens using fusion features. *IEEE Access*.

[23] G. Lakshmikanthan, S. S. Nair, J. Partha Sarathy, S. Singh, S. Santiago and B. Jegajothi, "Mitigating IoT Botnet Attacks: Machine Learning Techniques for Securing Connected Devices," 2024 International Conference on Emerging Research in Computational Science (ICERCS), Coimbatore, India, 2024, pp. 1-6, doi: 10.1109/ICERCS63125.2024.10895253