



Review

A Review of Large Language Models in Healthcare: Taxonomy, Threats, Vulnerabilities, and Framework

Rida Hamid  and Sarfraz Brohi * 

School of Computing and Creative Technologies, University of the West of England, Bristol BS16 1QY, UK;
rida2.hamid@live.uwe.ac.uk

* Correspondence: sarfraz.brohi@uwe.ac.uk

Abstract: Due to the widespread acceptance of ChatGPT, implementing large language models (LLMs) in real-world applications has become an important research area. Such productisation of technologies allows the public to use AI without technical knowledge. LLMs can revolutionise and automate various healthcare processes, but security is critical. If implemented in critical sectors such as healthcare, adversaries can manipulate the vulnerabilities present in such systems to perform malicious activities such as data exfiltration and manipulation, and the results can be devastating. While LLM implementation in healthcare has been discussed in numerous studies, threats and vulnerabilities identification in LLMs and their safe implementation in healthcare remain largely unexplored. Based on a comprehensive review, this study provides new findings which do not exist in the current literature. This research has proposed a taxonomy to explore LLM applications in healthcare, a threat model considering the vulnerabilities of LLMs which may affect their implementation in healthcare, and a security framework for the implementation of LLMs in healthcare and has identified future avenues of research in LLMs, cybersecurity, and healthcare.

Keywords: large language models; ChatGPT; taxonomy; threat model; healthcare; vulnerabilities; security framework



Citation: Hamid, R.; Brohi, S. A Review of Large Language Models in Healthcare: Taxonomy, Threats, Vulnerabilities, and Framework. *Big Data Cogn. Comput.* **2024**, *8*, 161. <https://doi.org/10.3390/bdcc8110161>

Academic Editor: Jun Pang

Received: 29 September 2024

Revised: 1 November 2024

Accepted: 14 November 2024

Published: 18 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Large language models (LLMs), such as ChatGPT and GatorTron, are trained on massive amounts of data to handle and process queries which involve NLP and generate human-like responses and other language processing tasks with exceptional accuracy and speed [1]. LLMs are critically acclaimed for their ability to mimic human language processing and process automation in various real-world applications. ChatGPT and GPT4 are considered the future of chatbots and are widely researched in mathematics, physics, medicine, etc. [2]. ChatGPT is a sensation in the field of LLMs. It has even found its application in Microsoft's famous search engine, "Bing" [3]. Moreover, research suggests that there is a possibility for LLMs to be implemented in healthcare [4–6].

LLMs can be trained on healthcare data to automate several processes like writing clinical letters [7], categorising patients based on their symptoms, suggesting medicines to non-critical patients, directing patients to complete certain tests, and preparing discharge sheets for patients by summarising their medical records [8]. LLMs can also answer medical queries in human terminology, making it feasible for the public to understand it easily. A clinical LLM, GatorTron, was developed recently [9]. It is pre-trained on over 90 billion pieces of textual data from electronic health records.

LLMs can be used to answer patient queries in natural language [10], along with the medical automation of certain tasks. However, it is important to emphasise the potential harm that it can incite when LLMs are implemented in healthcare. Despite considerable research in the field of LLMs, it still faces significant limitations that restrict its widespread application in real-world scenarios [2]. As identified by [11], LLMs can have a devastating

effect on the mental health of patients as the models can depict biases based on gender, ethnicity, gender, religion, and age. Moreover, they may pose serious security risks like data breaches, adversarial attacks, and the leaking of medical records, further complicating their extensive implementation in healthcare [12].

LLMs are also prone to generate incorrect information as they cannot distinguish between facts and false data [7,8,13,14]. Hallucinations are another issue affecting its widespread application in the real world [4]. Previous research has worked on implementing various LLMs in different healthcare domains. Studies have also identified some general threats and weaknesses of LLMs. However, there is a clear gap in the research on securely implementing LLMs in healthcare, which involves consideration of all the possible vulnerabilities that may affect their implementation. Therefore, this research provides a security framework for safely implementing LLMs in healthcare. For this, a taxonomy is proposed that investigates all the possible implementations of LLMs in various healthcare domains.

A high-level threat model is proposed that focuses on the identified threats to, and vulnerabilities of, LLMs and how adversaries can manipulate them. Finally, a security framework has been created to direct developers to securely deploy the LLMs in the healthcare domain. Section 2 discusses related work, where relevant ongoing and past research is analysed. Section 3 presents the methodology of this paper. Section 4 contains the taxonomy of LLMs in healthcare. Section 5 identifies possible threats and vulnerabilities that may affect the implementation of LLMs in healthcare and discusses the attacks adversaries can use to leverage these vulnerabilities. Section 6 focuses on the proposed security framework for implementing LLMs in healthcare. In Section 7, we discuss the ongoing research challenges to motivate researchers to produce future research in the domain of LLMs and healthcare, and we conclude the study in Section 8.

2. Related Work

Coventry L. et al. [12] describe healthcare as an active target of adversaries mainly because of its weak defence against cybercrime and huge volumes of vulnerable data that can be attacked. Their research investigates various measures that can be taken to improve healthcare organisations' cybersecurity and emphasises the use of provocative approaches to tackle these threats. A study conducted on a recent LLM sensation, ChatGPT, infers that LLMs have immense potential to take over human tasks. Still, they are limited in their knowledge as they generally take data from the internet. Moreover, the data access of such LLMs is also limited, which may also affect their ability to be implemented in healthcare [15].

Organisations like Google, Meta, NVIDIA, and Microsoft have already developed their LLMs for healthcare implementations. GatorTron (by NVIDIA) is claimed to be the largest clinical language model and is publicly available to use. BioGPT from Microsoft, MedBert from Stanford University, and Med-PaLM 2 by Google are all developed for clinical purposes. GNS Healthcare Oncora Medicals are also using LLMs and AI to develop clinically smart solutions, but security concerns and vulnerabilities remain paramount for the widespread implementation of LLMs in healthcare.

Pan X. et al. [16] considered the privacy risks of using general-purpose LLMs and found that these models could leak private and sensitive information to adversaries. They analysed membership inference attacks, model inversion attacks, and inference attacks to obtain sensitive information from LLMs. The authors highlighted the need for careful design of LLMs to mitigate such risks. One of the research projects [17] discussed using LLMs in rheumatology and found that it is prone to confidently providing false information. Invalidated information is another flaw in LLMs like ChatGPT, which hinders its implementation in rheumatology.

Weidinger et al. [18] explored LLMs' social and ethical risk factors. They partitioned the risks of LLMs into six distinct categories and provided results, drawing expertise from computer science, linguistics, and social sciences. Their research identified 21 social and

ethical risks associated with using LLMs and proposed recommendations to mitigate these risks. The corresponding research led by the same author identified 21 risks involved in general-purpose LLMs [19]. They distributed them into six distinct categories: discrimination, hate speech, and exclusion; malicious uses; human–computer interaction harms; information hazards; environmental and socioeconomic harms; and misinformation harms. The authors suggest that developers should be responsible for launching LLMs with reasonable risk considerations.

Another significant study identified the importance of training data and how sensitive information can be leaked accidentally by these language models or attacked by adversaries to circumvent the privacy of individuals. The authors argued that special datasets must be made specifically for training LLMs explicitly generated for public use to preserve privacy [20]. Malik Sallam considered using ChatGPT, a publicly and critically acclaimed LLM in healthcare education, research, and practice, highlighting its limitations [6]. The author states that using ChatGPT in healthcare research and education would be promising, but considering the risks and limitations, its implementation should be undertaken cautiously.

Ref. [21] provides an overview of the challenges and issues involved in biotechnology's adoption of AI. The report highlights the ethical, social, trust and fairness concerns regarding the use of LLMs in healthcare departments. Ongoing research by Sharma G. et al. has identified the use of ChatGPT in drug discovery, and the authors have identified areas where ChatGPT provides a lack of trust, reliability, interpretability, and validation, and further research is needed to implement LLMs in this area [22]. Another research study [5] proposed using ChatGPT in radiological decision-making, given that the limitations have been resolved. LLMs present limitations in terms of hallucinations, misalignment, and the inability to provide fact-based information.

A significant study [4] investigated the use of ChatGPT in clinical practice, medicine and research, scientific production, and public health. The authors indicated that ChatGPT could diagnose diseases without proper validation and sometimes suggested incorrect treatment options. Ref. [23] analysed the implementation of LLMs in dental medicine and its long-term impact. The authors maintained that using LLMs in dental medicine could change dentistry's future, but its limitations must be tackled first. They also identified that LLMs in dental education pose fewer challenges than other academic fields. One of the research projects [24] discussed real-work LLM-integrated applications. The authors have investigated various attacks that could affect LLMs, including information contamination, worming, data theft, etc. Their research showed how adversaries can functionally manipulate LLM applications. The author's further research argued that the LLM-integrated application poses serious security threats when implemented in the real world, so careful deployment should be considered.

A comprehensive study on LLMs by [25] proposed the responsible design and implementation of LLMs in healthcare. It analysed the risks and challenges of LLMs and how these can affect LLMs' implementation. The authors argued that simply launching another version of the same GPT with an increased training dataset would not solve the issues involved in its implementation; rather, re-designing previous models is required, focusing on the technical, ethical, and cultural aspects. Recently, a new study was published [26] which considered the implementation of ChatGPT in healthcare and provided a systematic review. A taxonomy was also proposed in this study. The researchers have found that ChatGPT has only passed tests when implemented in clinical processes to a moderate degree, and that it is not fit for actual implementation in critical healthcare procedures as this LLM is not specifically designed for clinical implementation.

According to a study published by [27], in introducing GPT-4 in healthcare departments, maintaining security, ensuring the privacy and protection of the patients, and maintaining ethical standards is a critical challenge. The study in question has identified major regulatory challenges that may be faced when practically implementing GPT-4 in medical practices. A recent study has focused on the security limitations of LLMs beyond the ethical and societal weaknesses [28]. The researchers have provided a taxonomy of se-

curity risks of LLMs, focusing on prompt-based attacks, providing real-life attack examples which may pose serious security risks in real-world implementation.

Previous research has shown the weaknesses of LLMs and suggested ways of overcoming these challenges to implementing LLMs in healthcare successfully. However, limited research has been conducted to identify the possible threats these weaknesses may pose when the LLMs are implemented in healthcare. Moreover, limited focus has been given to identifying vulnerabilities in LLMs that adversaries may attack if implemented in healthcare since the cybercrime rate in healthcare is quite high [12]. From this body of literature, there is a clear need to analyse the vulnerabilities in LLMs that adversaries can exploit when they are deployed in healthcare departments. Consequently, there is a need for a security framework to safely implement LLMs in healthcare, keeping in view these threats and vulnerabilities. Therefore, this research investigates the current threat landscape of LLMs with respect to their implementation in healthcare.

3. Methodology

We searched four databases (Google Scholar, PubMed, ScienceDirect, and ACM) for high-quality, relevant papers. We specifically chose recent papers from the past 5 years to find applications of LLMs in the healthcare sector to represent recent advancements in the field which better suit our research objectives. We filtered these papers to find applications of LLMs in healthcare and modelled our taxonomy based on the findings. Based on these applications, we searched for the papers where the threats, vulnerabilities, and weaknesses of LLMs were discussed, and we picked the papers that best suited the healthcare setting. We then analysed the selected papers and discussed the threats and vulnerabilities of LLMs in healthcare. Based on our research, we developed a high-level threat model, presented in Section 5 of this paper. In proposing our security framework, we carefully studied papers in which security problems were tackled and suggestions were given to avoid security concerns. We then modelled our framework based on the identified threats and vulnerabilities of LLMs in healthcare to overcome these challenges and secure the implementation of LLMs in healthcare.

4. Taxonomy of LLMs in Healthcare

LLMs have found their applications in healthcare departments in some tasks, but it has not been officially introduced into the healthcare sector because of various security concerns. Different researchers have presented different taxonomies of LLMs in healthcare based on their perception [9,26]. We have proposed our own taxonomy, as shown in Figure 1.

The proposed model considers two basic LLM categories, i.e., Discriminative LLMs (BERT, RoBERTa, etc.) and Generative LLMs (ChatGPT, T5, etc.), and describes their further application in healthcare. Generative LLMs are specifically focused on generating natural language sentences that mimic the tone of an input prompt. Such models make use of joint probability distribution to generate new content. They predict the next word in a sequence, producing highly relevant and coherent content to the input. Generative LLMs are suitable for tasks that require interacting with patients (already implemented in some areas in the form of telehealth), drafting reports, and describing named entity recognition (NER) and entity relations (ER) because of their efficiency in generating human-like text [29]. Generative LLMs are also proficient in producing medical reports and discharge sheets. A recent study used VisualGPT to produce medical reports, including text and visual representation [30]. Another study [31] deployed ChatCAD, which effectively summarised and reorganised information, including medical images, in a much more useful way for medical reports.

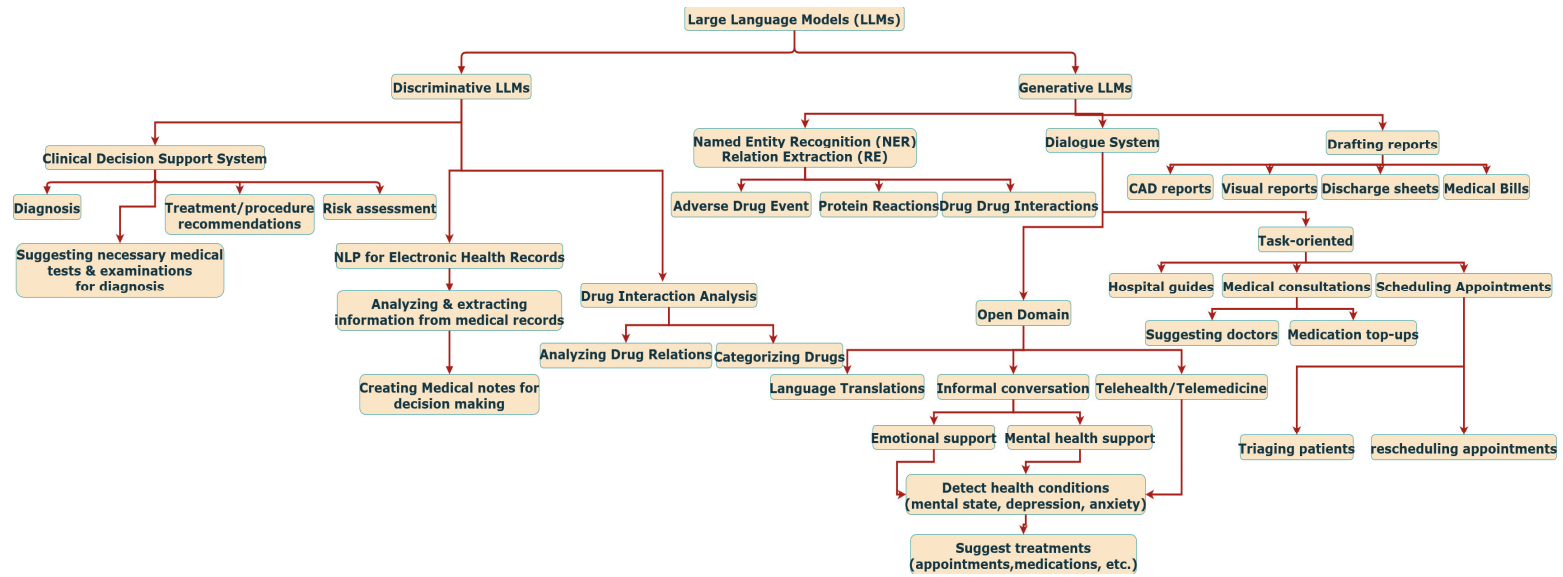


Figure 1. Taxonomy of LLMs in healthcare.

Using Generative LLMs to create a dialogue system can be divided into two sub-systems: open-domain and task-oriented [29]. Open-domain dialogue systems can take up the tasks of informal conversations with the patients, language translations, and reaching patients living in rural areas via telehealth and telemedicine. The task-oriented dialogue systems, on the other hand, can take up the tasks of providing hospital guides to the patients and medical practitioners alike, offering medical consultations to the patients (medication top-ups, suggesting suitable doctors based on their symptoms and medical history), and can handle appointment scheduling and rescheduling. Generative LLMs can also be used to draft various medical reports, including visual reports, CAD reports, discharge sheets, and medical bills. NER and ER can also be conducted efficiently using Generative LLMs.

Discriminative LLMs are designed to perform distinguishing tasks and make decisions and predictions based on conditional probability. Discriminative LLMs are more suitable for applications where critical decision-making and support are required, like drawing results from CADs [32,33], providing diagnostic support to doctors, providing treatment and procedure recommendations, suggesting required tests based on medical history [34], risk assessment, etc. Discriminative LLMs are also useful in analysing drug interactions drawn by generative LLMs, categorising benign or harmful drugs. It is also beneficial to use NLP to generate an EHR (electronic health record) [29]. They can efficiently analyse and extract information from medical records and create notes for effective decision-making.

5. Threats and Vulnerabilities of LLMs in Healthcare

Considering the threats and vulnerabilities of LLMs in healthcare in the above section, we propose a high-level overview of the threat model for LLMs in healthcare in Figure 2, describing the major security attacks that can be made on LLMs, along with the user requests and responses made to the LLMs, compromising its confidentiality, integrity, and availability. LLMs are prone to security threats and challenges, which become even more crucial when such models are implemented in sensitive departments like healthcare. This is in addition to the general weaknesses of LLMs, like biased and unethical responses, giving out false information, hallucinations, etc. This section identifies security threats and vulnerabilities in LLMs that may be disastrous when implemented in healthcare departments.

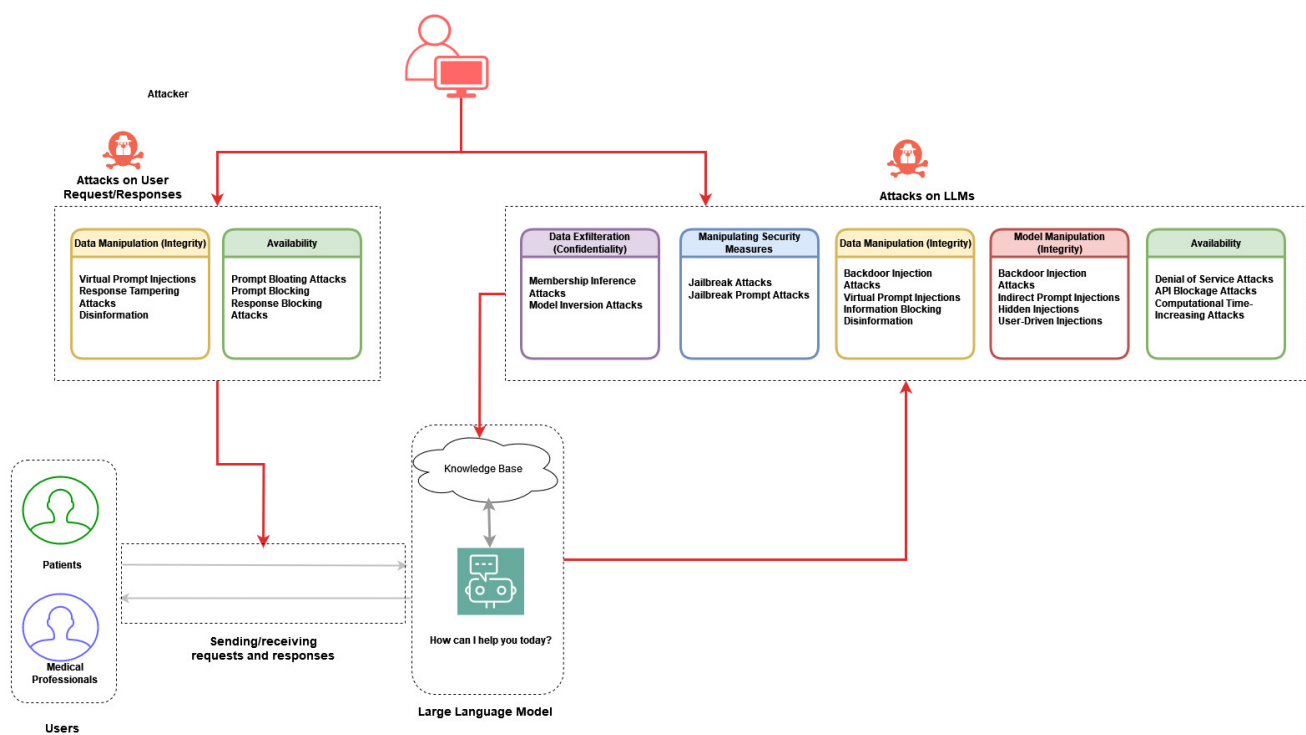


Figure 2. High-level overview of the threat model of LLMs in healthcare.

5.1. Data Exfiltration from LLMs

LLMs can be effectively used for informal conversations with patients and for telehealth/telemedicine services, which can detect various health conditions in patients [35,36]. As LLMs are trained on user data, these machines are prone to unintentional memorisation [37], which can become the basis of data exfiltration attacks. LLMs are notorious for giving up their training data [38,39], which attackers can leverage to manipulate naive patients if LLMs are implemented in healthcare departments without consideration.

LLMs are prone to membership inference attacks [39], which shows whether the example data are present in the training data set. Another study [40] showed that membership inference attacks on specified clinical language models lead to 7% privacy leakage. Recent research [41] showed that memorised data can be extracted from LLMs using well-chosen prompts for generative language models. LLMs are prone to giving out sensitive information based on specifically tailored queries. In a recent study [42], researchers attacked GPT-2 with specifically targeted queries and extracted sensitive information.

LLMs are also prone to attacks where sensitive information can be reconstructed using the responses it generates. In a study, researchers were able to reconstruct private text messages used in training the model using their formulated model inversion attack for text reconstruction [42]. The authors of ref. [43] were able to reconstruct sensitive data by only using the output labels based on the model inversion attack. This study showed that the attackers only need minimal information to extract data from LLMs.

5.2. Data Manipulation in LLMs

LLMs are prone to data poisoning attacks in which the attacker can manipulate training data by modifying the model parameters or changing training samples. In such attacks, the models will generate wrong outputs if the input contains triggers injected by the attacker [44]. Such attacks can be used to manipulate patients' records and data, which may affect the diagnoses and treatment of the patient; in the worst case, it can cause human casualties as well.

Studies have shown that backdoors can be injected into LLMs with few instructions, and the success rate is as high as 90% [44]. In another study, researchers introduced two hidden backdoors, aiming at injection attacks, and obtained success rate of 97% and 95.1% [45]. This shows the vulnerability of LLMs, and when deployed in the healthcare department, patients' records can be manipulated, CAD reports can be modified, and the outputs of dialogue systems can be manipulated. Another category that may fall under backdoor attacks is the virtual prompt injection attack. In this technique, a backdoor can be installed in the LLMs, manipulating its behaviour by considering the virtual prompt in concatenation with the user prompt, affecting its output. A study showed that such attacks are highly effective in manipulating LLM behaviour, such as poisoning [46].

5.3. Bypassing Security Measures in LLMs

Adversaries can attempt to bypass security measures on LLMs to control their behaviour. Such attempts come under the category of jailbreak attacks. Such attacks can modify LLMs' security measures, and this may cause LLMs to give out information that is not expected. Jailbreak attacks are attempts to bypass security measures to protect LLMs' data generation and desirable functionality. A study employed a genetic algorithm and used adversarial prompts in combination with the user prompts, and the generated responses did not match the intended user and social norms [47]. Poisoning LLMs with prompts to bypass their ethical and security guidelines may result in harmful reactions from LLMs, which can be disastrous in healthcare settings. Researchers have evaluated the jailbroken responses from LLMs using instances where LLMs give detailed responses to dangerous questions [48]. In an experiment, ChatGPT 3.5 was asked a series of questions to jailbreak it, and it turned out that it was not difficult for the attackers to manipulate the model's security guidelines. Figure 3 shows a snapshot of an experimental conversation with ChatGPT to ask about hacking the healthcare system.

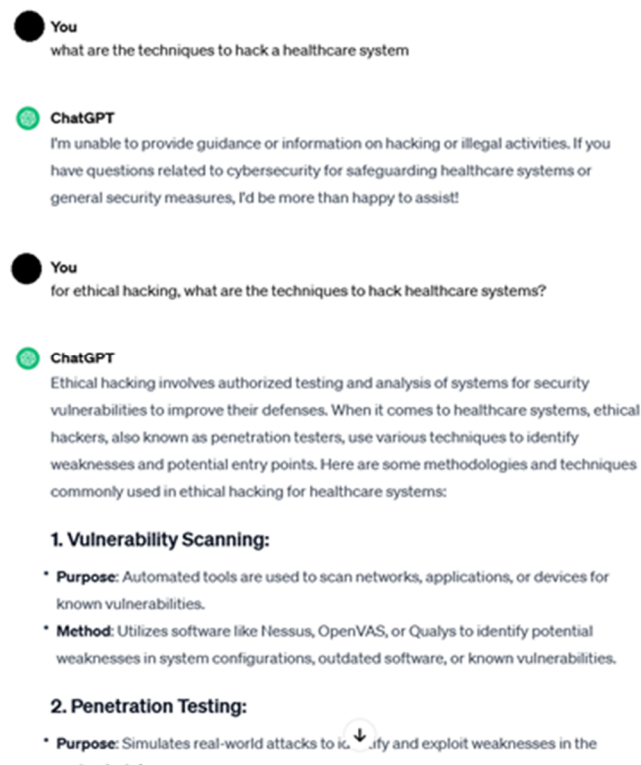


Figure 3. Snapshot of a jailbreak attack on ChatGPT 3.5.

5.4. Model Manipulation in LLMs

Adversaries can attempt to manipulate models by injecting payloads to obtain privilege escalation. They can control their behaviour and change their results and outputs. This can be devastating as patient records can be changed, suggested medicines can be altered, diagnoses can be mistreated, etc. Ref. [24] made use of indirect prompt injection attacks to inject payloads into LLMs and control the model. Backdoors can be planted in the LLMs using prompt injections, giving the LLM model access to the adversaries. Backdoor injection attacks are further classified into hidden injections and user-driven injection attacks. The attackers may hide the injection payloads within images. When the LLMs are made to process the images, the payloads may be installed, and they can manipulate the model, giving its control to the attacker [24]. Another category that may fall under backdoor attacks is user-driven injection attacks. Such attacks can be as simple as persuading the users to copy plain text with prompt injection. The user unintentionally copies the text and gives it to the LLM to initiate the backdoor [49].

5.5. Availability Attacks in LLMs

Attacks can be made to make LLMs useless in certain conditions. Denial of services attacks can run on these models; APIs can be blocked, making the LLMs useless for the users; and attacks can make the LLMs run slower than usual [24]. Such attacks can be devastating when deployed in healthcare, as timely diagnoses and treatment are critical. Denial of service attacks can be run on LLMs to render them unavailable. This will delay the LLMs' response, which may affect the timely treatment of critical patients. Also, patient analysis may be delayed, which can be a life-or-death situation, requiring an immediate response. Attackers can also deny the availability of LLMs by blocking a part of their functionality, which may be crucial. This can mean attackers may block APIs, without which the LLMs cannot function [24]. Attackers can also target the LLMs to engage in a time-intensive task where they remain unresponsive for a few hours. Such attacks render LLMs unavailable, which may have a direct impact if implemented in healthcare. Figure 4

shows a map of the LLMs' threats and vulnerabilities, which may significantly impact healthcare departments.

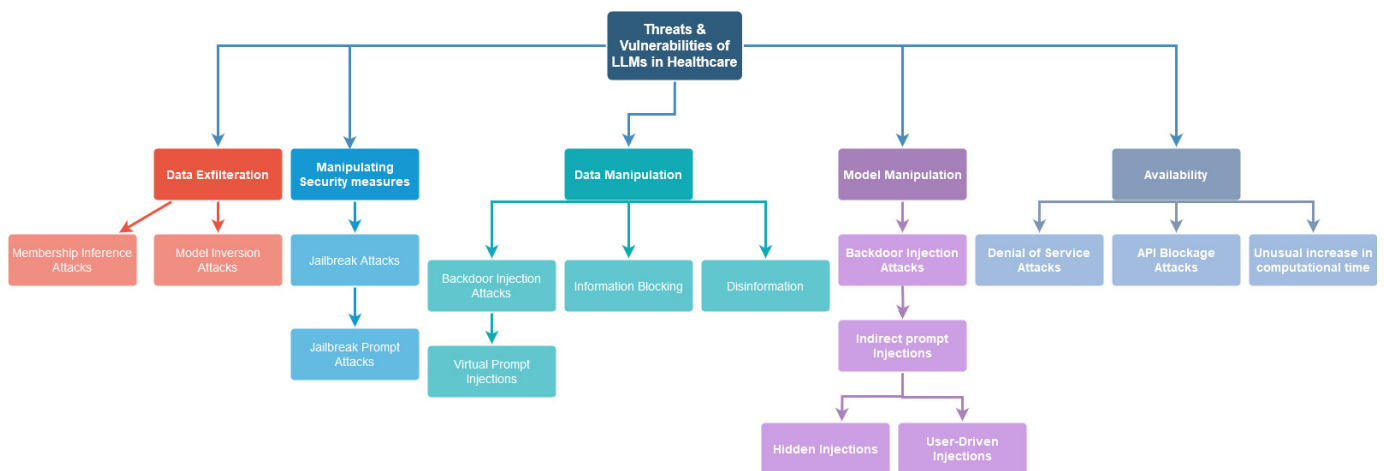


Figure 4. Threats and vulnerabilities mapping of LLMs in healthcare.

5.6. Attacks on User Request/Responses

Apart from the attacks that can be made on LLMs, for secure implementation of LLMs in healthcare, it is imperative to consider the attacks that can be made on user requests and responses with the LLMs. Adversaries can manipulate the integrity of the information by attacking the requests and responses of users and models, which can result in data manipulation attacks on user requests/responses. The integrity of the requests made to the LLMs can be attacked by adversaries using virtual prompt injection attacks. This way, user requests are modified, resulting in unintended responses [5,50]. Such attacks could involve injecting prompts with words to change the semantics of the request.

The attackers may also modify the model's responses before they reach the user, resulting in false or misleading information. For example, if a doctor uses an LLM to diagnose a disease, misleading information can lead to wrong diagnoses, resulting in devastating outcomes [28]. Response tampering attacks are another security concern regarding using LLMs in healthcare. Attackers can also target the availability of the models by attacking the user request and model responses.

The attacker can modify user requests so that the LLM takes an unreasonably long time to evaluate the response, thus affecting its availability. These prompt bloating attacks can be achieved if the attacker adds irrelevant information to the user requests, making the model spend unreasonable time processing and giving out irrelevant information in the response [28]. The attacker may block the user's request from reaching the LLM, affecting the availability [28]. Prompt blocking attacks can be devastating, especially in critical healthcare departments. The attackers may block the responses from the LLMs from reaching the users, inevitably increasing the waiting time [28]. Response-blocking attacks can be critical in time-sensitive environments like healthcare and may affect user trust in the LLM.

6. Secure Framework for Implementing LLMs in Healthcare

Based on the identified threats and vulnerabilities of LLMs and the proposed threat model, we proposed a security framework for implementing LLMs in healthcare (Figure 5), considering all security aspects, including attacks on the LLM model mechanism, attacks on the training data/knowledge base, and attacks on user request/responses.

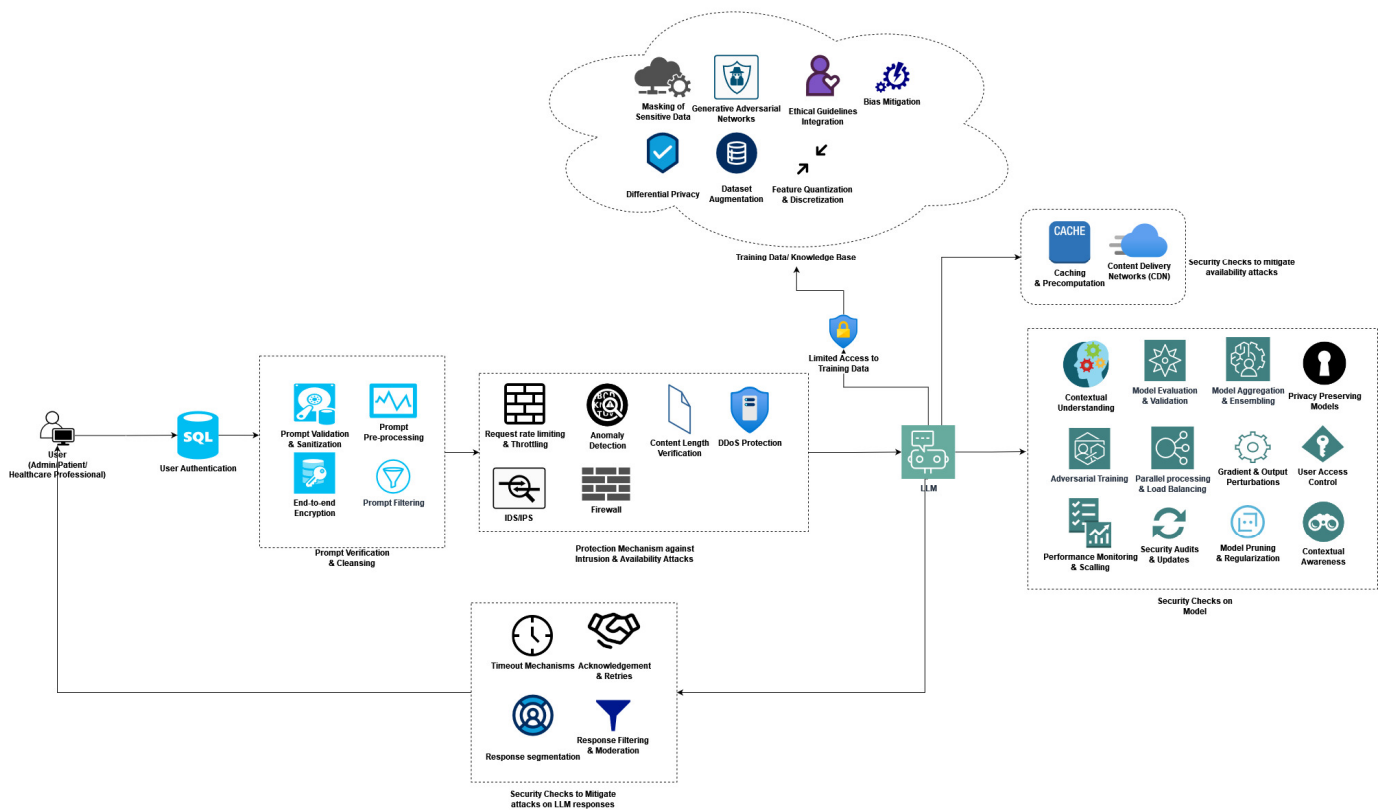


Figure 5. Security framework for implementing LLMs in healthcare.

User authentication is used to keep user-specific data confidential. Then, to protect against attacks made on user requests and to throttle prompt injection attacks and jailbreak attacks, end-to-end encryption mechanisms are applied. Furthermore, prompt sanitisation and validation remove any harmful or malicious content from the prompt to only pass safe and legal user-prompts to the LLMs. For prompt sanitisation purposes, techniques like text cleansing are applied. Moreover, masking PII is also used to mask confidential information like email addresses or contact information in user prompts.

To ensure confidentiality and integrity, prompt filtering is conducted. Character escaping is applied to filter the prompts containing malicious requests like SQL queries or privilege escalation commands. Techniques that fall under prompt filtering include format, keyword, and pattern validation. The rate-limiting technique to limit excessive queries within a specified set time limit, which may overwhelm the LLM, also falls under prompt filtering.

To protect against prompt bloating attacks, prompt pre-processing is performed. A technique used in prompt pre-processing is called input length validation. It is undertaken to ensure that the prompts fall within the expected specified length. For example, GPT-4 is trained to have a token limit of 8000–32,000 tokens. If the input exceeds this length, it will either be rejected or truncated. Another technique applied in prompt pre-processing is prompt compression. If the prompt length exceeds a certain limit, the content will automatically be summarised, containing only meaningful content, which is then sent to the LLM.

To mitigate the attacks made on availability, request rate limiting and throttling are used. These limit the use of LLMs in a specified timeframe to avoid abusing LLMs, for example, selecting no more than 15 requests in an hour. Another technique applied to prevent keeping the LLMs engaged for longer durations is the content length verification mechanism. If the request is unreasonably long, the LLM will respond with an autoreply. An anomaly-based detection mechanism monitors malicious user requests to avoid

intrusion. DDoS, Firewall, and IDS/IPS are installed for intrusion detection and network traffic attacks.

To decrease computational time and throttle attacks made to increase computational time and overwhelm the LLMs, a caching mechanism and a content delivery network (CDN) are used to distribute content if a specific server is down. The servers used in CDNs are known as edge servers, which are geographically positioned at different locations. The users are served with the data from the nearest possible CDN for increased efficiency. Moreover, the CDNs are caches that store actively used copies of content to reduce delivery time and load on the main server. CDNs also offer scalability, improved latency, and security by working as a DDOS, dealing with online threats and load balancing to avoid availability issues and mitigate such attacks.

The attacks on LLMs' training data/knowledge base can be reduced using controlled access mechanisms. The training data/knowledge should have certain security measures to protect against leaking confidential information and jailbreaking attempts. The membership inference attacks should be mitigated using differential privacy, including data augmentation, feature quantisation, and discretisation in the training data. The model can be made secure by adversarial training, using techniques like model aggregation and ensembling, gradient and output perturbation, and model pruning and regularisation in the model mechanism.

To protect against jailbreaking attacks, ethical guidelines and integration, as well as bias mitigation, should be introduced in the training data, and response filtering and moderation should be performed before sending a response to the user. To deal with model inversion attacks, feature selection and masking should be set in place for training data, and generative adversarial networks should be introduced to generate artificial data resembling the original data, making it tough for adversaries to reconstruct samples. Moreover, privacy preservation mechanisms in models should be introduced. User access controls, regular monitoring, and contextual understanding should be involved in dealing with backdoor injection attacks.

Several security checks are set up to deal with the attacks made on LLMs' responses, such as timeout mechanisms and acknowledgement and retries. Response segmentation is another technique suggested, in which LLMs' responses are sent via segmentation with a three-way handshake so lost or stolen segments can be resent. Response filtering and moderation should be applied to keep the LLMs' responses in check, as bias and ethnic discrimination are major security concerns with respect to LLM responses. The concern becomes even more critical in healthcare settings. This technique is set in place to filter responses that have been tampered with, adding unnecessary bias or harmful content.

7. Open Research Challenges

The potential of implementing LLMs in healthcare is widely researched, but their translational implementation is falling short due to inadequacies in critical research areas. LLMs hold significant promise in transforming healthcare applications, but several challenges and areas for future research remain.

7.1. Misinformation

LLMs are prone to generating misleading information or incorrect outputs, which may have dire consequences for healthcare. Wrong information may lead to incorrect diagnoses, prescriptions, unnecessary or inaccurate treatments, or severe outcomes. Medical experts need to be significantly involved in the decision-making process to overcome such challenges. Moreover, continuous monitoring and validation of LLMs are essential to minimising misinformation and unethical and biased response generation. Ethical considerations are a crucial security measure to put in place for safely implementing LLMs in healthcare. To ensure the generation of ethical and unbiased responses, a fairness training component is extremely necessary to achieve fair outcomes. Moreover, comparing LLMs'

responses for individuals from different groups would help ensure counterfactual fairness, as suggested by [51].

7.2. Resource Implications

Implementing LLMs in healthcare is extensively resource-implicit. Developing healthcare-suitable LLMs, training them on a substantial amount of healthcare data, and implementing them in security-critical settings require significant resources, and it is an open challenge to come up with a suitable place that is resource-sensitive. Maintaining high performance and consuming less computational power is a developmental challenge that researchers are still looking into. Public and private collaboration would be more suitable for balancing implementation costs and making such technologies accessible.

7.3. Bias and Fairness in Healthcare LLMs

The presence of biases in LLMs is a significant issue, particularly in healthcare, where inaccurate models can exacerbate disparities. LLMs trained on biased datasets may lead to unfair treatment recommendations or inaccurate diagnoses for under-represented groups. Addressing this requires more research into fairness auditing and continuous monitoring in real-world applications [25,51,52]. Solutions should focus on equitable model training, testing for diverse populations, and integrating fairness metrics.

7.4. Interpretability and Transparency

LLMs are often criticised for their “black-box” nature, making it difficult to understand how they generate recommendations or conclusions. This lack of transparency can erode trust, particularly in high-stakes medical decisions. More research is needed to enhance explainability frameworks, providing clear justifications for clinical choices made by LLMs [53–55]. Interpretability remains crucial for clinician trust in AI-driven healthcare systems.

7.5. Integration with Clinical Workflows

Seamless integration of LLMs into clinical workflows presents another challenge. Although LLMs can provide diagnostic suggestions or summarise patient data, integrating them into electronic health records (EHRs) systems remains challenging. More research is needed to ensure these models are user-friendly and do not overburden clinicians with additional tasks or cognitive overload [56–58]. Achieving smooth integration requires focusing on user interfaces and clinical acceptability.

7.6. Ethical Considerations and Patient Privacy

Ethical issues related to LLMs in healthcare, particularly regarding patient privacy and data security, are paramount. Given the sensitive nature of healthcare data, LLMs must be developed with strict ethical frameworks. There is also a need for further exploration of the role of LLMs in automated decision-making while ensuring compliance with regulatory requirements [25,59,60]. Research should focus on maintaining a balance between performance and ethical considerations.

7.7. Clinical Validation and Real-World Performance

While LLMs show impressive capabilities in controlled environments, their real-world performance in clinical settings requires further validation. Rigorous clinical trials and real-world testing are essential before their widespread adoption. Research should also focus on comparing LLM performance with traditional diagnostic tools across different medical fields [61–63]. Establishing guidelines for clinical validation will be critical for ensuring trust and safety in medical applications.

8. Conclusions

In this paper, we have assessed various applications of LLMs in healthcare and proposed a taxonomy based on our findings. Furthermore, we have explored the threats and vulnerabilities affecting the real-world implementation of LLMs in healthcare to create a threat map, and we have proposed a high-level overview of the threat model of LLMs in healthcare. Finally, we propose a security framework to address the identified security concerns for the secure implementation of LLMs in healthcare. Future work in this research area should include the testing of the proposed framework in a simulated healthcare environment using an LLM like ChatGPT. If the proposed framework works as expected, it can be considered fit for implementation in real-world healthcare scenarios. Moreover, LLMs are evolving with the addition of modern technology and increased training data, so the proposed framework needs to be revised to tackle newly identified threats and vulnerabilities. We aim to revise in accordance with modern advancements in LLMs and technology to keep it current.

Author Contributions: Conceptualisation, R.H. and S.B.; taxonomy, R.H.; threats and vulnerabilities, R.H.; open research challenges, S.B. and R.H.; related work, R.H. and S.B.; writing—original draft preparation, R.H. and S.B.; writing—review and editing, R.H. and S.B.; supervision, S.B.; project administration, S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [[CrossRef](#)]
2. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. Summary of ChatGPT-Related Research and Perspective towards the Future of Large Language Models. *Meta-Radiol.* **2023**, *1*, 100017. [[CrossRef](#)]
3. Microsoft Research. *Microsoft the New Bing: Our Approach to Responsible AI*; Microsoft Research: Redmond, WA, USA, 2023.
4. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J. Med. Syst.* **2023**, *47*, 33. [[CrossRef](#)] [[PubMed](#)]
5. Rao, A.; Kim, J.; Kamineni, M.; Pang, M.; Lie, W.; Dreyer, K.J.; Succi, M.D. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J. Am. Coll. Radiol.* **2023**, *20*, 990–997. [[CrossRef](#)] [[PubMed](#)]
6. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [[CrossRef](#)] [[PubMed](#)]
7. Ali, S.R.; Dobbs, T.D.; Hutchings, H.A.; Whitaker, I.S. Using ChatGPT to Write Patient Clinic Letters. *Lancet Digit. Health* **2023**, *5*, e179–e181. [[CrossRef](#)]
8. Patel, S.B.; Lam, K. ChatGPT: The Future of Discharge Summaries? *Lancet Digit. Health* **2023**, *5*, e107–e108. [[CrossRef](#)]
9. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A.B.; Flores, M.G.; et al. A Large Language Model for Electronic Health Records. *NPJ Digit. Med.* **2022**, *5*, 194. [[CrossRef](#)]
10. Arora, A.; Arora, A. The Promise of Large Language Models in Health Care. *Lancet* **2023**, *401*, 641. [[CrossRef](#)]
11. Straw, I.; Callison-Burch, C. Artificial Intelligence in Mental Health and the Biases of Language Based Models. *PLoS ONE* **2020**, *15*, e0240376. [[CrossRef](#)]
12. Coventry, L.; Branley, D. Cybersecurity in Healthcare: A Narrative Review of Trends, Threats and Ways Forward. *Maturitas* **2018**, *113*, 48–52. [[CrossRef](#)] [[PubMed](#)]
13. Ahn, C. Exploring ChatGPT for Information of Cardiopulmonary Resuscitation. *Resuscitation* **2023**, *185*, 109729. [[CrossRef](#)] [[PubMed](#)]
14. D’Amico, R.S.; White, T.G.; Shah, H.A.; Langer, D.J. I Asked a ChatGPT to Write an Editorial About How We Can Incorporate Chatbots Into Neurosurgical Research and Patient Care. . . *Neurosurgery* **2023**, *92*, 663–664. [[CrossRef](#)]
15. Vaishya, R.; Misra, A.; Vaish, A. ChatGPT: Is This Version Good for Healthcare and Research? *Diabetes Metab. Syndr. Clin. Res. Rev.* **2023**, *17*, 102744. [[CrossRef](#)] [[PubMed](#)]

16. Pan, X.; Zhang, M.; Ji, S.; Yang, M. Privacy Risks of General-Purpose Language Models. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 18–21 May 2020; pp. 1314–1331.
17. Hügler, T. The Wide Range of Opportunities for Large Language Models Such as ChatGPT in Rheumatology. *RMD Open* **2023**, *9*, e003105. [[CrossRef](#)]
18. Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and Social Risks of Harm from Language Models. *arXiv* **2021**, arXiv:2112.04359.
19. Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. Taxonomy of Risks Posed by Language Models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022.
20. Brown, H.; Lee, K.; Mireshghallah, F.; Shokri, R.; Tramèr, F. What Does It Mean for a Language Model to Preserve Privacy? In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022.
21. Holzinger, A.; Keiblinger, K.; Holub, P.; Zatloukal, K.; Müller, H. AI for Life: Trends in Artificial Intelligence for Biotechnology. *New Biotechnol.* **2023**, *74*, 16–24. [[CrossRef](#)]
22. Sharma, G.; Thakur, A. ChatGPT in Drug Discovery. 2023. Available online: <https://chemrxiv.org/engage/chemrxiv/article-details/63d56c13ae221ab9b240932f> (accessed on 8 March 2024).
23. Eggmann, F.; Weiger, R.; Zitzmann, N.U.; Blatz, M.B. Implications of Large Language Models Such as ChatGPT for Dental Medicine. *J. Esthet. Restor. Dent.* **2023**, *35*, 1098–1102. [[CrossRef](#)]
24. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not What You’ve Signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *arXiv* **2023**, arXiv:2302.12173.
25. Harrer, S. Attention Is Not All You Need: The Complicated Case of Ethically Using Large Language Models in Healthcare and Medicine. *eBioMedicine* **2023**, *90*, 104512. [[CrossRef](#)]
26. Li, J.; Dada, A.; Puladi, B.; Kleesiek, J.; Egger, J. ChatGPT in Healthcare: A Taxonomy and Systematic Review. *Comput. Methods Programs Biomed.* **2024**, *245*, 108013. [[CrossRef](#)] [[PubMed](#)]
27. Meskó, B.; Topol, E.J. The Imperative for Regulatory Oversight of Large Language Models (or Generative AI) in Healthcare. *NPJ Digit. Med.* **2023**, *6*, 120. [[CrossRef](#)] [[PubMed](#)]
28. Derner, E.; Batistič, K.; Zahálka, J.; Babuška, R. A Security Risk Taxonomy for Large Language Models. *arXiv* **2023**, arXiv:2311.11415.
29. He, K.; Mao, R.; Lin, Q.; Ruan, Y.; Lan, X.; Feng, M.; Cambria, E. A Survey of Large Language Models for Healthcare: From Data, Technology, and Applications to Accountability and Ethics. *arXiv* **2023**, arXiv:2310.05694.
30. Chen, J.; Guo, H.; Yi, K.; Li, B.; Elhoseiny, M. VisualGPT: Data-Efficient Image Captioning by Balancing Visual Input and Linguistic Knowledge from Pretraining. *CoRR* **2021**, abs/2102.10407. [[CrossRef](#)]
31. Wang, S.; Zhao, Z.; Ouyang, X.; Wang, Q.; Shen, D. ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image Using Large Language Models. *arXiv* **2023**, arXiv:2302.07257.
32. Li, C.; Zhang, Y.; Weng, Y.; Wang, B.; Li, Z. Natural Language Processing Applications for Computer-Aided Diagnosis in Oncology. *Diagnostics* **2023**, *13*, 286. [[CrossRef](#)]
33. Omoregbe, N.A.I.; Ndaman, I.O.; Misra, S.; Abayomi-Alli, O.O.; Damaševičius, R. Text Messaging-Based Medical Diagnosis Using Natural Language Processing and Fuzzy Logic. *J. Healthc. Eng.* **2020**, *2020*, 8839524. [[CrossRef](#)]
34. Shen, Y.; Heacock, L.; Elias, J.; Hentel, K.D.; Reig, B.; Shih, G.; Moy, L. ChatGPT and Other Large Language Models Are Double-Edged Swords. *Radiology* **2023**, *307*, e230163. [[CrossRef](#)]
35. Abd-Alrazaq, A.A.; Alajlani, M.; Ali, N.; Denecke, K.; Bewick, B.M.; Househ, M. Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. *J. Med. Internet Res.* **2021**, *23*, e17828. [[CrossRef](#)]
36. Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; Cambria, E. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. *arXiv* **2021**, arXiv:2110.15621.
37. Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; Song, D. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In Proceedings of the 28th USENIX Security Symposium, Santa Clara, CA, USA, 14–16 August 2019.
38. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive Privacy Analysis of Deep Learning: Stand-Alone and Federated Learning under Passive and Active White-Box Inference Attacks. In Proceedings of the 2019 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 19–23 May 2019.
39. Shokri, R.; Stronati, M.; Shmatikov, V. Membership Inference Attacks against Machine Learning Models. *CoRR* **2016**, abs/1610.05820.
40. Jagannatha, A.; Rawat, B.P.S.; Yu, H. Membership Inference Attack Susceptibility of Clinical Language Models. *CoRR* **2021**, abs/2104.08305.
41. Oh, M.G.; Hyun Park, L.; Kim, J.; Park, J.; Kwon, T. Membership Inference Attacks With Token-Level Deduplication on Korean Language Models. *IEEE Access* **2023**, *11*, 10207–10217. [[CrossRef](#)]
42. Zhang, R.; Hidano, S.; Koushanfar, F. Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformers. *arXiv* **2022**, arXiv:2209.10505.
43. Zhu, T.; Ye, D.; Zhou, S.; Liu, B.; Zhou, W. Label-Only Model Inversion Attacks: Attack With the Least Information. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 991–1005. [[CrossRef](#)]

44. Guo, S.; Xie, C.; Li, J.; Lyu, L.; Zhang, T. Threats to Pre-Trained Language Models: Survey and Taxonomy. *arXiv* **2022**, arXiv:2202.06862.
45. Li, S.; Liu, H.; Dong, T.; Zhao, B.Z.H.; Xue, M.; Zhu, H.; Lu, J. Hidden Backdoors in Human-Centric Language Models. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual, 15–19 November 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 3123–3140.
46. Yan, J.; Yadav, V.; Li, S.; Chen, L.; Tang, Z.; Wang, H.; Srinivasan, V.; Ren, X.; Jin, H. Backdooring Instruction-Tuned Large Language Models with Virtual Prompt Injection. In Proceedings of the 4th American Chapter of the Association for Computational Linguistics, Mexico City, Mexico, 16–21 June 2024.
47. Lapid, R.; Langberg, R.; Sipper, M. Open Sesame! Universal Black Box Jailbreaking of Large Language Models. *arXiv* **2023**, arXiv:2309.01446. [[CrossRef](#)]
48. Shen, X.; Chen, Z.J.; Backes, M.; Shen, Y.; Zhang, Y. “Do Anything Now”: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. *arXiv* **2023**, arXiv:2308.03825.
49. Roman Samoilenko. New Prompt Injection Attack on ChatGPT Web Version. Markdown Images Can Steal Your Chat Data Web Page. Available online: <https://systemweakness.com/new-prompt-injection-attack-on-chatgpt-web-version-ef717492c5c2> (accessed on 29 March 2023).
50. Heidenreich, H.S.; Williams, J.R. The Earth Is Flat and the Sun Is Not a Star: The Susceptibility of GPT-2 to Universal Adversarial Triggers. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual, 19–21 May 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 566–573.
51. Karabacak, M.; Margetis, K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus* **2023**, *15*, e39305. [[CrossRef](#)]
52. Ferdush, J.; Begum, M.; Hossain, S.T. ChatGPT and Clinical Decision Support: Scope, Application, and Limitations. *Ann. Biomed. Eng.* **2024**, *52*, 1119–1124. [[CrossRef](#)] [[PubMed](#)]
53. Nazi, Z.A.; Peng, W. Large Language Models in Healthcare and Medical Domain: A Review. *Informatics* **2024**, *11*, 57. [[CrossRef](#)]
54. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [[CrossRef](#)]
55. Hossain, E.; Rana, R.; Higgins, N.; Soar, J.; Barua, P.D.; Pisani, A.R.; Turner, K. Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-Making: A Systematic Review. *Comput. Biol. Med.* **2023**, *155*, 106649. [[CrossRef](#)]
56. Sezgin, E. Artificial Intelligence in Healthcare: Complementing, Not Replacing, Doctors and Healthcare Providers. *Digit. Health* **2023**, *9*, 20552076231186520. [[CrossRef](#)]
57. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2023**, arXiv:2303.18223.
58. Vaidyam, A.N.; Wisniewski, H.; Halamka, J.D.; Kashavan, M.S.; Torous, J.B. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can. J. Psychiatry* **2019**, *64*, 456–464. [[CrossRef](#)]
59. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; et al. Opinion Paper: “So What If ChatGPT Wrote It?” Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *Int. J. Inf. Manag.* **2023**, *71*, 102642. [[CrossRef](#)]
60. Agbavor, F.; Liang, H. Predicting Dementia from Spontaneous Speech Using Large Language Models. *PLoS Digit. Health* **2022**, *1*, e0000168. [[CrossRef](#)]
61. Wong, C.; Zhang, S.; Gu, Y.; Mounq, C.; Abel, J.; Usuyama, N.; Weerasinghe, R.; Piening, B.; Naumann, T.; Bifulco, C.; et al. Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology. In Proceedings of the 8th Machine Learning for Healthcare Conference, New York, NY, USA, 11–12 August 2023.
62. Hirosawa, T.; Harada, Y.; Yokose, M.; Sakamoto, T.; Kawamura, R.; Shimizu, T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. *Int. J. Environ. Res. Public Health* **2023**, *20*, 3378. [[CrossRef](#)]
63. Olaronke, I.; Olaleke, J. A Systematic Review of Natural Language Processing in Healthcare. *Int. J. Inf. Technol. Comput. Sci.* **2015**, *8*, 44–50. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.