RESEARCH PAPER

Securing AI Chatbots: A Framework for Trust and Standardization

Ahmad Musa Bulama[1], Mohammad Faisal[2], Huzaifa Muhammad Bello[3], Muhammad Muhammad Fadama[4]

[1]Department of computer Application Integral University, Dasauli, Bas-ha Kursi Road, Lucknow – 226026, India ahmadmusa@student.iul.ac.in ORCID: 0009-0007-5543-8367

[2]Department of computer Application Integral University, Dasauli, Bas-ha Kursi Road, Lucknow – 226026, India mdfaisal@iul.ac.in ORCID: 0000-0002-6120-5259

[3]Department of Computer Science and Engineering Integral University, Dasauli, Bas-ha Kursi Road, Lucknow – 226026, Indiahuzaifabe@student.iul.ac.in ORCID ID: 0009-0008-4175-1704

[4]Department of computer Application Integral University, Dasauli, Bas-ha Kursi Road, Lucknow – 226026, India fadama@student.Iul.ac.in ORCID iD: 0009-0008-4652-3321

Abstract

AI-based chatbots have taken over industries such as health care, finance, and customer services. However, the implementation of AI chatbots poses many issues regarding
trust, privacy, and proper ethical use of this technology. This paper discusses a framework to standardize trust and security in AI-based chatbots. The framework combines insights from
technical, psychological, philosophical, and cultural perspectives. The main problems tackled include issues such as lack of privacy, robustness of the models against adversarial
attacks, and ethical considerations. Some of the privacy enhancement measures discussed in this study include the use of techniques such as homomorphic encryption and
federated learning. Measures aimed at making the models resilient to attacks by adversaries such as prompt injection and data poisoning have been discussed. The framework seeks
to address the privacy paradox and other aspects such as emotional aspects and the limitations of epistemology.

Keywords: AI chatbots, chatbot security, chatbot privacy, adversarial attack, privacy paradox, artificial Intelligence

How to cite this article: Bulama AM, Faisal M, Bello HM, Fadama MM. Securing AI Chatbots: A Framework for Trust and Standardization. Int J Drug Deliv Technol. 2026;16(49s): 908-914. DOI: 10.25258/ijddt.16.49s.102

1. Introduction

Chatbots that employ Artificial Intelligence, especially the ones using LLMs, have drastically transformed Human Computer Interaction. The growing complexity of these chatbots and their use in crucial fields such as health care, banking, and customer relations require strict guidelines for chatbot trustworthiness and security. In absence of any unified standards, chatbots will remain susceptible to data theft, adversarial attacks, and unethical behavior [1], [2]. Many previous studies have focused on different facets of chatbot trust and security, such as technical vulnerabilities [3], [4], psychological elements [5], [6], culture, and philosophical limitations of chatbots [8]. Yet, these studies work in silos, leaving a notable void in knowledge regarding the interaction between these domains.

This study seeks to fill the above-discussed gaps by presenting a holistic standardization framework for trust and security in AI chatbots. The proposed framework draws on the findings of various disciplines to bring together the previously separated research streams. The suggested framework features some important contributions such as the discussion of the privacy paradox using emotional and rational privacy calculus [7], consideration of cultural factors in ethics standards [7], and resolution of epistemological shortcomings of chatbots based on philosophical arguments [8]. In essence, this framework is expected to provide practical recommendations for developers, regulators, and end-users that will enable them to deploy chatbots in a secure manner. Apart from tackling security issues such as attacks and lack of proper permission management [3], this paper highlights the need for social and emotional trust-based strategies. Some recent research has established the presence of parasocial relationships between the users and chatbots, where emotional gratification outweighs privacy considerations [7]. Also, anthropomorphism is known to increase engagement levels among the users, while at the same time creating misplaced trust towards the outputs generated by the chatbots [9], [10].

*Author for Correspondence: Ahmad Musa Bulama

2. Challenges in ensuring trust and security

A. Data Privacy Concerns

Chatbots are likely to deal with private user data, thereby being prone to cyber-attacks. Protection of data requires encryption, anonymity, and compliance with such policies as GDPR [11] and CCPA [12], but the approaches vary, thus creating different degrees
of protection. Transparency and controllability were highlighted by Guo et al., who noted that [13] users need access to their data history and an ability to correct errors simply. Alagarsamy and Mehrolia stated [6] that issues related to privacy and security affect chatbot trust

adversely because users are worried about the handling and use of their private data.

Privacy issues have made it necessary to ensure that privacy issues are addressed. For instance, there was an issue of leaking data due to a bug found in the open-source library of ChatGPT. This involved leaking sensitive data, including names, email addresses, and incomplete credit card information [14]. The GDPR and CCPA stipulate that organizations must obtain explicit consent from users, minimize data collection, and apply appropriate security mechanisms to ensure that their data is safe [15], [16]. However, complete anonymity is difficult to achieve due to challenges with irreversible anonymization of datasets [17].

Furthermore, Meng & Liu also bring to light [7] the above-mentioned difficulty by proving empirically that there is the privacy paradox phenomenon in the case of AI chatbots, where users persistently share private data recognizing their potential privacy threats. Specifically, the findings show that the stronger the parasocial relations that users build with AI chatbots, the lower their recognition of privacy threats can become due to the primacy of emotion over reasoning.

B. Adversarial Attacks

The use of adversarial attack techniques, such as prompt injection and data poisoning, may affect the behavior of chatbots, causing adverse impacts [18]. There is a need for standard test cases to detect and solve such vulnerabilities. According to Alagarsamy and Mehrolia, risk perception has a significant negative effect on the trust of chatbots ($\beta = 0.174$; $p < 0.01$). Thus, risk perception becomes one of the most significant factors influencing the trust of chatbots.

In their paper, Edu et al. [3] outline an attack vector which relies on hosting chatbots directly on developers' machines as opposed to having the chatbot run on the user's device. The advantage for the developers with such an architecture is that they have the capability of changing the chatbot at any time after installation without the users knowing about the change, unless the modification requires additional permissions. One type of attack in such a situation is jailbreak, whereby the user breaks the chatbot's security measures to create malicious content.

C. Ethical Considerations

Issues with biased training data and opaque decision-making processes can undermine user trust [21]. Ethical considerations have to be embedded within chatbot design to ensure that they are fair and accountable. Engagement and intimacy are important aspects for chatbot design, according to Guo et al., who recommend that chatbots adopt empathic and participatory communication styles [13]. According to [6], information quality ($\beta = 0.099$; $p < 0.05$) and service quality ($\beta = 0.151$; $p < 0.01$) affect chatbot trust positively.

The study findings by Edu et al [3] uncover other ethical considerations regarding permission systems within chatbot applications. In their assessment of the permissions of chatbots on Discord, 55% were found to require "administrator" permissions, which include all other permissions. These permissions raise numerous ethical concerns since the chatbots have access to a lot of information and activities within the channels without proper measures being put in place. One example is the Samsung data breach by ChatGPT [22].

D. Fundamental Trust Mechanism

The work by Magnus [8] offers a philosophical discussion of the reasons why traditional trust mechanisms fall apart when they are used for chatbots. Magnus discusses five strategies which are normally employed while examining the source of information, namely authority,

consistency of style, consistency of content, calibration, and sampling. The problem with using each one of these strategies is that they all fail when employed for chatbots. Magnus claims that chatbots are not authorities because they have no regard for the truth at all; they just try to predict a string of words without being responsible for the truth behind them.

This ontological restriction implies that chatbots will be able to produce output texts that are stylistically plausible (imitating expert language), yet substantively incorrect. Magus uses Frankfurt's technical term [8] "bullshit," defined as statements made without regard to whether they are true or false, to describe chatbot outputs by saying that "chatbots can't possibly care about the truth" because they have no world view that can serve as the benchmark for truth. This ontological restriction poses serious problems for any framework

aiming at standardization, addressing both technological security and epistemic foundations of trust.

3. Proposed framework for standardization

Proposed framework of standardization includes four pillars namely Data Privacy, Robustness, Ethics and Social Emotional Trust, based on the design aspects discussed in [13], security and privacy discussed in [24], vulnerability discussed in [4], social emotional aspects discussed in [5], empirical results discussed in [6], interpersonal trust mechanism discussed in [9], systematic review discussed in [10], parasocial interaction theory discussed in [7] and philosophy of trust mechanisms discussed in [8].

A. Data Privacy Standards

Encryption and Anonymization: User data should be encrypted both during transfer and storage phases. Anonymization approaches are to be used so as to avoid re-identification risks [25]. This is critical especially considering the findings from [6], which state that privacy and security concerns reduce chatbot trust.

Training Employees: There is a need for the training of employees on risks involved in the entry of personal information in chatbots, and also on compliance with company policy regarding data management issues. The other training programs might involve phishing, secure data entry, regulatory compliance, incident response, among others.

Contractual Agreement: Create a contractual agreement that would define terms that will limit the amount of data that is stored or used by the chatbots in order to follow

the GDPR and other similar laws. Regulatory Compliance: The chatbot has to be able to comply with various global data protection laws, like GDPR and CCPA. As indicated in [3], most of the chatbots available in Discord do not have any privacy policy which means that they are not in compliance with such laws. Data Minimization: Gather data that is needed for the specific case scenario to minimize the risks to privacy associated with data gathering and to follow the laws on privacy. For example, chatbots should only gather data that is required to provide their services.

Data Transparency: The users should be made aware of the purpose for which their data will be used through the chatbot, along with the procedures and methods involved in collecting and protecting their data. This helps in gaining the users' trust as well as allowing them to make better decisions concerning their data. This would include having a privacy policy statement and making sure that the users have access to it and can easily manage their data.

Strong Security Procedures: The use of effective security protocols is vital to safeguard sensitive data from any unauthorized access or data breaches. This would require encryption, conducting regular audits, and employing proper data storage mechanisms.

Management of Privacy Paradox: Following the findings by Meng & Liu [7], it is recommended that privacy management strategies consider both rational and emotional components of privacy decision making. It is suggested that any communication related to privacy risks should be geared towards the involvement of emotions in addition to rational awareness of privacy risks. Emotional factors tend to prevail in privacy decisions. Privacy Enhancing Technologies (PETs): The use of PETs such as homomorphic encryption, differential privacy, and secure multiparty computation can be used to improve data protection [26]. The above-mentioned techniques make it possible to perform computations using encrypted data.

Homomorphic Encryption: This approach enables computation to be conducted on encrypted data without decrypting it, which makes it very useful for preserving privacy in cloud-based chatbot applications. Homomorphic encryption, for example, can be used to perform analytics on user feedback or sentiment analysis without compromising privacy

[26]. Both Microsoft and Google have started to integrate homomorphic encryption technology into their cloud offerings as a measure to protect client data [27]. Several studies have also established the efficacy of using homomorphic encryption in chatbot security frameworks. In a recent study, Jalali and Chen [28], for example, developed an extensive framework based on blockchain, federated learning, and fully homomorphic encryption techniques for securing and protecting user information from security threats associated with chatbots. Not only does this framework improve chatbot efficiency but it also preserves user data integrity, confidentiality, and authenticity by ensuring secure encryption of large data sets (up to 1792 MB) with a record encryption rate of 1240 blocks per second, with each block representing computation on encrypted data [26]. Computational costs associated with this approach can be alleviated using hybrid models.

Differential Privacy: The technique involves adding noise to the data to ensure that no single record is identifiable, thereby allowing for the computation of statistics from the data without revealing any individual data point. Differential privacy is an indispensable requirement for chatbots designed to gather user information for statistical analysis and customization, as it guarantees that no single user's data can be deduced from the computed statistic [26]. Apple successfully adopted the approach when designing its iOS platform [29].

Federated Learning (FL): Federated learning is a possible solution in order to improve the privacy of chatbots by means of collaborative machine learning without exchanging raw data. Data in federated learning stays on devices, while updates of machine learning models are sent to a central server [30]. However, traditional federated learning is faced with problems like single-point failure and model inversion attack because of its centralized nature. There has been some work recently suggesting decentralized federated learning (DFL) with the help of blockchain technology to solve the problem. As an example,

Su et al proposed a DFL scheme in which the whole training process was orchestrated using smart contract of blockchain technology. In such a way, user data stays private and secure. Moreover, secure multi-party computation (SMPC) in blockchain networks is applied after every round of local learning. This way of doing allows achieving comparable results with centralized models [31].

Secure Multi-Party Computation (SMPC): The SMPC technique allows various parties to carry out computations on their private data without exposing it to the rest of the parties. The use of such technology is ideal for chatbots that need to collaborate among different organizations while keeping their data private. In health care applications, SMPC can be used to ensure privacy when sharing medical data between hospitals and insurance companies [26]. Generative artificial intelligence has been demonstrated to employ SMPC to ensure privacy of data in a decentralized manner in recent literature [32]. This technique can be employed in chatbot applications where generative AI algorithms are widely applied to generate responses.

B. Robustness Standards

Adversarial Testing: Chatbots should be subjected to stringent testing involving attacks from adversaries such as prompt injection and data poisoning [33]. Recent studies have proven the susceptibility of LLM-powered chatbots to data poisoning attacks. For example, through red teaming, Fernandez et al. found out [34] that the BarkPlug chatbot is vulnerable to poisoned documents which cause erroneous outputs.

Continuous monitoring and testing must be done in order to identify vulnerabilities like those identified in red teaming activities. Continuous Monitoring: Monitoring techniques that provide real-time feedback in case of threats must be put into place. This is because according

to Edu et al., chatbots can be easily altered at any time after deployment in cases where they are hosted by developers.

Epistemic Guardrails: Following Magnus's approach [8] for trust systems, chatbots need to establish definite epistemic guardrails that restrict their capacity to make any claim without evidence where the training data is limited. This involves using the notion of "negative use of appeal to plausibility" discussed by Magnus in his paper, in which he recommends flagging any content produced with low internal confidence as possibly misleading. For implementation, chatbots could utilize an entropy measure for quantifying uncertainty in their predictions, which would filter out any predictions made with low confidence. This can be done using similar ideas proposed in calibrated classifiers [35], where the system will label its response as "likely generated" or "lacks evidence" alongside the response itself. Additionally, techniques like layer-wise relevance propagation or SHAP values might provide information regarding which prompt tokens were relevant to generate the output.

Adversarial Prompt Engineering: In recent times, it has been observed that chatbots such as ChatGPT are susceptible to adversarial prompts which force them to generate certain output, sometimes undesirable ones [36]. For instance, in Wang et al., it was revealed that the use of adversarial prompts could lead to chatbots such as ChatGPT generating harmful content [36]. Therefore, it becomes imperative to adopt a sound adversarial testing procedure to avoid this problem. Out-of-Distribution (OOD) Robustness: It is necessary for a chatbot to be able to handle out-of-distribution inputs. Recent research has found that even though LLMs are effective, they can have problems with handling out-of-distribution inputs, resulting in erratic behavior [36].

Robustness via Question Generation: The employment of controllable question generation algorithms, such as the Diversity Controllable Semantically Valid Adversarial Attacker (DCSA), introduced by Pan et al. [37], can contribute towards improving the robustness of the chatbot by producing adversarial question-answer pairs. This approach involves using a semantic graph to guarantee fluency and semantic validity in the generated questions [37].

Mitigation Techniques: Using robustness improvement techniques like adversarial training and data augmentation may increase the robustness of the chatbot towards any kind of attack [36]. With adversarial training, models are trained using adversarial samples to make them more robust, whereas data augmentation introduces new data diversity for better generalization.

Periodic Security Audits: Perform periodic security audits to find weaknesses and make sure that your chatbot is still secure. This includes not only testing against known security threats but also those that are not yet discovered [3].

C. Ethical Standards

Mitigating Bias: The training data used to train chatbots should be checked for any biases, and there should be inclusion of fairness metrics while evaluating the models [21]. It becomes very relevant since, according to Edu et al., chatbots tend to become over-privileged and hence, may end up propagating certain biases [3]. Transparency: Chatbots should explain the reasoning behind their actions or outputs so that users can understand and believe them [38]. As mentioned by Alagarsamy & Mehrolia, information quality is one of the factors that affect chatbot trust [6].

Cultural Sensitivity: According to Meng & Liu's observations [7], regarding cultural differences in attitudes towards AI chatbots, the proposed framework should include awareness and adaptability with regard to culture. The creation of AI chatbots should take into account the way in which various cultures approach AI technology and ensure that ethical standards are developed in consideration of their unique cultural backgrounds.

Transparency of Discourse Map: Based on the classification by Magnus [8], of "maps of discourse" versus "models of the world," it is essential for chatbots to make this difference clear to users. In particular, chatbot responses that are formulated based on linguistic structures should be distinguished from statements made based on facts. This is addressed by the inherent weakness of chatbots mentioned by Magnus [8].

Transparency of discourse maps can be ensured by applying the metadata tagging approach, which categorizes bot responses as either retrieval based (factual sources) or generative based (inferred patterns). For instance, OpenAI employs its system message injection, and Google employs its Bard chatbot, which both apply distinctive visual markers that indicate whether the source backs up a claim. Adopting such mechanisms within an open domain can be achieved by tagging each response with provenance metadata markers [38].

D. Social-Emotional Trust Standards

Trust Conceptualization Consistently: According to the results of studies by Ng & Zhang [10], there is an inconsistency in conceptualizations of trust. Thus, to overcome such discrepancies in definition, it would be reasonable to develop a multi-dimensional framework that would take into account different kinds of trust: behavior-based, emotional, attributes, and situational. Management of Parasocial Relationships: In line with the conclusions reached by Meng & Liu [7], who argue about the importance of parasocial relationships and their role in privacy choices, the framework should include guidelines for parasocial relationships management. Specifically, in relation to chatbots, one should pay attention to designing emotionally engaging bots while ensuring proper consideration of privacy issues.

Clarity of Trust Mechanisms: As per the trust mechanism failure analysis presented by Magnus [8], chatbots should have transparency in the form of their epistemic shortcomings. These include the implementation of what Magnus calls "calibration" mechanisms that ensure users do not make an incorrect

trust expansion from tasks where the chatbot is competent to those where it isn't.

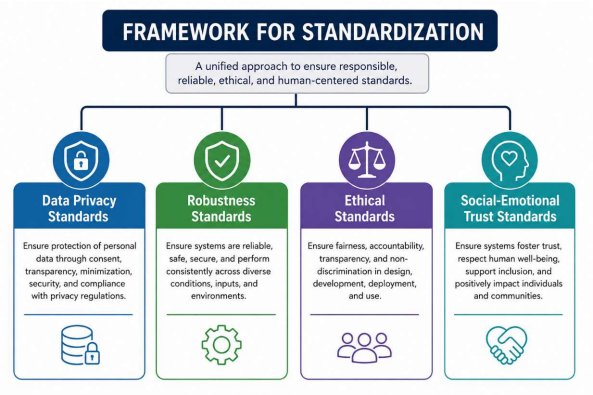


Fig 1: Framework for Standardization

4. Implementation and validation

For verifying the real-world usability of the proposed framework, the following two phases of an evaluation approach are suggested:(1) Technical testing for adversarial prompt injections and epistemic guards via prototyping, and (2) User evaluations on trust calibration and privacy behavior. As part of Phase 1, the chatbot prototypes need to be tested against adversarial prompt injection attacks through benchmarking tests with Text Attack or AutoPrompt [36].

When it comes to epistemic guards, experiments on model calibration can be conducted with ECE (Expected Calibration Error) and Brier scores. Explainability techniques such as SHAP (SHapley Additive exPlanations), Captum (explainability for PyTorch models), and Transformers Interpret (explainability for transformer-based language models) can also help reveal the inputs that had the most impact on a specific output response from the model. The explainability methods mentioned here are useful for trust calibration because they can make the model more transparent and thus provide greater interpretability to users.

The systematic review by Ng & Zhang [10] indicates that there are serious methodological shortcomings in existing studies investigating trust in AI chatbots. According to their findings, the majority of the research (80%) used cross-sectional survey methods, while only one paper used the longitudinal survey approach. Since the formation of trust is a complicated and evolving process, it is necessary for the framework to be implemented and validated with longitudinal studies.

The methodology employed by Meng and Liu [7] through questionnaire surveys conducted among ChatGPT users (N=480) constitutes an important benchmark for assessing some components of the framework, especially in relation to parasocial relations and privacy disclosure behaviors. The fact that privacy calculus mediates the connection between parasocial interactions and privacy disclosure is a critical variable to measure the validity of privacy standards.

In evaluating the epistemological aspects of the framework, the trust mechanism model by Magnus [8] offers some useful criteria. In achieving the standards of "epistemic guardrails"

and "transparency of discourse map," success will depend upon whether users properly calibrate their trust in chatbot outputs by applying "negative use of appeal to plausibility" to weed out implausible outputs without being misled into false trust based on superficial plausibility in output style. In Phase 2, user-facing metrics can be tested through controlled experiments on trust calibration, based on the method suggested in [10]. For example, assessing the users' accuracy in recognizing chatbot output plausibility before and after the application of transparency cues can help validate the metrics empirically. A longitudinal study, drawing from [7], can also be conducted to examine how parasocial trust is developed or eroded over multiple interactions.

5. Future challenges and research directions

Going forward, there are a number of challenges that must be considered regarding the evolution of AI chatbots:

- Multimodal AI: The combination of text-based AI with other modalities like voice or visual chatbots creates some additional trust and security concerns. Further research must be undertaken to establish standards for multimodal AI.
- Decentralized AI: With the development of decentralized AI, trust and security considerations take on greater significance. More research must be done on regulating and securing decentralized AI chatbots.
- Autonomous Agents: Beyond simple chatbots, there will be autonomous agents in many different industries (such as robotics or drone technology) that will require trust and security systems.

6. Conclusion

Given the increasing use of AI-driven chatbots, the issues of their reliability and security become critical. In this respect, the framework suggested, which takes into account several viewpoints discussed by technical, empirical, and philosophical studies, serves as a starting point for standardizing such practices and helping to ensure the safety of chatbots.

For instance, recent studies focused on robustness involving controllable question generation models [37] reveal the significance of developing diverse examples of attacks targeting the ability of chatbots to deal with challenging and potentially hostile questions. This also supports the necessity of creating robust tests including the assessment of vulnerabilities, which should be comparable with red team evaluations [34], and out-of-distribution robustness testing [36].

The philosophical analysis by Magnus [8] of the means of trust shows one core problem: chatbots operate without any interest in truth since they possess only "maps of discourse" and not "models of the world." According to him, "a chatbot is simply not the kind of thing which could be an authority." In order to solve the

problem, we need a standardization solution considering both security and epistemology.

Meng and Liu's empirical verification [7] of the privacy paradox in AI chatbot communications reveals the immediate requirement for guidelines that cater to both rational and emotional considerations when it comes to privacy choices. Emotional elements usually weigh more heavily on individuals, who voluntarily give away personal details even when they are concerned about their privacy.

Future research should concentrate on enhancing the framework according to the results of pilot implementations, extending its use into other chatbot applications, and creating automatic mechanisms for evaluating adherence to the standards. By working together, researchers, businesses, and regulatory agencies can create a strong standardization framework that guarantees the trustworthiness and usefulness of chatbots for everyone involved.

References

1. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," in ICLR, 2015.
2. Luciano Floridi et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," Minds and Machines, pp. 689–707, 2018.
3. Jide Edu et al., "Exploring the security and privacy risks of chatbots in messaging services," in IMC '22: Proceedings of the 22nd ACM Internet Measurement Conference, Nice, 2022, pp. 581-588.
4. Winson Ye and Qun Li, "Chatbot Security and Privacy in the Age of Personal Assistants," in IEEE/ACM Symposium on Edge Computing (SEC), San Jose, 2020, pp. 388-393.
5. Moez Ltifi, "Trust in the chatbot: a semi-human relationship," Future Business Journa, 2023.
6. Subburaj Alagarsamy, Mehrolia, and Sangeeta, "Exploring chatbot trust: Antecedents and behavioural outcomes," Heliyon, vol. 9, no. 5, May 2023.
7. Xiaoxiao Meng and Jiaxin Liu, "'"Talk to me, I'm secure": investigating information disclosure to AI chatbots in the context of privacy calculus," Online Information Review, Feb. 2025.
8. P.D. Magnus, "On trusting chatbots," Episteme, pp. 111, Jan. 2025.
9. A. M. Bulama and M. Faisal, "Human-AI collaboration in education: Enhancing learning environments," in Proc. Int. Conf. Trends and Development in Science and Engineering, Sep. 29, 2025, doi: 10.17577/IJERTCONV13IS06039.
10. Sheryl Wei Ting Ng and Renwen Zhang, "Trust in AI chatbots: A systematic review," Telematics and Informatics, vol. 97, Feb. 2025.
11. European Union. (2016, May) GDPR. [Online]. https://gdpr-info.eu/
12. State of California. (2018) CCPA. [Online]. https://oag.ca.gov/privacy/ccpa
13. Yunsan Guo, Jian Wang, Runfan Wu, Zeyu Li, and Lingyun Sun, "Designing for trust: a set of design principles to increase trust in chatbot," CCF Transactions on Pervasive Computing and Interaction, vol. 4, pp. 474-481, May 2022.
14. OpenAI. (2023, Mar.) March 20 ChatGPT outage. [Online]. https://openai.com/index/march-20-chatgptoutage/
15. Dialzara. (2024, May) Dialzara. [Online]. https://dialzara.com/blog/ai-chatbot-privacy-data security-best-practices/
16. Glorin Sebastian, "Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information," International Journal of Security and Privacy in Pervasive Computing, vol. 15, no. 1, Aug. 2023. [Online]. 10.4018/IJSPPC.325475
17. Suntherasvaran Murthy, Asmidar Abu Bakar, Fiza Abdul Rahim, and Ramona Ramli, "A Comparative Study of Data Anonymization Techniques," in IEEE International Conference on BigDataSecurity,HPSC,IDS, Washington DC, Mar. 2019, pp. 306-309.
18. Nicolas Papernot et al., "The Limitations of Deep Learning in Adversarial Settings," in IEEE European Symposium on Security and Privacy (EuroS&P), Saarbrücken, 2016.
19. Gelei Deng et al., "MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots," in Network and Distributed System Security Symposium, 2023.
20. Lara Alotaibi, Sumayyah Seher, and Nazeeruddin Mohammad, "Cyberattacks Using ChatGPT: Exploring Malicious Content Generation Through Prompt Engineering," in ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), Bahrain, 2024.
21. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys, vol. 54, no. 6, July 2021.
22. Lindsey Wilkinson. (2023, Apr.) Samsung employees leaked corporate data in ChatGPT: report. [Online]. https://www.ciodive.com/news/Samsung-ElectronicsChatGPT-leak-data-privacy/647137/
23. Harry G. Frankfurt, On Bullshit.: Princeton University Press, 2005.

24. Martin Hasal et al., "Chatbots: Security, privacy, data protection, and social aspects," Concurrency and Computation: Practice and experience, vol. 33, 2021.
25. Narayanan, Arvind, Shmatikov, and Vitaly, "Myths and fallacies of "Personally Identifiable Information"," Communications of the ACM, vol. 53, no. 6, pp. 24-26, June 2010.
26. Ali Reza Ghavamipour, Fatih Turkmen, and Jiang Xiaoqian, "Privacy-preserving logistic regression with secret sharing," BMC Med Inform Decis Mak, Apr. 2022. [Online]. https://doi.org/10.1186/s12911-02201811-y
27. Microsoft. (2018, Feb.) Microsoft SEAL. [Online]. https://www.microsoft.com/en us/research/project/microsoft-seal/
28. Nasir Ahmad Jalali and Hongsong Chen, "Comprehensive framework for implementing
29. blockchain-enabled federated learning and full homomorphic encryption for chatbot security system," Cluster Computing, vol. 27, pp. 10859–10882, May 2024.
30. A. M. Bulama and M. Faisal, “The evolution of cybersecurity: Analyzing modern threats, overcoming challenges, and shaping future directions,” International Journal of Engineering Research & Technology (IJERT), vol. 15, no. 2, Feb. 2026, doi: 10.17577/IJERTV15IS020645.
31. Ait-Mlouk, Alawadi,Addi Sadi , Salman Toor, and Andreas Hellander, "FedBot: Enhancing Privacy in Chatbots with Federated Learning," arXiv, Apr. 2023.
32. Hongxu Su, Cheng Xiang, and Bharath Ramesh, "Towards Confidential Chatbot Conversations: A Decentralised Federated Learning Framework," The Journal of The British Blockchain Association, vol. 7, no. 1, pp. 1-15, 2024.
33. Manil Shrestha, Yashodha Ravichandran, and Edward Kim, "Secure Multiparty Generative AI," in arXiv, 2024.
34. Nicholas Carlini and David Wagner, "Towards Evaluating the Robustness of Neural Networks," in IEEE Symposium on Security and Privacy, San Jose, 2017.
35. Ivan A. Fernandez, Subash Neupane, Sudip Mittal, and Shahram Rahimi, "Poison Attacks and Adversarial Prompts Against an Informed University Virtual Assistant," in arXiv, 2024.
36. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, "On calibration of modern neural networks," in International Conference on Machine Learning, Sydney , 2017.
37. Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, and Runkai Zheng, "On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective," IEEE Data Engineering Bulletin, Feb. 2023.
38. Yan Pan, Mingyang Ma, Bernhard Pflugfelder, and Georg Groh, "How to Build Robust FAQ Chatbot with Controllable Question Generator?," arXiv, Nov. 2021.
39. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016, pp. 1135 - 1144.