

## Homomorphic encryption for secure conversations with AI bots over cloud to prevent Social Engineering attacks

Yameen Ajani

Computer Engineering, Fr. Conceicao Rodrigues College of Engineering Mumbai, India

Krish Mangalorkar

Computer Engineering, Fr. Conceicao Rodrigues College of Engineering Mumbai, India

Yohann Nadar

Computer Engineering, Fr. Conceicao Rodrigues College of Engineering Mumbai, India

Sunil Chaudhari

Computer Engineering, Fr. Conceicao Rodrigues College of Engineering Mumbai, India

Mahendra Mehra

Computer Engineering, Fr. Conceicao Rodrigues College of Engineering Mumbai, India

### ABSTRACT

The advancement in the field of Artificial Intelligence has opened the gates for AI-based applications, especially chatbots, to be commercially used in a variety of industries due to the exceptional features that these chatbots offer. However, in many industries like finance and healthcare, sensitive and valued information is exchanged between the user and the AI bot. These bots, although powerful, generally lack sufficient security leaving the conversation exposed to hackers. If such a system gets compromised, it can be used as a means of psychologically manipulating the users and hence tricking them into giving out sensitive information which can be the basis of a social engineering attack. Our paper aims to provide a unique solution to this problem by using homomorphic encryption to secure the conversation and data exchange between the user and a bot hosted on the cloud. Homomorphic encryption allows us to perform computation on data which is encrypted. This data can be of various types including text and images. This enables us to use and manipulate encrypted data without the need to decrypt it. Since no information is left unencrypted at any point, if any data is sniffed by a hacker, no information is compromised and hence the threat of any kind of social engineering attack is nullified.

**Index Terms**—Homomorphic Encryption, chatbots, social engineering attacks, data security, encryption

### I. INTRODUCTION

#### A. *The evolution of AI*

Artificial Intelligence is no longer the dystopian future it once was considered to be. It has indeed become a reality. Looking at the current use cases of AI in the real world, it becomes clear that AI has massive potential in improving human life. One of the applications of AI, which has gained eminence and is currently on the rise are chatbots. Chatbots have gained immense popularity for both Business-to-Business and Business-to-Client brands that aim to provide a richer experience for their customers. According to a research published by Gartner, it has been predicted approximately 67% of customer service and support operations will integrate and utilize chatbot technology by 2022 [1]. AI has proven its potential to provide meaningful innovation which

can be used to empower businesses, researchers and innovators but the key piece to this utopian puzzle is to enforce reasonable regulations around this technology.

#### B. *Threats due to AI chatbots*

With the ever increasing power of AI technology, it is now capable of replicating human aspects like speech and conversation with high accuracy. Due to this high degree of resemblance, it has become increasingly harder to differentiate between AI bots and human representatives thus opening doors for social engineering attacks and other felonious activities with the help of compromised or malicious bots. One such case of chatbots being compromised and used for malicious activity is that of Tinder. The hackers used the chatbot as a front to impersonate

an individual with the motive to trick them into giving out their payment details [2]. This is a perfect example of the impersonation technique used in social engineering attacks. The victim blindly trusts the requests of the chatbot, not being aware of the fact that it is being controlled by a cybercriminal. With numerous conversations everyday, chatbots have access to large amounts of user data which makes them a prime target for cybercriminals and thus making security of the chatbot as well as the conversations a matter of great importance [3].

### C. Onset of Social Engineering

In the recent years, the field of cybersecurity has become extremely powerful owing to the fact that it can and already has been combined with the domain of Artificial Intelligence in order to provide more secure techniques for software and data security. This has also proved to be of great help towards the prevention of zero-day attacks [4]. However, even with the increase in the awareness campaigns regarding the prevalent glooming threats related to information security, there still continues to be data security compromises which has evidently evaded our competency to defend against. These security barrier breaches have led to billions of dollars being lost yearly both for corporations and individuals [5].

One such threat to data confidentiality is evidently the social engineering attack, which has become popular for using sophisticated psychological manipulation techniques in order to extract sensitive and personal user information. Social engineering attacks can be seen as methods to exploit individuals psychologically and thus tricking them into giving out sensitive information to achieve a malicious objective. These attacks are so efficient, that presently, these exploits are one of the major contributing factors which support the majority of the cyber attacks, even more so than the traditional technical cyber exploits [6]. With the growing use of customer support and collaboration tools in business environments, the exposure of sensitive and personal information of the users through social exploitation by cybercriminals is now a pressing concern. These attacks in unison with zero-day exploits endangers user data at best and can devastate the organisation at worst.

### 1) SOCIAL ENGINEERING

**METHODOLOGY:** Even though individuals who are victims of social engineering attacks may have different vulnerabilities, the cybercriminals generally follow a set of conventional procedures as far as social engineering attacks are concerned.

1) Information gathering - The attacker seeks

to obtain personal information from the victim's social profiles for example, date of birth, photos, friend groups, locations, likes/dislikes e.t.c. The attacker can then generate deeper insights from all the information gathered.

2) Trust building - With the information that the attacker has gathered, he enters the same social circles as the victim and now tries to build relations and establish trust with the victim.

3) Social Exploitation - At this stage, the attacker uses the trust built to manipulate the victim and tries to extract the desired information for malicious use.

### 2) COMMON SOCIAL ENGINEERING ATTACK TECHNIQUES:

1) Phishing - These scams generally contain emails and messages that highlight a certain urgency to respond to the aforementioned message, thus redirecting to malicious website clones or opening malware-containing attachments [7].

2) Man in the middle (MitM) attack - MitM attacks intercept the messages between the two parties having a conversation and replaces it with malicious messages [8]. In this type of attack, the chatbot is made to seem as if it is representing a credible organization, whereas in reality it is being operated by a malicious attacker. An example of a MitM bot is Honeybot [9].

3) Corrupted communication channel - In case of self learning chatbots, if a large number of users knowingly provide false or inappropriate information to the chatbot, it affects the learning thus influencing its decision making ability. For example in 2016 Microsoft's Tay bot got affected by this attack and was used to target a particular group by spreading inappropriate and false messages. AI chatbots, which are used as customer support tools, are also vulnerable to such threats and thus making security of such software tools a matter of high priority.

### D. Why is Homomorphic Encryption needed here?

The objective now is to be able to run these AI applications over the cloud architecture by having the data encrypted throughout the transit in a persistent manner, so that even in the case of unauthorised access by a cybercriminal over these data centers, there wouldn't be a way for the attacker to access any sensitive data. Traditional encryption algorithms implement a concept of a distributed key i.e both a public key and a private key which is shared among the communicating parties so that they can access the data exchange. This implementation however introduces privacy concerns as the service providers of this communication channel also have access to this

data. This brings about another concern where malicious providers may have an unregulated tab on user data. Moreover these conventional encryption methodologies need to decrypt the data before performing any task on it for example, storing data on a cloud based storage system. This brings the need for practically implementing an encryption scheme which allows operations to be performed on encrypted data itself isolating the need for knowing the unencrypted contents of the data. Here's where Homomorphic Encryption comes into the picture by providing a way for enterprises and large organizations to exponentially strengthen their resistance to cyber attacks and data breaches. Homomorphic encryption gives the ability to perform computation on encrypted data, the result of which, also in the encrypted form, when decrypted is identical to the result of the same computation performed on the original data [10].

Homomorphic encryption, when used particularly in a cloud based environment excels as an encryption methodology. Homomorphic encryption also proves beneficial if the system is working with different data types like numbers, text, images and other multimedia at the same time. Apart from text, a cloud storage proves more efficient and scalable when images and their processing is concerned [11]. Taking multimedia objects like images into consideration is important and they often contain important information and if acquired by malicious third party service providers can be used to identify the victim, their friends, family, etc. which can be used for social engineering [12]. Such compromise of data in the form of images needs to be prevented. Since homomorphic encryption supports all different data types, it mitigates the problem of privacy breach while also making image processing over the cloud possible apart from working on textual data. Since chatbots are AI applications, it is extremely likely that they accept image-based data from the users and also forward it to deep learning models for further processing. The results from the aforementioned models then return to the chatbot and are displayed to the user. If these images are homomorphically encrypted, they make the system airtight in terms of security. While the raw image cannot be accessed by the third-party deep learning model, it does not cause a problem to the processing required to be done by the model due to the homomorphic properties.

Moving forward in the future, any organisation that seeks to attain excellence with respect to data integrity standards will most certainly have to embrace homomorphic encryption.

## II. LITERATURE SURVEY

Chatbots, as previously mentioned, are like every other technology. They have vulnerabilities which can be exploited thus compromising sensitive and important information.

Malicious users can hamper the response generation module by providing it with corrupted input messages which might trigger the chatbot and result in undesired responses in the form of foul language or untrustworthy information. Thus, attacks directed at the chatbot's dialogue system with skillfully crafted malicious input statements can break the chatbot [13]. Language model attacks work in a similar fashion by using adversarial language models which can cause the NLP system to malfunction in certain ways [14]. A practical approach to performing such attacks is described by Neekhara et al. [15] in which they have trained a reinforcement learning agent that feeds carefully crafted input messages to the system resulting in undesired behaviour of the chatbot. Also pose a threat to chatbot security. In such attacks, the feedback network used by the chatbots to learn from user feedback is taken advantage of. A chatbot can diverge from its desired behaviour if a hacker keeps providing it with bad feedback thus compromising the chatbot. Zhang et al. [16] describes an attack on a system that uses reinforcement learning to learn from the feedback by interfering with the reward system of the reinforcement model. This makes the chatbot follow the learning policy that the hacker desires. Another form of feedback network is where the chatbot retrains itself to learn from the feedback provided. An attack on such a system works in a way such that some input words trigger malicious behaviour from the chatbot. However, this behaviour will not be evident enough for the user to detect [17]. Another area where chatbots can get compromised is highlighted by Zhang et al. [18] in which the attacker can directly target the training dataset of the agent in order to unlearn certain aspects of the data or learn new unwanted and harmful intents.

Machine Learning technology has been on rise due to its ability to make predictions, even more so accurately than humans in some cases [19]. It proves as an aide to human intelligence thus helping in complex decision making by giving mathematical analysis of the situation. The underlying principle of a machine learning scheme is to be able to train models based on the large amounts of the necessary data acquired [20]. This large amount of data needs to be secure if a third-party machine learning model is going to get access to it. Homomorphic encryption allowing computations on encrypted data is a feasible solution to this problem [21]. This has resulted in the possible practical use of homomorphic



L is defined as  $L(x) = \frac{x-1}{n}$   
 Public key is  $(n, g)$   
 Private key is  $(\lambda, \mu)$   
 In order to encrypt a message  $(0 \leq m < n)$ , select a random  $r(0 < r < n)$  and  $gcd(r, n) = 1$   
 The ciphertext

$$c = g^m \cdot r^n \text{ mod } n^2 \quad (5)$$

In order to decrypt the above ciphertext  $c$ , so as to obtain the plaintext  $m$  such that

$$m = L(c^\lambda \text{ mod } n^2) \cdot \mu \text{ mod } n \quad (6)$$

Fig 1. shows the public and private key used in this implementation.

We have used the 'phe' and 'dialogflow' Python packages for Paillier encryption and DialogFlow respectively. Using the dialogflow library, we detect if the triggered intent is the one that takes in the user's information. If so, the data input to the chatbot is encrypted using the user's public key. Fig. 2. show the data before and after encryption.

This encrypted data as well as the public key is now sent to the Linear Regression model on the cloud for prediction. The public key is extracted from the data. That leaves us with the

user-input data in the encrypted form which can now be used to compute the result. The result is computed by multiplying the encrypted data with the regression model coefficients and is summed together to form a single encrypted ciphertext. This single value is the predicted salary of the user in the encrypted form. Fig 3. shows the encrypted result after computation done by the machine learning model.

A dictionary with the public key and ciphertext result is now sent back to the chatbot where the received public key is matched with the user's public key. If there is a match, the ciphertext will be decrypted using the user's private key. Fig 4. shows the decrypted result that is displayed to the user.

The same inputs in plaintext form were tested on the machine learning model and the result was identical

In our proposal, we suggest homomorphic encryption be used only in scenarios involving extremely sensitive data like health records, medical prescriptions, financial data, etc. Since chatbots work on intent-triggering mechanisms, it is possible to identify the intents that use the aforementioned sensitive data and hence apply the homomorphic encryption technique only when required. The remainder of the chat will not be left exposed and will be encrypted with a different

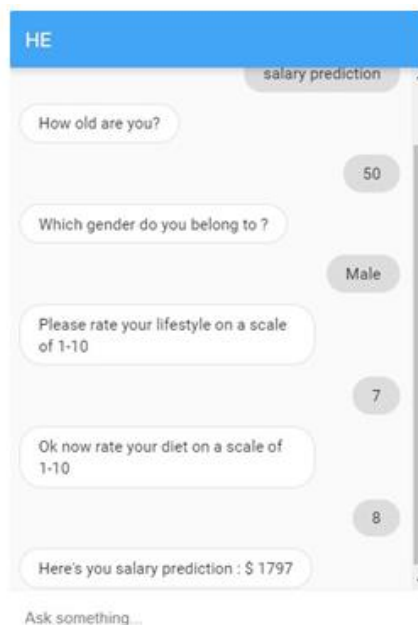


Fig. 4. The plaintext result in the chatbot after decryption

encryption mechanism so ensure maximum efficiency and security of the system. This system is not required in cases where users' personal data is not being shared or only information related to the services provided by the organisation is being served through the chatbot. Apart from the above implementation, a more advanced example of how the proposed system would work can be explained by taking a scenario

involving a healthcare chatbot which has a feature of predicting diseases based on the symptoms experienced by the user with the help of a third-party API. In such a case, the experienced symptoms suggested by the user to the chatbot needs to be given to the API over cloud for processing and making predictions. If the data is not homomorphically encrypted then the third-party has access to this sensitive data about the user. Hence, the data needs to be encrypted homomorphically using partial homomorphic techniques like RSA, ElGamal and Paillier encryption or use open-source libraries that provide fully homomorphic encryption from Microsoft and IBM as mentioned previously. Since the data is now homomorphically encrypted, it can be used by the third-party API without the need to be decrypted. The results provided by the API are also in the encrypted form and can be decrypted only by the user. If the data is encrypted in any other format, it leaves the data exposed at the third-party endpoint from where it can be compromised. If a hacker gets access to this exposed data, it can be used to perform phishing, baiting or MitM attacks on the user as he has access to the symptoms as well as the predicted ailment of the user. For example, a user uses this chatbot and complains of having trouble breathing. If this data is not homomorphically encrypted but uses a different encryption technique, two major scenarios can be possible. In the first scenario, the hacker gains access to this information at the third-party endpoint where it is decrypted for processing. Therefore, he

now has access to the symptoms experienced by the user as

#### IV. CONCLUSION

Cloud computing is a new paradigm which is revolutionizing the way in which computation over the internet takes place. According to the current trends and forecast, it is believed that by 2026 more than 67% of enterprises will switch to the cloud [33]. This implies that a customer service tool like chatbot used by these enterprises will also work over the cloud. Since the chatbot will have access to the cloud where all data is stored and will itself access and process user data, it needs to be made as secure as possible.

Social engineering attacks have gotten to a point where not even an organization's state of the art data center security systems can keep these attackers at bay. Given this significant improvement in the effectiveness of these attacks, enterprises too need to equip themselves with encryption standards that can truly mitigate such risks.

Apart from preserving the privacy of sensitive data, homomorphic encryption unlocks a suite of additional benefits. It indirectly promotes collaborative work by allowing use of third-party services without dwelling into the risk of data compromise. It also minimizes the loss incurred by organizations in terms of money and reputation as far as data breaches are concerned. The advances in research in the field of homomorphic encryption has bridged the gap between "theoretically possible" and "practically implementable" [34] thus finally allowing the use of this revolutionary technology in practice.

#### REFERENCES

- [1]. <https://www.gartner.com/en/newsroom/press-releases/2018-02-19-gartner-says-25-percent-of-customer-service-operations-will-use-virtual-customer-assistants-by-2020>
- [2]. <https://portswigger.net/daily-swig/i-chatbot-a-prime-target-for-cybercriminals>
- [3]. Bozic J., Wotawa F. (2018) Security Testing for Chatbots. In: Medina-Bulo I., Merayo M., Hierons R. (eds) Testing Software and Systems. ICTSS 2018. Lecture Notes in Computer Science, vol 11146. Springer, Cham. [https://doi.org/10.1007/978-3-319-99927-2\\_3](https://doi.org/10.1007/978-3-319-99927-2_3)
- [5]. Das, Rishabh Morris, Thomas. (2017). Machine Learning and Cyber Security. 1-7. 10.1109/ICCECE.2017.8526232.
- [6]. Workman, M. (2007). Gaining Access with Social Engineering: An Empirical Study of the Threat. Information Systems Security, 16(6), 315–331. doi:10.1080/10658980701788165
- [7]. Breda, Filipe Barbosa, Hugo Morais, Telmo. (2017). SOCIAL ENGINEERING AND CYBER SECURITY. 4204-4211. 10.21125/inted.2017.1008.
- [8]. Tom N. Jagatic, Nathaniel A. Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social phishing. Commun. ACM 50, 10 (October 2007), 94–100. DOI:<https://doi.org/10.1145/1290958.1290968>
- [9]. M. Conti, N. Dragoni and V. Lesyk, "A Survey of Man In The Middle Attacks," in IEEE Communications Surveys Tutorials, vol. 18, no. 3, pp. 2027-2051, thirdquarter 2016, doi: 10.1109/COMST.2016.2548426.
- [10]. T. Lauinger, V. Pankakoski, D. Balzarotti, and E. Kirda, "Honeybot, your man in the middle for automated social engineering." in USENIX.
- [11]. Vankudoth, Biksham Vasumathi, D.. (2017). Homomorphic Encryption Techniques for securing Data in Cloud Computing: A Survey. International Journal of Computer Applications. 160. 1-5. 10.5120/ijca2017913063.
- [12]. Altarawneh, Mokhled Al-Qaisi, Aws. (2019). EVALUATION OF CLOUD COMPUTING PLATFORM FOR IMAGE PROCESSING ALGORITHMS. Journal of Engineering Science and Technology.
- [13]. M. T. I. Ziad, A. Alanwar, M. Alzantot and M. Srivastava, "CryptoImg: Privacy preserving processing over encrypted images," 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, 2016, pp. 570-575, doi: 10.1109/CNS.2016.7860550.
- [14]. H. Liu, T. Derr, Z. Liu, and J. Tang, "Say what i want: Towards the dark side of neural dialogue models," arXiv preprint arXiv:1909.06044, 2019.
- [15]. X. Zhang, Z. Zhang, and T. Wang, "Trojanning language models for fun and profit," arXiv preprint arXiv:2008.00312, 2020
- [16]. P. Neekhara, S. Hussain, S. Dubnov, and F. Koushanfar, "Adversarial reprogramming of text classification neural networks," in ACL 2019.
- [17]. X. Zhang, Y. Ma, A. Singla, and X. Zhu, "Adaptive reward-poisoning attacks against reinforcement learning," arXiv preprint arXiv:2003.12613,
- [18]. Chen, Xiaoyi Salem, Ahmed Backes, Michael Ma, Shiqing Zhang, Yang. (2020). BadNL: Backdoor Attacks Against NLP Models.



- [19]. H. Zhang, T. Zheng, J. Gao, C. Miao, L. Su, Y. Li, and K. Ren, "Data poisoning attack against knowledge graph embedding," in *IJCAI 2019*
- [20]. Holzinger A., Kieseberg P., Weippl E., Tjoa A.M. (2018) Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. In: Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) *Machine Learning and Knowledge Extraction. CD-MAKE 2018. Lecture Notes in Computer Science*, vol 11015. Springer, Cham. [https://doi.org/10.1007/978-3-319-99740-7\\_1](https://doi.org/10.1007/978-3-319-99740-7_1)
- [21]. Qiu, Junfei Wu, Qihui Ding, Guoru Xu, Yuhua Feng, Shuo. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*. 2016. 10.1186/s13634-016-0355-x.
- [22]. Minelli, Michele. (2018). Fully Homomorphic Encryption for Machine Learning.
- [23]. Kahrobaei, Delaram [orcid.org/0000-0001-5467-7832](https://orcid.org/0000-0001-5467-7832), Wood, Alexander and Najarian, Kayvan (2020) Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics. *ACM Comput. Surv.* ISSN 0360-0300
- [24]. Yi, Xun Paulet, Russell Bertino, Elisa. (2014). Homomorphic Encryption and Applications. 10.1007/978-3-319-12229-8.
- [25]. Badawi AA, Chao J, Lin J, Mun CF, Jie SJ, Tan BHM, Nan X, Aung KMM, Chandrasekhar VR (2018) The AlexNet moment for homomorphic encryption: HCNN, the first homomorphic CNN on encrypted data with GPUs. *IACR cryptology ePrint archive*, vol 1056
- [26]. Chen H, Gilad-Bachrach R, Han K, Huang Z, Jalali A, Laine K, Lauter KE (2018) Logistic regression over encrypted data from fully homomorphic encryption. *IACR cryptology ePrint archive*, vol 462
- [27]. Graepel T, Lauter KE, Naehrig M (2012) ML confidential: machine learning on encrypted data. *ICISC* 1–21
- [28]. Kocabas O, Soyata T (2014) Private predictive analysis on encrypted medical data. *J Biomed Inform* 50:234–243. <https://doi.org/10.4018/978-1-4666-5864-6.ch019>
- [29]. Carpov S, Nguyen TH, Sirdey R, Costantino G, Martinelli F (2016) Practical privacy-preserving medical diagnosis using homomorphic encryption. *CLOUD* 593–599
- [30]. Bos JW, Lauter KE, Naehrig M (2014) Private predictive analysis on encrypted medical data. *J Biomed Inform* 50:234–243
- [31]. Peng H-T, Hsu WWY, Ho J-M, Yu M-R (2016) Homomorphic encryption application on FinancialCloud framework. *SSCI* 1–5
- [32]. Lu W, Kawasaki S, Sakuma J (2017) Using fully homomorphic encryption for statistical analysis of categorical, ordinal and numerical data, *NDSS*
- [33]. Papadimitriou A, Bhagwan R, Chandran N, Ramjee R, Haeberlen A, Singh H, Modi A, Badrinarayanan S (2016) Big data analytics over encrypted datasets with seabed. *OSDI* 587–602
- [34]. Chaudhary, Sanjay Somani, Gaurav Buyya, Rajkumar. (2018). *Research Advances in Cloud Computing*. 10.1007/978-981-10-5026-8
- [35]. Michael Naehrig, Kristin Lauter, and Vinod Vaikuntanathan. 2011. Can homomorphic encryption be practical? In *Proceedings of the 3rd ACM workshop on Cloud computing security workshop (CCSW '11)*. Association for Computing Machinery, New York, NY, USA, 113–124. DOI:<https://doi.org/10.1145/2046660.2046682>