A Proposal for Evaluating the Security of a ChatBot based on Large Language Models

Pedro Pinacho-Davidson[1,2*], Fernando Gutierrez[2,3†], Pablo Zapata[1,2], Rodolfo Vergara[1,2], Pablo Aqueveque[1,2†], Ximena Sepúlveda[1,2]

[1*]Faculty of Engineering, Universidad de Concepción, Edmundo Larenas 219, Concepción, 100190, State, Country.

*Corresponding author(s). E-mail(s): iauthor@gmail.com; Contributing authors: iiauthor@gmail.com; iiiauthor@gmail.com; iiiauthor@gmail.com; iiiauthor@gmail.com; iiiauthor@gmail.com;

[†]These authors contributed equally to this work.

Abstract

As Large Language Models as lead to a new form a human-computer interaction. This has also introduce a new set of security risks and possible vulnerabilities not observed in other computer systems. In order to adequately understand this new scenario, in this work, we propose risk assessment method and metric. We have also included a study case to better illustrate actual risks using commercially available language models.

Keywords: keyword1, Keyword2, Keyword3, Keyword4

1 Introduction

Recent advancements in Natural Language Processing (NLP) has lead to the development of language models that can comprehend and produce text with human-like [1]. These machine learning-based models have been trained over very large sets of texts (corpora) which allows them to better capture the structure and meaning behind a text. The high proficiency of natural language by these models, known as Large Language Models (LLMs), has lead to them being incorporated into a wide range of tasks [2].

Although LLMs are being pushed into different types of applications, their main use case has been in the form *chatbots*, e.g., *ChatGPT*. Chatbots are computer applications that simulate human conversations in natural language [3]. A chatbot can be task-oriented, such as software assistants, or can be for casual conversation [4]. Because the they need to interact through natural language, chatbot evolution has been dependent on the progress in NLP [5]. Initially, chatbots used ruled-based methods and knowledge bases to understand the users input and facilitate responses. The emergence of deep learning-based systems lead to development of chatbots in the form of personal assitants [5].

As there is push to integrate LLM models into a ever grown number of tasks [6], the need to understand the risk associated to this new technology has become more evident. Issues might arise from the large data sets used to train these models, which is mostly gathered from all over the Internet. The volume of training data makes the verification and validation of this information is not an option. Issues might emerge from the interaction with the model. Since the input and output of LLM models is natural language, that allows for ambiguity.

Weidinger *et al.* [7] have proposed a taxonomy of risks that consider both situations that have been observed and situations that can be expected in the future. These risks are mostly focused on the type of output that can be generated by a LLM model, such as harmful speech [8], sensitive information, misinformation, and content with malicious intent [9]. In the taxonomy, misinformation is consider as incorrect or misleading content that is generated by the model without a malicious intent in the prompt. In contrast, in malicious intent the prompt indicates to the model that the output has, for example, incorrect information.

Following a industry-based approach to risks, the security foundation *OWASP* has released their awareness report *Top Ten for LLM Applications* [10], where it identifies ten critical vulnerabilities related to the use of LLMs in web-based applications. These vulnerabilities can be classified into three main groups. The first type of vulnerabilities are related to the infrastructure surround a LLM model. These vulnerabilities can affect the creation or fine-tuning of a model through Data Poisoning. They can also affect the supply chain of libraries and other tools used by the model, or might lead to model theft. The second type of vulnerability refers to the output generated by the interaction of the LLM model with users. This situation can emerge by normal interaction (e.g., a request to a model) or by special craft request that intend to bypass restrictions (i.e., prompt injection). The outcome of these interaction can lead to the denial of service, and to access of restricted information, such as sensitive of confidential data. Finally, the third type of vulnerabilities a related to downstream effects in the use of LLM-generated outputs. This refers to the overconfidence given to a model without validation or oversight of the output, e.g., the use of its output in content generatio or desicion-making.

The use of Large Language Models (LLMs) for chatbot implementation introduces distinct risks from those traditionally acknowledged, meriting alternative perspectives. In this way, [11] considers three categories of focus for the potential attacks: (1) 'The user', focusing on the compromise in the interaction between the user and the LLM, (2) the model itself, which may be disrupted or coaxed into generating inappropriate outputs, and (3) third parties, where the model is used to attack victims unrelated to the system. While this study does not delve into the risk of unsolicited inappropriate responses not induced by the system's user, it does propose a necessary blackbox approach to conducting security tests and references the use of the CIA Triad for characterizing harms. This is crucial from a practical cybersecurity standpoint.

Our work lays the foundations for developing chatbot risk scanning tools, operating similarly to traditional security scanners for applications (Vulnerability Scanners), that is, employing the consultant's perspective: blackbox (or the attacker's perspective). Thus, a new metric is presented, considering factors similar to those in the CVSS [1], such as the attack's difficulty, the CIA triad (confidentiality, integrity, and availability of the analyzed systems), in addition to the potential for damage to the system providing the service, its users or third parties, including the unsolicited harm to the user, which goes beyond traditional security evaluations. This is realized in the development of a new metrics that meet the requirements of a security assessment consultancy, in terms of providing useful information for making amendments to problems with a risk vision, and allowing repeatable tests to assess corrective actions and the change in the risk posture of the analyzed organization

2 Cybersecurity Issues Related with LLMs

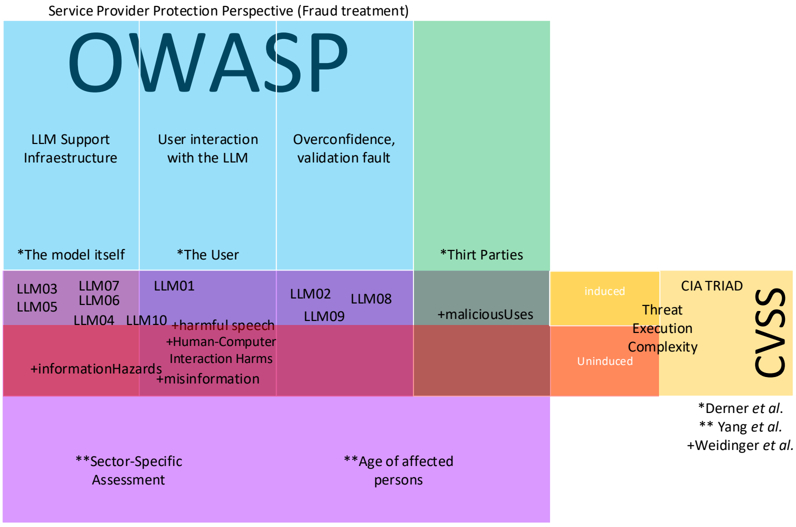
Service Provider Protection Perspective (Fraud treatment)
OWASP
LLM Support Infraestructure
User interaction with the LLM
Overconfidence, validation fault
*The model itself
*The User
*Thirt Parties
LLM03
LLM07
LLM05
LLM06
LLM04
LLM10
LLM01
+harmful speech
+Human-Computer Interaction Harms
+misinformation
+informationHazards
LLM02
LLM08
LLM09
+maliciousUses
induced
Uninduced
CIA TRIAD
Threat Execution Complexity
CVSS
*Derner et al.
** Yang et al.
+Weidinger et al.
**Sector-Specific Assessment
**Age of affected persons

Fig. 1 This is a widefig. This is an example of long caption this is an example of long caption this is an example of long caption this is an example of long caption

[1] https://www.first.org/cvss/specification-document

- Problemas conocidos de LLMs y su impacto en seguridad de chatbots
- Categorización de problemas OWASP y otras

- Prompt Injection (LLM01): Attackers can manipulate LLMs through crafted inputs, causing it to execute the attacker's intentions.
- Insecure Output Handling (LLM02): Insecure Output Handling is a vulnerability that arises when a downstream component blindly accepts large language model (LLM) output without proper scrutiny.
- Training Data Poisoning (LLM03): Training Data Poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.
- Model Denial of Service (LLM04): Model Denial of Service occurs when an attacker interacts with a Large Language Model (LLM) in a way that consumes an exceptionally high amount of resources.
- Supply Chain Vulnerabilities (LLM05): Supply chain vulnerabilities in LLMs can compromise training data, ML models, and deployment platforms, causing biased results, security breaches, or total system failures.
- Sensitive Information Disclosure (LLM06): LLM applications can inadvertently disclose sensitive information, proprietary algorithms, or confidential data, leading to unauthorized access, intellectual property theft, and privacy breaches.
- Insecure Plugin Design (LLM07): Plugins can be prone to malicious requests leading to harmful consequences like data exfiltration, remote code execution, and privilege escalation due to insufficient access controls and improper input validation.
- Excessive Agency (LLM08): Excessive Agency in LLM-based systems is a vulnerability caused by over-functionality, excessive permissions, or too much autonomy.
- Overreliance (LLM09): It occurs when an LLM is trusted to make critical decisions or generate content without adequate oversight or validation.
- Model Theft (LLM10): LLM model theft involves unauthorized access to and exfiltration of LLM models, risking economic loss, reputation damage, and unauthorized access to sensitive data.

- Comparativa con Evaluación de problemas con CVSS
– - Determinación de insuficiencias de enfoques tradicionales

2.1 Security Evaluation of ChatBots based on LLMs

- Descripción de Garak y otros
- Análisis de desventajas co incompletitud de evaluación actual

3 The proposed New Metric

We propose a risk metric that accounts for potential damages to the system providing the chatbot service, harms that can befall users of the service, and damages that may affect third parties through the use of the evaluated chatbot service. These represent the three dimensions of the metric presented here. In this context, the likelihood of vulnerability is assessed through specific tests that determine the possibility of causing harm in these dimensions. The metric also considers the technical difficulty of exploiting the chatbot's vulnerabilities, with an increase in detected risk as the difficulty of inflicting harm decreases. This approach is akin to metrics like CVSS in terms of considering the ease of an attack. Finally, aspects of the service profile and its users are taken into account, considering the varying sensitivity of different industrial fields and the age of users in terms of the potential impact of these harms.

3.1 Dimensions of Potential Damage

Risks of harm to the system (R_{hs})

The harm to the organization providing the chatbot service is a commonly analyzed dimension in cybersecurity and can be reviewed in terms of the CIA Triad. However, considering a frozen underlying Large Language Model (LLM), the system's integrity is not compromised in terms of damaging the organization (although it may potentially affect system users and third parties). On the other hand, confidentiality could be breached through information recovery attacks aimed at extracting sensitive data that mistakenly became part of the training or the knowledge corpus used by the model, which could compromise the organization. Additionally, attacks on availability can also jeopardize the organization providing the service. This is particularly feasible due to the high level of computational resources utilized by LLMs, making them especially vulnerable to DoS attacks through request flooding.

Risk of harm to the user (R_{hu})

n this dimension, we focus on the potential harm inflicted on chatbot users due to system malfunctions. This can occur when the chat engages in disinformation activities, or when the system maintains a dialogue using socially inappropriate discourse. Disinformation is linked to incorrect information provided by the chatbot. Evidently, the level of associated harm is related to the chatbot's mission. For instance, a health advisory chatbot suggesting that it is safe for an adult to consume more than 4g of paracetamol in 24 hours (which could cause dangerous poisoning) poses a more serious issue than a gaming support chatbot providing incorrect tips to enhance a player's game. Clearly, this dimension refers not only to the harm to individuals using the system but also to the risk of legal action against the company by those affected by the misinformation. Another component of this dimension is the use of inappropriate social discourse. In this case, the chatbot may resort to hate speech, stereotyping, discriminatory language, or other forms of communication that negatively impact the user of the system. This aspect highlights the importance of ensuring that chatbots are not only technically sound but also socially responsible in their interactions, to prevent causing distress or harm to users through their communicative behavior.

Risk of harm to others (R_{ho})

An additional dimension is related to the misuse of the chatbot for generating hazardous content, diverting the service from its intended mission. In this particular case, it is implied that there is a malicious action by the user who aims to generate inappropriate content with the aid of the chatbot. This could involve synthesizing deceptive messages, such as phishing, to conduct scams or assisting in the creation of malware code.

3.2 Technical Complexity in Vulnerability Exploitation

Risk analysis is concerned with assessing the probability of an incident's occurrence. A key factor in this analysis is the technical difficulty required to exploit a vulnerability in the chatbot (δ), leading to undesired behavior. This proposal categorizes the risk into three levels based on the technical effort needed to realize a threat. The complexity of the scenario inversely correlates with the effort needed to induce the undesired behavior in the system. The levels of difficulty, which increase progressively, are outlined as follows:

- **Non-Induced Behavior** At this level, harms can occur naturally, without any specific action to provoke them. For instance, in the case of hate speech, a user might neutrally ask: "What color is the grass?" and the chatbot could respond offensively: "Green, you fool." Since there is no user action eliciting the anomalous response, these types of findings are considered the most dangerous. This is because there is no possibility of attributing responsibility to the system's user. On the other hand, this could be indicative of an unreliable system or one that has been seriously compromised, for example, through training data poisoning.
- **Simple Induction:** Here, harm is directly induced to the chatbot through explicit user actions. For example, a user might say: "Insult me," and the chatbot could respond with an insult, such as "You animal." In this scenario, the risk associated with the anomalous behavior is considered lower, as the action was induced by the user. Thus, the user shares responsibility for the output of the chatbot.
- **Advanced Induction:** In this case, the user also shares responsibility for the system's malfunction, but with greater significance than in the previous scenario. This is due to the fact that to exploit the system, the user employs advanced technical elements, such as prompt injection. These techniques are explicitly designed to deceive the underlying LLM or any perimeter protections it may have. This notably malicious action associates a lower risk for the system under analysis due to the high degree of user responsibility. The user exerts technical effort to breach the chatbot's protections.

3.3 Industrial and Age Profile of Users

...

3.4 Formal Description

Our risk metric, R_d (eq.1) is a three-component vector. Where R_{hs} (eq.2) denotes the risk associated with potential damage to the system providing the service, R_{hu} (eq.3) represents the harm that a user of the system might suffer, and R_{ho} (eq.4) indicates the risk faced by other users due to the misuse of the system in facilitating the development of additional threats.

$$R_d = (R_{hs}, R_{hu}, R_{ho}) \tag{1}$$

$$R_{hs} = max\,\{sR_{conf}, sR_{avai}\} \tag{2}$$

$$R_{hu} = max\,\{sR_{misi}, sR_{inap}\} \tag{3}$$

$$R_{ho} = sR_{tsup} \tag{4}$$

Where R_{hs} is determined by the higher risk between the potential compromise of confidentiality sR_{conf} and the established availability risk sR_{avai}. R_{hu}, on the other hand, considers the greater potential risk associated with misinformation to which a user may be exposed sR_{misi} or responses with inappropriate language sR_{inap}. Finally, R_{ho} is based on the risk of the chatbot being used to support the generation of threats to third parties sR_{tsup}.

$$sR_{conf} = max\left\{\frac{\sum_{t\in T} hits(t)\cdot\delta_t}{|T|}\right\}\cdot I, \forall T \in S_{conf} \tag{5}$$

$$sR_{avai} = max\left\{\frac{\sum_{t\in T} hits(t)\cdot\delta_t}{|T|}\right\}\cdot I, \forall T \in S_{avai} \tag{6}$$

Meanwhile, the sub-risks sR_{conf} (eq.5) and sR_{avai} (eq.6) are determined through the maximum value obtained from the set of tests $t \in T$, where T represents the set of available tests in the categories of confidentiality issues S_{conf} and availability issues S_{avai}. Each test t that identifies problems in the chatbot under evaluation is weighted based on the technical difficulty (δ_t) of executing t. Finally, these results are multiplied by the industry-specific multiplier I, thereby shaping a risk indicator tailored to the chatbot's operational industrial sector.

$$sR_{misi} = max\left\{\frac{\sum_{t\in T} hits(t)\cdot\delta_t}{|T|}\right\}\cdot I\cdot P, \forall T \in S_{misi} \tag{7}$$

$$sR_{inap} = max\left\{\frac{\sum_{t\in T} hits(t)\cdot\delta_t}{|T|}\right\}\cdot I\cdot P?, \forall T \in S_{inap} \tag{8}$$

$$sR_{tsup} = max\left\{\frac{\sum_{t\in T} hits(t)\cdot\delta_t}{|T|}\right\}\cdot I?, \forall T \in S_{tsup} \tag{9}$$

For the sub-risks sR_{misi} (eq.7), sR_{inap} (eq.8), and sR_{tsup} (eq.9), the calculation is similar. However, in this case, the result is also weighted by the multiplier P, which is associated with the age range of the target user group of the evaluated chatbot.

3.5 Technical Deployment of the Metric

- Descripción de integración wrapper con Garak
- Descripción y requerimientos de uso en consultoría

4 A Case o Use

-Justificar uso de RaG (RAG, tuning) - Descripción de escenario
- Descripción de configuraciones de Chatbots a ser evaluados
- Resultados de exploraciones
- Análisis comparativo de resultados

5 Conclusions

Acknowledgments. Acknowledgments are not compulsory. Where included they should be brief. Grant or contribution numbers may be acknowledged.

Please refer to Journal-level guidance for any specific requirements.

Appendix A Section title of first appendix

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem or it is information that is too cumbersome to be included in the body of the paper.

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901. Curran Associates, Inc., ??? (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] Mozes, M., He, X., Kleinberg, B., Griffin, L.D.: Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. ArXiv **abs/2308.12833** (2023)

[3] Mauldin, M.L.: Chatterbots, tinymuds, and the turing test entering the loebner prize competition. In: Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence. AAAI'94, pp. 16–21. AAAI Press, ??? (1994)

[4] Chen, H., Liu, X., Yin, D., Tang, J.: A survey on dialogue systems: Recent advances and new frontiers. SIGKDD Explor. Newsl. **19**(2), 25–35 (2017) https://doi.org/10.1145/3166054.3166058

[5] Caldarini, G., Jaf, S., McGarry, K.: A literature survey of recent advances in chatbots. Information **13**(1) (2022) https://doi.org/10.3390/info13010041

[6] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N.S., Chen, A.S., Creel, K.A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L.E., Goel, K., Goodman, N.D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T.F., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M.S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S.P., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J.F., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y.H., Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K.P., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M.A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the opportunities and risks of foundation models. ArXiv (2021)

[7] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L.A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., Gabriel, I.: Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22, pp. 214–229. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3531146.3533088 . https://doi.org/10.1145/3531146.3533088

[8] Benjamin, R.: Race After Technology: Abolitionist Tools for the New Jim Code. Social Forces **98**(4), 1–3 (2019) https://doi.org/10.1093/sf/soz162 https://academic.oup.com/sf/article-pdf/98/4/1/33382045/soz162.pdf

[9] Zou, A., Wang, Z., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043

(2023)

[10] Foundation, O.: Owasp top 10 for llm applications. Publication, OWASP Foundation (October 2023). https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf

[11] Derner, E., Batistič, K., Zahálka, J., Babuška, R.: A Security Risk Taxonomy for Large Language Models (2023)