# Risk Assessment of Large Language Models Beyond Apocalyptic Visions

**Clara Maathuis[1] and Sabarathinam Chockalingam[2]**
[1]Open University of the Netherlands, Heerlen, Netherlands.
[2]Institute for Energy Technology, Halden, Norway.

clara.maathuis@ou.nl
sabarathinam.chockalingam@ife.no

**Abstract:** The remarkable development of Large Language Models (LLMs) continues to revolutionize various human activities in different societal domains like education, communications, and healthcare. While facilitating the generation of coherent and contextually relevant text across a diverse plethora of topics, LLMs became a set of instruments available in different toolboxes of decision makers. In this way, LLMs moved from a hype to an actual underlying mechanism for capturing valuable insights, revealing different perspectives on topics, and providing real-time decision-making support. As LLMs continue to increase in sophistication and accessibility, both societal and academic effort from AI and cyber security is projected in this direction, and a general societal unrest is seen due to their unknown consequences. Nevertheless, an apocalyptic vision towards their risks and impact does not represent a constructive and realistic approach. Contrarily, this could be an impediment to building LLMs that are safe, responsible, trustworthy, and have a real contribution to the overall societal well-being. Hence, understanding and addressing the risks of LLMs is imperative for building them in an ethical, social, and legal manner while making sure to consider control mechanisms for avoiding, mitigating, accepting, and transferring their risks and harmful consequences. Taking into consideration that these technological developments find themselves in an incipient phase, this research calls for a multi-angled perspective and proposes a realistic theoretical risk assessment method for LLMs.

**Keywords**: Artificial Intelligence, AI Risks, Large Language Models, Risk Assessment, Security, Privacy.

## 1. Introduction

"The only limit to our realization of tomorrow will be our doubts of today." (Franklin D. Roosevelt)

A significant milestone in the Artificial Intelligence (AI) domain is the development of the transformer model (Vaswani et al., 2017) that serves as basis for well-known LLMs (Large Language Models) such as ChatGPT. Through LLMs and in combination with other Generative AI models (e.g., diffusion), an accelerating trend towards increasingly sophisticated language understanding as well as video, audio, and image processing and creation, for fostering innovation is seen across various domains. Nevertheless, building and using LLMs raises ethical, social, and legal risks related to aspects like transparency, dis/misinformation, security, privacy, and fairness. Recognizing these issues is crucial for building responsible and trustworthy LLMs systems (Mikalef et al., 2022; EU Commission, 2019). This represents the first step for building risk assessment and management mechanisms for LLMs as they further support the improvement of strategic planning and decision-making, highlight common and cross-cutting risks, optimize resource allocation, and assist legislative/governing bodies (UN, 2020). After recognizing the risks, their evaluation and analysis should be conducted to establish proper treatment for avoiding, controlling, or mitigating them. This represents a difficult process given the complexity, uncertainty, and multifaceted nature of LLMs and their application context. Concurrently, this is a pressing challenge in technical and governance terms which needs to be addressed through joint research, multi/inter/transdisciplinary perspective, and collaborative practitioner efforts given the implications of their action. This further supports the development of LLMs audit for assuring a systematic, proactive, and transparent development, use, and facilitation of awareness and training solutions on their potential risks and impact (Mökander et al., 2023).

While academic and practitioner efforts on building risk assessment and management frameworks and mechanisms for AI systems exist, they are in an incipient stage for Generative AI systems. In particular, for LLMs, this represents the knowledge gap that this research aims to tackle as LLMs-based systems do not only instantiate existing socio-technological risks, but also pose new ones that need to be addressed. These risks that account both technical and human-related aspects can be directly seen in activities taken, for instance, by military experts when building, executing, and assessing military (cyber) operations, and by cyber security field experts involved when building solutions for preventing and responding to various cyber security incidents. Hence, this research aims to propose a theoretical risk assessment method for LLMs by adopting a multidisciplinary stance for merging knowledge from the AI, ethics, cyber security, and risk management domains. This implies conducting a comprehensive literature review on academic and practitioner resources while considering compliance with existing AI and risk management standards and instruments like (ISO/IEC,

2008; ISO/IEC, 2023; NIST, 2023; EU Parliament, 2023). This research stresses the need for assessing and addressing risks in all the life cycle phases of LLMs in a transparent manner for making sure that the LLMs are built and used in a safe, responsible, and trustworthy way.

The remainder of this article is structured as follows. Section 2 presents the context of this research together with relevant academic and practitioner studies carried out in this domain. Section 3 presents the risk assessment method proposed in this research. At the end, Section 4 discusses concluding remarks and future research perspectives.

## 2. Background and Related Research

A groundbreaking deep learning model architecture was introduced in 2017 (Vaswani et al., 2017): the transformer model which marks a new era in the AI domain. This model relies on the self-attention mechanism and feed-forward neural networks to process sequential data, is suitable for various Natural Language Processing (NLP) tasks (Shanahan, McDonell, & Reynolds, 2023), and represents the foundation of state-of-the-art models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), especially through broadly used models like RoBERTa, GPT 3.5, and GPT 4. This mechanism implies that a model is pre-trained on a vast corpus for a primary task and is subsequentially fine-tunned on a second task for transferring knowledge gained from one task to another, helping to deal with the challenges posed by limited labelled data. Due to their mediatic visibility and exponential use, these models developed into LLMs that conquered attention of a significant part of the academic and industry efforts that are working on their development, deployment, and use for assuring a responsible digital behaviour (Chang et al., 2023; Maathuis & Chockalingam, 2023). However, LLMs pose a series of challenges, issues, and risks that are addressed in this section considering a socio-technical stance for capturing relevant social, ethical, and legal dimensions. These are collected based on a comprehensive literature review using search keywords like *LLMs, risk assessment, LLMs security, and LLMs privacy*. The review was conducted in the IEEE Digital Library, ACM, Scopus, Wiley, Google Scholar scientific databases and in public governance available resources from relevant global stakeholders like UN, EU Commission, and NIST.

The UN AI advisory Group (UN, 2023) proposed the following five guiding principles for building AI systems for humanity: AI should be governed inclusively; AI must be governed in the public interest; AI governance should be built in step with data governance and the promotion of data commons; AI must be universal, networked, and rooted in adaptive multi-stakeholder collaboration; AI governance should be anchored in the UN charter, International Human Rights Law, and other agreed international commitments like the Sustainable Development Goals. At the same time, the group stresses that AI also implies issues/risks like bias, surveillance expansion, automated decision-making that blurs accountability of public officials, stressing that such risks can manifest globally, which calls for dedicated mechanisms to mitigate and avoid them. Nevertheless, such mechanisms need to be developed with care to prevent harming the further development of AI technologies. Seeing the agreement to adopt the AI Act, (EY, 2023) stresses the importance of assessing the risk of AI systems, classifying AI systems as: prohibited (systems that pose a risk to people's safety, security, and fundamental rights), high-risk (systems that carry the majority of compliance obligations, including the establishment of risk and quality management systems, data governance, human oversight, and cyber security measures), and minimal risk (systems that do not impose additional obligations next to the initial risk assessment and transparency requirements). While the risk approach is relatively novel in the AI domain, it has a rich history in the field of cyber security. For instance, in the UN Security Management Manual (UNSMS, 2017), the risk matrix is used to assess possible risks taking into consideration the likelihood and impact components. The likelihood levels contain very unlikely, unlikely, moderately likely, likely, and very likely, while the impact component contains the negligible, minor, moderate, severe, and critical levels. When building risk assessment solutions, NIST argues (NIST, 2018) that these should be done together for security and privacy aspects of the systems assessed. SANS (SANS, 2021) relates the level of severity of vulnerabilities with the risks from minor severity, moderate severity, high severity with minor exposure, moderate exposure, and high exposure. A well-known standard in the security risk management domain is ISO/IEC 27001 where the risk assessment is the core part of the risk management framework that includes risk identification, risk analysis, risk estimation, and risk evaluation phases. Once the risks are addressed, risk treatment measures can be applied (ISO/IEC, 2008).

The risk management process associated with AI systems by ITU implies the following steps: (i) risk assessment where the assets are considered together with their value and utility, threats and vulnerabilities are associated with these assets, risk of exposure of these assets to the threats and vulnerabilities is considered, and the risk and impact resulting from the risk of exposure is determined; (ii) risk treatment; (iii) risk management decision;

(iv) risk monitoring, review, and communication; (vi) update and improvement of risk controls. NIST proposed the AI Risk Management Framework (NIST, 2023) that contains four components: govern by cultivating and adopting a culture of risk management, map where the context is recognized and risks related to the context are identified, measure where the identified risks are assessed, analysed, and tracked, and manage where risks are prioritized and acted upon based on the projected impact. The ISO/IEC 23894 on guidance for risk assessment for AI considers the following risk management principles: risk management is an integral part of all organizational activities; a structured and comprehensive approach to risk management contributes to consistent and comparable results; the risk management framework and process are customized and proportionate to the organization's external and internal context related to its objectives; appropriate and timely involvement of stakeholders that enables their knowledge, views, and perceptions to be considered; risk can emerge, change, or disappear as an organization's external and internal context changes; the inputs to risk management are based on historical and current information, as well as on future expectations; human behaviour and culture significantly influence all aspects of risk management at each level and stage; and risk management is continually improved through learning and experience (ISO/IEC, 2023). From a general standpoint, Somer & Thalmann (2023) consider the following four design requirements for building AI risk management mechanisms: identification of AI use cases, evaluation and adaptation of existing risk management frameworks, risk management and quantification, and the consideration of legal aspects.

While conducting a systematic mapping study for concrete AI risk assessments, Xia (2023) considers that the Responsible AI principles can be seen as quality metrics for operationalizing the risk assessment process for AI systems. Moreover, they author identifies the following categories of stakeholders at risk: industry-level stakeholders (e.g., AI technology producers/procurers, AI solution producers/procurers, and AI users/consumers), organization-level stakeholders (e.g., board members, executives, and managers), and team-level stakeholders (e.g., product managers, project managers, and team leaders). CLTC (2021) introduces a guidance framework for building risk and impact assessment mechanisms for AI systems including the following factors: data related (its sensitivity, appropriateness, and timeliness), nature of impact (of the potential impacted parties, of the impact on the affected parties, potential safety risks, individual fundamental and legal rights, individual physical or mental well-being, individual economic stability, ecological impact, whether the overall effect of the impact is positive or negative), scale of the impact (number of individuals affected, severity of the impact), are the harmful effects permanent (reversibility of the effects), likelihood of harm (likelihood of the impact occurring), role of the system in making decisions, transparency of the system (explainability, auditability). Based on these factors, the guidance framework proposes to assess the number of riskiness level which AI systems are classified to. WEF (2020) built an AI project risk assessment tool that focuses on the planning phase of an AI procurement including the following five categories of issues: data, field of use, socio-economic impact, financial consequence for agency and individuals, and business function of the AI system. Schuett (2023) discusses the phases of conducting a risk management in the EU AI Act. The first phase deals with the identification of known and foreseeable risks (Article 9(2), sentence 2, point (a)), the second phase is estimation and evaluation of risks that may emerge from intended uses or foreseeable misuses, or risks that have identified during post-market-monitoring (Article 9(2), sentence 2, points (b), (c)), and the third phase is the adoption of risk management measures (Article 9(2), sentence 2, point (b)). Building risk assessment methods for AI systems also requires defining valuable metrics to be considered. To this end, Maathuis, Pieters, & Van den Berg stress the importance of assessing the risks and impact of intelligent systems on systems, data, humans, and processes; and Piorkowski, Hind, & Richards (2021) define the following attributes of risk assessment metrics: reliability, validity, significance, applicability, monotony, understandability, explainability, and context-awareness.

Mökander et al. (2023) propose a three-layered audit framework for LLMs that focuses on governance audit of technology providers, model audit of the LLMs after the pre-training phase and prior to their release, and application audit for the applications based on LLMs. Weidinger et al. (2021) identify the following categories of socio-ethical risks of LLMs: (i) discrimination, equity, and toxicity; (ii) information hazards; (iii) misinformation harms; (iv) malicious uses; (iv) human-computer interaction harms; (v) automation, access, and environmental harms. For ChatGPT, Zhou et al. (2023) identify the following ethical issues: the existence of bias, privacy and security attacks, transparency, abuse, and authorship issues. Fundamental values like transparency and safety are values addressed by models like LLaMA 2 (Touvron et al., 2023) and BLOOM (Workshop et al., 2022). The groups that introduced them shared details about the dataset used for training, challenges faced in the development process, and the evaluation metrics used. IEAI (2022) builds a practical risk assessment method for AI systems structured in three phases: risk self-assessment and quantification, risk visualization, and risk categorization. In the first phase, the reaction urgency index calculation is based on risk intensity that contains

the prevalence, magnitude, and probability of the risks; and reduction demand that includes proximity, social discourse, and temporal immediacy. In the second phase, the risk reaction urgency matrix is based on associations made to all the seven requirements considered for building trustworthy AI systems. In the third phase, the risk potential categorization is used for the overall AI application based on the proposed previous calculations, and considering four dimensions: low reaction demand, high reaction demand, low-risk intensity, and high-risk intensity. From a practical standpoint, TechLaw (2021) proposes the RAIIA (Responsible Artificial Intelligence Impact Assessment) framework which considers six risk levels (zero, very low, low, medium, high, and very high) while recommending the application of the impact assessment framework starting with the medium risk level. Furthermore, the categories of risk are governance, people, process, and technology, calling further for defining dedicated risk mitigation strategies for each of them based on their likelihood and impact. Next to the general risk assessment and management perspectives, Khlaaf (2023) proposes the introduction of the notion of ODD (Operational Design Domain) which represents the description of the domain where the AI system is designed considering as core elements, application, users/agents, protected characteristics, and assets. Moreover, a taxonomy of security risks for LLMs is proposed by (Derner et al., 2023) where the entities involved in an attack are the user, the model, and a third party. Moreover, the security risks are considered in relation to the CIA (confidentiality, integrity, and availability) triad in an experimental setting.

To this end, Kaddour et al. (2023) address those risks of LLMs in relation to three core dimensions: design due to the use of unfathomable datasets, tokenizer-reliance, and fine-tunning overhead; science behind them due to evaluations made based on static human-written ground truth, lacking experimental design, and lack of reproducibility; and behaviour through prompt brittleness, misaligned behaviour, and outdated knowledge. Liu et al. (2023) structure the risks of LLMs according to the principles of building AI systems: they associate misinformation and hallucination to reliability, unlawful conduct and privacy violation to safety, preference bias and disparate performance to fairness, cyber-attack misuse and social engineering misuse to resistance to misuse, lack of interpretability and limited causal reasoning to explainability and reasoning, toxicity and cultural insensitivity to social norms, prompt attacks and poisoning attacks to robustness. Zhao et al. (2023) reflect on the meaning and implications of transparency for LLMs pointing the importance of building explainability AI methods for elucidating their behaviours, limitations, and social impact. Deldjoo (2023) addresses fairness for assuring that the models are not only transparent to users, but that they are not biased towards a specific category. Staab et al., (2023) explore the privacy risk through experiments that show personal data leakage even if they only imply seemingly benign prompt questions. Guo et al., (2022) also assess the integrity risk through possible backdoors. For ChatGPT, Wu, Duan, & Ni (2023) stress the fact that ChatGPT is introducing new threats to its users and the public that call directly for heightened vigilance and protective measures. The authors focused on security, privacy, and ethical issues of ChatGPT discussing risks and threats like malware creation, hallucination, and propaganda threat. Begou et al., (2023) conduct an experiment for developing a phishing attack focusing on cloning a targeted website, integrating code for stealing credentials, obfuscating code, automating website development on a hosting provider, registering a phishing domain, and integrating the website with a reverse proxy. Along these lines, Sison et al., (2023) consider the following categories of ethical risks for ChatGPT: possibility of being sentient or displaying sentience, privacy, bias, employment, and automation as job displacement can occur, and social media and public discourse through the creation of echo-chambers, the production of emotional contagion, exploitation of psychological vulnerabilities and risk of manipulation. Stahl & Eke (2024) structure the ethical risks of ChatGPT as social justice and rights (e.g., fairness, freedom of speech and repression, harm to society), individual needs (e.g., safety, autonomy, psychological harm), culture and identity (e.g., cultural difference, discrimination, and social sorting), and environmental (e.g., sustainability, environmental harm). Wach et al. (2023) see as core threats and risks of LLMs the urgency for building and adopting AI regulation; lack of quality control, disinformation, deepfake content, algorithmic bias; automation-spurred job issues; personal data violation, social surveillance, and privacy violation; social manipulation, weakening ethics and goodwill; widening socio-economic inequalities; and technostress.

As this comprehensive literature review reveals, various valuable dimensions, and concepts to building risk assessment methods for LLMs were addressed at scientific and practitioner levels. Nevertheless, a unified structured method that considers socio-technical risks applicable to all the phases of the AI's life cycle that benefits from previous research done not only in the AI domain, but also in the cyber security domain, is lacking. This would be beneficial to LLMs systems given their complex and uncertain nature and represents the knowledge gap that this research aims to tackle.

## 3. Risk Assessment Method

To ensure that LLMs systems are built in a safe, responsible, and trustworthy manner, this article stresses the need for conducting the risk assessment of LLMs across all their life cycle phases, i.e., starting from their definition and data collection phases and going to their deployment and monitoring phases. Such a comprehensive approach is imperative to capture, identify, and evaluate corresponding risks that manifest in different ways and forms at various stages. This in turn would allow to be able to properly deal with potential adverse implications and consequences of their action. Accordingly, the risk assessment method proposed in this research is designed considering the life cycle phases of LLMs captured in Figure 1 and defined as follows (UN, 2023; NIST, 2023; Mökander et al., 2023; Xia et al., 2023; Maathuis et al., 2023; Abbas, 2023; Maathuis, 2022; Pamula, 2023).

The LLMs life cycle contains the following phases:

- Model definition: The goals and requirements of the model are established. These include determining specific tasks to be accomplished, considering possible architectural components, establishing working and performance parameters.
- Data collection and pre-processing: Relevant data is collected, pre-processed by cleaning, formatting, and transforming it as preparation for training and validating the model.
- Model development: The working model architecture is established as part of the design that is further developed using the dataset(s) prepared in the previous phase. Further, the model is trained to learn patterns and tasks from the available data and context meaning, validated for assessing model's performance on unseen data, and fine-tunned to optimize the parameters for improving the results obtained.
- Model deployment: After the model is developed and validated, the model is deployed in the production environment so that it is accessible to end-users. To this end, aspects that need to be considered are real-time performance monitoring, efficiency, and scalability.
- Model maintenance and update: Ongoing continuous model maintenance and preparations for potential future necessary model update(s) are carried out in relation to real-world conditions by arising the issues that arise. This could include re-training the model with new available data for assuring its relevance and effectiveness in evolving contexts.

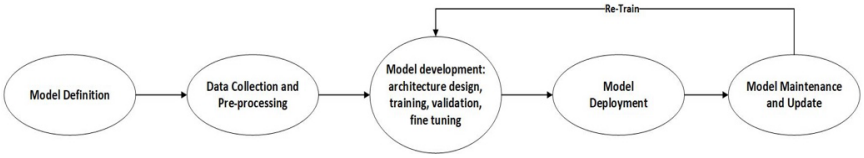


**Figure 1: LLMs lifecycle**

The risk assessment method proposed in this research is compliant with the ISO/IEC 27001 and 23894 standards and is transparent to its users. The method is illustrated in Figure 2 and contains the following phases:

- *Risk Identification:* Implies for each phase of the AI life cycle, identifying corresponding risks for the assets that can be posed at risk (humans as individuals or collective, organizations, systems or processes that embed the LLMs-based model, data used for building, evaluating, fine tuning, or querying the model), how the risk is defined, the source of risk, the risk metrics defined to assess when and how the risks occur and what is their duration, associated vulnerabilities and recognizing the potential impact. In this process, the following information is gathered (see Table 1):

**Table 1: Risk Identification phase**

| Asset at Risk Category | Risk Definition | Risk Source | Risk Metrics | Associated Vulnerability | Associated Impact |
|---|---|---|---|---|---|
| Human | | | | | |
| Organization | | | | | |
| System/Process | | | | | |
| Data | | | | | |
| Model | | | | | |

For instance, in this phase two categories of data privacy and model security risks are identified when building an LLM-based model for decision-making support for military or law enforcement purposes. In this sense,

unauthorized access and potential misuse could occur on sensitive data which represents a direct threat to personal information used in the training data of the model.

- *Risk Analysis:* Once that the risks are identified, they can be further analysed in a qualitative or quantitative way based on their perceived severity and likelihood or involving assigning numerical values to the probability and impact of risks, respectively using the risk analysis matrix captured in Table 2. This implies analysing factors like the likelihood of a specific threat exploiting an existing vulnerability and the associated impact on humans if the risk materializes. This produces a prioritization of lists and further considerations to be addressed in the assessment process.

**Table 2: Risk analysis matrix**

| Likelihood/Severity | | | | | |
|---|---|---|---|---|---|
| Very High | | | | | |
| High | | | | | |
| Moderate | | | | | |
| Low | | | | | |
| Very Low | | | | | |
| | Very Low | Low | Moderate | High | Very High |

For the same example, the data privacy is analysed in relation to personal data being leaked and the model security in relation to its resilience against potential security attacks. For instance, adversarial attacks could be carried out on the model by malicious actors to manipulate input data and produce erroneous outputs, potentially leading to incorrect decisions in critical settings.

- *Risk Evaluation:* The risks identified are analysed considering the goal of the AI system, socio-ethical and legal considerations, plus the source and metrics adopted in relation to the goals of the AI system that poses risks and associated vulnerabilities and potential impact. In this process, a risk score is calculated considering internal and external risks, and two important variables are defined: (i) risk criteria which point to predefined factors that support decision-makers for taking informed decisions about the significance and impact of the risk assessed; and (ii) the risk threshold to establish the predefined level or boundary of risk that is acceptable or tolerated for the AI system.

  In the above-mentioned example, both risks are assessed at very high level of severity as it could directly imply harm and damage to data and people involved, issues to security, privacy, safety, robustness, transparency, responsibility, and reputation damage to the stakeholders involved.

- *Risk Decision:* Once the risks are identified, analysed, and evaluated, informed decisions are made based on the risk score obtained which informs how the risks need to be handled in relation to their potential consequences. From this, the initial decision is made to establish if the risks require treatment or can be directly accepted. In case that the risks require treatment, the most appropriate mechanisms for managing, mitigating, avoiding, and responding to them, are defined, adopted, and further communicated and documented.

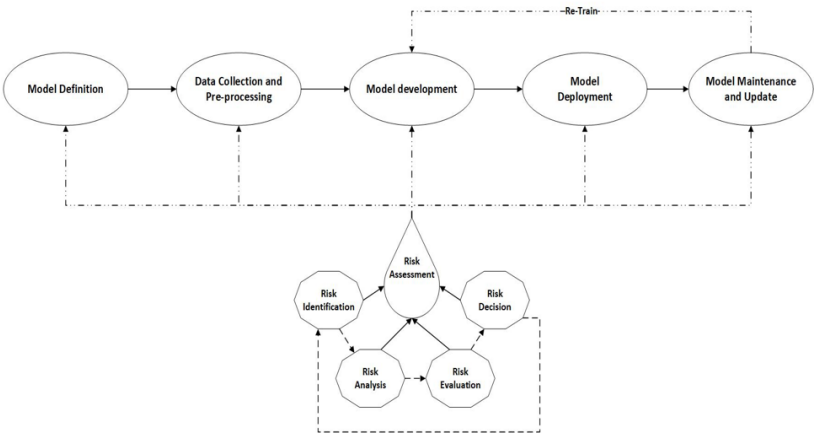


**Figure 2: Risk Assessment Method for LLMs**

For the same above-mentioned example, given the severity level of the risks considered and potential impact, it could be decided that risk treatment is necessary. This implies that corresponding mechanisms and measures for avoiding, mitigating, controlling, and responding to such risks need to be considered. In this sense, among the measures that could be included is incorporating adversarial training techniques to enhance the model's resilience against adversarial attacks and ensuring fairness and transparency in the decision-making process to mitigate the risk of biased or erroneous outcomes.

## 4. Conclusions

The multifaceted nature of AI techniques and technologies introduces unique challenges and issues, often captured through associated risks like potential biases, lack of transparency, and potential harm to society. As the implications and consequences of AI systems are experienced at all phases of their life cycle, thus from their design up to their deployment and use, it is important to have a comprehensive understanding of their risks so that the stakeholders involved and/or the ones impacted by their action can foster responsible and trustworthy AI practices by directly addressing them when necessary. The first step in this direction is building transparent risk assessment mechanisms that represent the basis for establishing corresponding risk evaluation and treatment decisions, and demonstrate a commitment to accountability, ethical use, and audit of AI systems.

To effectively address this fact, it is important to build risk assessment and risk management mechanisms that are tailored to specific AI specific paradigms and (set of) techniques given their complex nature. To this end, given the increase in development and use of LLMs in various societal domains, and aiming at preventing confusion or mis/disinformation (Fard & Maathuis, 2021), based on a comprehensive literature review conducted on scientific and professional literature, this research acknowledges that the ongoing efforts for building risk assessment and management mechanisms are in an incipient stage, but much needed. Hence, adopting the perspective of learning from the past, others, and other domains (CLTC, 2021), this research proposes a risk assessment method for LLMs building upon existing AI, cyber security, and risk assessment efforts. This research continues by further developing and evaluating the method proposed in real scenarios as it aims to serve as a contribution to building safe, robust, responsible, and trustworthy AI systems that are beneficial to society across different industries and applications.

## References


Abbas, S. (2023). LLMC: LLM life cycle https://medium.com/@SyedAbbasT/llm-lc-large-language-model-life-cycle-99cbbb12771f

Begou, N., Vinoy, J., Duda, A., & Korczyński, M. (2023). Exploring the Dark Side of AI: Advanced Phishing Attack Design and Deployment Using ChatGPT. In *2023 IEEE Conference on Communications and Network Security (CNS)* (pp. 1-6). IEEE.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., ... & Xie, X. (2023). A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

CLTC (2021). Guidance for the development of AI risk and impact assessments.

Deldjoo, Y. (2023). Fairness of chatgpt and the role of explainable-guided prompts. *arXiv preprint arXiv:2307.11761*.

Derner, E., Batistič, K., Zahálka, J., & Babuška, R. (2023). A security risk taxonomy for large language models. *arXiv preprint arXiv:2311.11415*.

EU Commission (2019). Ethics guidelines for Trustworthy AI. High-level expert group in Artificial Intelligence.

EU Parliament (2023). Artificial Intelligence Act.

EY (2023). Political agreement reached on the EU Artificial Intelligence Act.

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Guo, S., Xie, C., Li, J., Lyu, L., & Zhang, T. (2022). Threats to pre-trained language models: Survey and taxonomy. *arXiv preprint arXiv:2202.06862*.

Khlaaf, H. (2023). Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems. *Trail of Bits*.

IEAI (2022). On a risk-based assessment approach to AI ethics governance.

ISO/IEC (2008). 27005 Information technology – security techniques – information security risk management.

ISO/IEC (2023). 23894 Information technology – Artificial Intelligence – guidance on risk management.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Cheng, R. G. H., Klochkov, Y., ... & Li, H. (2023). Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv preprint arXiv:2308.05374*.

Maathuis, C., Pieters, W., & van den Berg, J. (2018). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.

Maathuis, C. (2022). An Outlook of Digital Twins in Offensive Military Cyber Operations. In *European Conference on the Impact of Artificial Intelligence and Robotics* (Vol. 4, No. 1, pp. 45-53).

Maathuis, C., Kerkhof, I., Godschalk, R., & Passier, H. (2023). Design Lessons from Building Deep Learning Disinformation Generation and Detection Solutions. In *European Conference on Cyber Warfare and Security* (Vol. 22, No. 1, pp. 285-293).

Maathuis, C., & Chockalingam, S. (2023, May). Modelling Responsible Digital Security Behaviour for Countering Social Media Manipulation. In *ECSM 2023 10th European Conference on social media*. Academic Conferences and publishing limited.

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side'of AI. *European Journal of Information Systems*, *31*(3), 257-268.

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. *AI and Ethics*, 1-31.

NIST (2018). Risk Management Framework for information systems and organizations.

NIST (2023). Artificial Intelligence risk management framework.

Pamula, V. (2023). An introduction to LLMOps: operationalizing and managing LLMs using Azure ML. https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/an-introduction-to-llmops-operationalizing-and-managing-large/ba-p/3910996

Piorkowski, D., Hind, M., & Richards, J. (2022). Quantitative ai risk assessments: Opportunities and challenges. *arXiv preprint arXiv:2209.06317*.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, 1-6.

SANS (2021). An overview of threat and risk assessment.

Schuett, J. (2023). Risk management in the artificial intelligence act. *European Journal of Risk Regulation*, 1-19.

Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). ChatGPT: More than a weapon of mass deception, ethical challenges and responses from the human-Centered artificial intelligence (HCAI) perspective. *arXiv preprint arXiv:2304.11215*.

Somer, P., & Thalmann, S. (2023). Risk Management of AI in Industry: a Literature Review.

Staab, R., Vero, M., Balunović, M., & Vechev, M. (2023). Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT–Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, *74*, 102700.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models, 2023. *URL https://arxiv. org/abs/2307.09288*.

TechLaw (2021). Responsible AI impact assessment (RAIIA).

UN (2020). Enterprise risk management approaches and uses in UN system organizations.

UN (2023). UN Internal Report Governing AI for humanity.

UNSMS (2017). Security Policy Manual.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., ... & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(2), 7-30.

WEF (2020). AI procurement in a box: workbook.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models (2021). *arXiv preprint arXiv:2112.04359*.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., ... & Bari, M. S. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Wu, X., Duan, R., & Ni, J. (2023). Unveiling security, privacy, and ethical concerns of ChatGPT. arXiv. *URL: https://arxiv. org/abs/2308.10092*.

Xia, B., Lu, Q., Perera, H., Zhu, L., Xing, Z., Liu, Y., & Whittle, J. (2023). Towards Concrete and Connected AI Risk Assessment (C 2 AIRA): A Systematic Mapping Study. In *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)* (pp. 104-116). IEEE.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2023). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.

Zhou, J., Müller, H., Holzinger, A., & Chen, F. (2023). Ethical ChatGPT: Concerns, challenges, and commandments. *arXiv preprint arXiv:2305.10646*.