ABSTRACT

The thesis titled Analysis of Security and Privacy Risks in Large Language Models, presented by student Ana Șarapova as a Master's project, was developed at the Technical University of Moldova. It is written in English and contains 67 pages, 15 tables, 10 figures, and 48 references. The thesis consists of a an introduction, three chapters, a conclusion, a list of references, and a list of appendices.

The research begins with an introduction into the domain of artificial intelligence (AI) systems followed by domain analysis of large language models (LLMs). The first chapter examines both the positive and negative impacts of this technology. Through the examination of over 74 scientific articles, and books, the literature review emphasized the lack of structure related to problems that appear during the utilization of LLMs. This section also explores existing data protection techniques, identifies key challenges, and proposes solutions to ensure the safe development and use of LLM systems.

The second chapter presents the initial deliverables of the research. It introduces the theoretical research framework, followed by detailed explanation of the 3-dimensional taxonomy, extended glossary, and decision tree. The chapter concludes with a discussion of the taxonomy's validation, conducted by three domain experts. Their analysis, based on specific criteria, offers constructive feedback and recommendations for future work.

The final chapter focuses on the design and implementation of an expert system which has the purpose of demonstrating the practical application of the taxonomy. This system's design is outlined through the definition of functional and non-functional requirements, along with the development of its architecture and components. Additionally, the chapter describes the implementation of the expert system, achieving the second objective, and discusses project results and observations.

REZUMAT

Teza cu titlul Analiza riscurilor de securitate și confidențialitate în modelele lingvistice mari, prezentată de studenta Ana Șarapova ca proiect de masterat, a fost elaborată în cadrul Universității Tehnice a Moldovei. Este redactată în limba engleză și conține 67 de pagini, 15 tabele, 10 figuri și 48 de referințe. Teza este structurată în: o introducere, trei capitole, o concluzie, lista referințelor și o listă de anexe.

Cercetarea începe cu o introducere în domeniul sistemelor de inteligență artificială (AI), urmată de o analiză a domeniului modelelor lingvistice mari (LLM). Primul capitol examinează atât impacturile pozitive, cât și cele negative ale acestei tehnologii. Prin examinarea a peste 74 de articole științifice și cărți, analiza literaturii de specialitate a evidențiat lipsa unei structuri organizate privind problemele care apar în timpul utilizării LLM-urilor. Această secțiune explorează, de asemenea, tehnicile existente de protecție a datelor, identifică principalele provocări și propune soluții pentru a asigura dezvoltarea și utilizarea în siguranță a sistemelor LLM.

Capitolul al doilea prezintă primele livrabile ale cercetării. Acesta introduce cadrul teoretic al cercetării, urmat de o explicație detaliată a taxonomiei tridimensionale, a glosarului extins și a arborelui decizional. Capitolul se încheie cu o discuție privind validarea taxonomiei, realizată de trei experți din domeniu. Analiza lor, bazată pe criterii specifice, oferă feedback constructiv și recomandări pentru lucrări viitoare.

Capitolul final se concentrează pe proiectarea și implementarea unui sistem expert care are scopul de a demonstra aplicarea practică a taxonomiei. Proiectarea acestui sistem este detaliată prin definirea cerințelor funcționale și nefuncționale, împreună cu dezvoltarea arhitecturii și a componentelor sale. În plus, capitolul descrie procesul de implementare a sistemului expert, realizând al doilea obiectiv al cercetării, și analizează rezultatele proiectului și observațiile obținute.

TABLE OF CONTENTS

INTRODUCTION

Artificial intelligence (AI) technology brings a new dimension to human-computer interaction. It is a transformative field designed to replicate human-like thought processes capabilities in machines. Today, AI's integration into various sectors, including healthcare, finance, and education, highlights its potential to solve complex problems efficiently and cost-effectively, reshaping industries and questioning the notion of intelligence as being an exclusively human trait [1]. However, its rapid evolution presents to the society a range of opportunities and challenges, that are more relevant than ever.

The positive impact can be seen in healthcare field, where AI-powered systems assist in early disease detection, personalized treatment plans, and robotic surgeries. In business, it is used during decision-making and strategic planning performing predictive analysis, optimization of supply chain processes, enhancing security measures through anomalies detection. It also helps people with disabilities through assistive technologies like voice recognition, gesture control, emotional support encouraging their inclusion in day-to-day life [1].

While AI brings numerous benefits to society, the other side of the coin reveals significant risks and challenges. For example, autonomous vehicles may prioritize optimization goals like travel time over safety, decision that can lead to lethal consequences [2]. As happened in March 2018, in Arizona, where a self-driving Uber fatally collided a woman, marking the first pedestrian death caused by an autonomous car [3]. Also, this can be an instrument for generating fake news, biassed recommendations, propaganda or discrimination. For example, in 2015, the Amazon’s AI recruiting tool, developed to assess job applications, showed bias against women, due to training of the algorithm mostly on male resumes [4]. This underscores the bias in predictions and insensitivity of AI systems.

Beside all the risks arising, AI tools are largely used by a lot of people. The most popular products, ranked by total web visits (from Sept 2022 to Aug 2023) are: ChatGPT, Character.AI, QuillBot, Midjourney, Hugging Face, Google Bard, and others [5]. Where, ChatGPT resulted in 60% visits and Character.AI in 15.8%, dominating the web industry.

Recent years have introduced a new addition to this well-established domain: a class of foundation models trained on enormous amounts of data named large language models (LLMs) [6]. This technology “reshapes industries” through processes automation, engaging customers through intelligent chatbots or troubleshooting issues. However, through the mass utilization of this type of tool, a lot of risks show up.

These days, due to the free access, even children can use LLMs, depending on the positive or negative intentions. With the apparition of these models, cybersecurity issues went to a new level. Now, LLMs can help dealing with these issues, and also help malicious actors in performing cybersecurity attacks.

This research aims to analyze the current situation regarding security and privacy risks in LLMs, identifying a suited framework for defining the safe LLM utilization and deployment.

BIBLIOGRAPHY

[1] S. Mahajan, “Artificial Intelligence and its Impacts on the Society,” *CSS*, vol. 32, no. 4, pp. 135–151, Dec. 2023, doi: 10.62047/CSS.2023.12.31.135.

[2] S. Saisubramanian, S. C. Roberts, and S. Zilberstein, “Understanding User Attitudes Towards Negative Side Effects of AI Systems,” in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–6. doi: 10.1145/3411763.3451654.

[3] “Driver in deadly Tempe Uber crash pleads guilty to endangerment | FOX 10 Phoenix.” Accessed: Jan. 17, 2025. [Online]. Available: https://www.fox10phoenix.com/news/driver-in-deadly-tempe-uber-crash-in-court-settlement-possible

[4] “Amazon scrapped ‘sexist AI’ tool.” Accessed: Jan. 17, 2025. [Online]. Available: https://www.bbc.com/news/technology-45809919

[5] “Ranked: The Most Popular AI Tools.” Accessed: Jan. 17, 2025. [Online]. Available: https://www.visualcapitalist.com/ranked-the-most-popular-ai-tools/

[6] “What Are Large Language Models (LLMs)? | IBM.” Accessed: Oct. 09, 2024. [Online]. Available: https://www.ibm.com/topics/large-language-models

[7] M. A. K. Raiaan *et al.*, “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges,” Sep. 27, 2023. doi: 10.36227/techrxiv.24171183.

[8] M. U. Hadi *et al.*, “Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects,” Aug. 12, 2024, *Preprints*. doi: 10.36227/techrxiv.23589741.v6.

[9] A. Qammar, H. Wang, J. Ding, A. Naouri, M. Daneshmand, and H. Ning, “Chatbots to ChatGPT in a Cybersecurity Space: Evolution, Vulnerabilities, Attacks, Challenges, and Future Recommendations,” May 29, 2023, *arXiv*: arXiv:2306.09255. Accessed: Oct. 09, 2024. [Online]. Available: http://arxiv.org/abs/2306.09255

[10] Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan, “A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions,” *IEEE Open J. Comput. Soc.*, vol. 4, pp. 280–302, 2023, doi: 10.1109/OJCS.2023.3300321.

[11] Professor, College of Management, Kyung Hee University, Korea, K. J. Lee, T. Hong, H. Ahn, T. Kim, and C. Koo, “Special Topic: The Impact of ChatGPT in Society, Business, and Academia,” *Asia Pacific Journal of Information Systems*, vol. 33, no. 4, pp. 957–976, Dec. 2023, doi: 10.14329/apjis.2023.33.4.957.

[12] D. G. Poalelungi *et al.*, “Advancing Patient Care: How Artificial Intelligence Is Transforming Healthcare,” *JPM*, vol. 13, no. 8, p. 1214, Jul. 2023, doi: 10.3390/jpm13081214.

[13] X. Wang, N. Anwer, Y. Dai, and A. Liu, “ChatGPT for design, manufacturing, and education,” *Procedia CIRP*, vol. 119, pp. 7–14, 2023, doi: 10.1016/j.procir.2023.04.001.

[14] F. He, T. Zhu, D. Ye, B. Liu, W. Zhou, and P. S. Yu, “The Emerged Security and Privacy of LLM Agent: A Survey with Case Studies,” Jul. 27, 2024, *arXiv*: arXiv:2407.19354. Accessed: Oct. 09, 2024. [Online]. Available: http://arxiv.org/abs/2407.19354

[15] F. Allwein, “ChatGPT - A critical view,” no. 1, 2024.

[16] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, “A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly,” *High-Confidence Computing*, vol. 4, no. 2, p. 100211, Jun. 2024, doi: 10.1016/j.hcc.2024.100211.

[17] X. Wu, R. Duan, and J. Ni, “Unveiling security, privacy, and ethical concerns of ChatGPT,” *Journal of Information and Intelligence*, vol. 2, no. 2, pp. 102–115, Mar. 2024, doi: 10.1016/j.jiixd.2023.10.007.

[18] A. Habbal, M. K. Ali, and M. A. Abuzaraida, “Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions,” *Expert Systems with Applications*, vol. 240, p. 122442, Apr. 2024, doi: 10.1016/j.eswa.2023.122442.

[19] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, “Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs,” Sep. 03, 2023, *arXiv*: arXiv:2308.13387. Accessed: Oct. 09, 2024. [Online].

Available: http://arxiv.org/abs/2308.13387

[20] A. Kumar, S. V. Murthy, S. Singh, and S. Ragupathy, “The Ethics of Interaction: Mitigating Security Threats in LLMs,” Jul. 10, 2024, *arXiv*: arXiv:2401.12273. Accessed: Oct. 09, 2024. [Online]. Available: http://arxiv.org/abs/2401.12273

[21] R. Pasupuleti, R. Vadapalli, and C. Mader, “Cyber Security Issues and Challenges Related to Generative AI and ChatGPT,” in *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Abu Dhabi, United Arab Emirates: IEEE, Nov. 2023, pp. 1–5. doi: 10.1109/SNAMS60348.2023.10375472.

[22] V. Chang, L. Draper, and S. Yu, “Generative AI Risk Management in Digital Economy:,” in *Proceedings of the 6th International Conference on Finance, Economics, Management and IT Business*, Angers, France: SCITEPRESS - Science and Technology Publications, 2024, pp. 120–127. doi: 10.5220/0012729800003717.

[23] G. 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, pp. 1–14, Jul. 2023, doi: 10.4018/IJSPPC.325475.

[24] C. Piri, “DATA PRIVACY IN THE AGE OF LLM-BASED SERVICES IN EDUCATION: CURRENT CHALLENGES, IMPROVEMENT GUIDELINES AND FUTURE DIRECTIONS”.

[25] Y. Wang *et al.*, “HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection,” Aug. 06, 2024, *arXiv*: arXiv:2408.02927. Accessed: Oct. 11, 2024. [Online]. Available: http://arxiv.org/abs/2408.02927

[26] “What is AI TRiSM? What It Means and Why It Matters,” Abnormal. Accessed: Nov. 19, 2024. [Online]. Available: https://abnormalsecurity.com/glossary/ai-trism

[27] “Gartner Survey Revealed 34% of Organizations Are Already Using or Implementing AI Application Security Tools,” Gartner. Accessed: Nov. 20, 2024. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2023-09-18-gartner-survey-revealed-34-percent-of-organizations-are-already-using-or-implementing-ai-application-security-tools

[28] “Theoretical and conceptual frameworks in research – Dr Lynette Pretorius.” Accessed: Nov. 21, 2024. [Online]. Available: https://www.lynettepretorius.com/the_scholars_way_blog/theoretical-and-conceptual-frameworks-in-research/

[29] A. Nucci, “Large Language Model (LLM) Security: OWASP Checklist Guide,” Aisera: Best Generative AI Platform For Enterprise. Accessed: Nov. 21, 2024. [Online]. Available: https://aisera.com/blog/llm-security/

[30] S. Ray, “Samsung Bans ChatGPT Among Employees After Sensitive Code Leak,” Forbes. Accessed: Oct. 10, 2024. [Online]. Available: https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

[31] Click4Assistance, “What countries is ChatGPT banned in?” Accessed: Oct. 10, 2024. [Online]. Available: https://www.click4assistance.co.uk/what-countries-is-chatgpt-banned-in-and-how-this-will-impact-live-chat-for-businesses

[32] Y. Huang *et al.*, “TrustLLM: Trustworthiness in Large Language Models,” Sep. 30, 2024, *arXiv*: arXiv:2401.05561. Accessed: Nov. 21, 2024. [Online]. Available: http://arxiv.org/abs/2401.05561

[33] “taxonomy.” Accessed: Jan. 13, 2025. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/taxonomy

[34] T. Schoormann, F. Möller, and D. Szopinski, “Exploring Purposes of Using Taxonomies”.

[35] B. Oksengendler, K. Mukimov, R. Letfullin, N. Turaeva, G. Abdurakhmanov, and S. Yuldashev, “Periodic table of elements, Mendeleev’s periodic table: history, achievements and problems,” *Bulletin of National University of Uzbekistan: Mathematics and Natural Sciences*, vol. 2, no. 2, pp. 94–112, Jun. 2019, doi: 10.56017/2181-1318.1023.

[36] M. R. Kibler, “From the Mendeleev periodic table to particle physics and back to the periodic table,” *Found Chem*, vol. 9, no. 3, pp. 221–234, Oct. 2007, doi: 10.1007/s10698-007-9039-9.

[37] M. Usman, R. Britto, J. Börstler, and E. Mendes, “Taxonomies in software engineering: A Systematic mapping study and a revised taxonomy development method,” *Information and Software Technology*, vol. 85, pp. 43–59, May 2017, doi: 10.1016/j.infsof.2017.01.006.

[38] M. Shakheen, “Glossary: A Valuable Resource for Knowledge Sharing,” Document360. Accessed: Jan. 13, 2025. [Online]. Available: https://document360.com/blog/glossary-in-knowledge-base/

[39] “What is a Glossary: Definition and Purpose.” Accessed: Jan. 13, 2025. [Online]. Available: https://www.timelytext.com/what-is-a-glossary-and-why-is-it-important/

[40] M. K. Tackabery, "Defining Glossaries," *Technical Communication*, vol. 52, no. 4, pp. 427-433, Nov. 2005. [Online]. Available: https://www.researchgate.net/publication/233620484

[41] K. Mittal, D. Khanduja, and P. C. Tewari, “An Insight into ‘Decision Tree Analysis’”.

[42] “What Is a Decision Tree and How Is It Used?” Accessed: Jan. 14, 2025. [Online]. Available: https://careerfoundry.com/en/blog/data-analytics/what-is-a-decision-tree/

[43] M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel, “Visual classification: an interactive approach to decision tree construction,” in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego California USA: ACM, Aug. 1999, pp. 392–396. doi: 10.1145/312129.312298.

[44]BOLUN, Ion, BULAI, Rodica, CIORBĂ, Dumitru. Support of education in cybersecurity. In: Pro Publico Bono – Public Administration. 2021, nr. 1(9), pp. 127-147. ISSN 2063-9058. DOI: https://doi.org/10.32575/ppb.2021.1.8

[45]BULAI, Rodica, ŢURCANU, Dinu, CIORBĂ, Dumitru. Education in Cybersecurity. In: Central and Eastern European eDem and eGov Days. 2-3 mai 2019, Budapesta. Viena, Austria: Facultas Verlags- und Buchhandels, 2019, pp. 33-44. ISBN 978-3-7089-1898-3; 978-3-903035-24-9. DOI: https://doi.org/10.24989/ocg.v335.2

[46] “Building Expert Systems.” Accessed: Nov. 21, 2024. [Online]. Available: https://engineering.purdue.edu/~engelb/abe565/es.htm

[47] P. J. Lucas and L. C. van der Gaag, *Principles of Expert Systems*, Amsterdam: Addison-Wesley, 1991. [Online]. Available: https://www.researchgate.net/publication/259867658_Principles_of_Expert_Systems