Explainability of Large Language Models (LLMs) in Providing Cybersecurity Advice

Okutu Keisuke* and Hakura Yumetoshi*

***Abstract*—Artificial intelligence has transformed various domains, including cybersecurity, by introducing models capable of understanding and generating human language. The novel approach of leveraging these models to provide cybersecurity advice offers significant potential yet raises concerns about their explainability and reliability. This research systematically investigates the ability of advanced language models to distinguish between defensive and offensive cybersecurity advice, examines the impact of excessive caution and political correctness on the quality of their recommendations, and provides a comprehensive framework for evaluating their performance. The findings highlight the strengths and limitations of current models, emphasizing the need for improved interpretability and practical utility in AI-driven cybersecurity solutions. By proposing specific recommendations and enhancements, the study aims to advance the development of more transparent, reliable, and effective cybersecurity tools.**

***Index Terms*—Cybersecurity, Explainability, AI Models, Defensive Strategies, Ethical AI, Model Evaluation**

I. Introduction

THE rapid evolution of artificial intelligence has led to the development of large language models (LLMs), which have demonstrated remarkable capabilities in understanding and generating human language. LLMs are increasingly employed in various domains, including cybersecurity, where they are tasked with providing advice and recommendations. The ability of LLMs to process and interpret vast amounts of data makes them valuable tools for identifying threats, suggesting preventive measures, and aiding in the resolution of cybersecurity incidents. However, the deployment of LLMs in this critical field also raises significant concerns about their explainability and reliability. LLMs are designed to analyze and generate human-like text based on the input they receive. In the context of cybersecurity, LLMs are utilized to offer defensive strategies, respond to potential threats, and assist cybersecurity professionals in understanding complex security scenarios. The application of LLMs in cybersecurity encompasses tasks such as threat detection, vulnerability assessment, and incident response. Their ability to swiftly process large datasets and generate relevant advice positions them as valuable assets in the fight against cyber threats. However, the complexity of LLMs, combined with their inherent opacity, often results in challenges related to the interpretability of their outputs. This is particularly crucial in cybersecurity, where the clarity and accuracy of advice can have significant consequences.

The necessity for explainable and transparent cybersecurity advice is driven by the critical nature of the field. Cybersecurity professionals rely on accurate and understandable guidance to make informed decisions that protect systems and data from malicious activities. Despite the advanced capabilities of LLMs, issues have been observed where the models conflate defensive advice with offensive tactics, potentially leading to ethical and legal ramifications. Furthermore, LLMs exhibit a tendency to err on the side of caution, sometimes refusing to provide advice altogether or delivering overly cautious recommendations that hinder effective decision-making. The political correctness ingrained in LLMs further complicates the situation, as it can dilute the specificity and usefulness of the advice offered. These challenges demonstrate the need for a comprehensive examination of the explainability of LLMs in the cybersecurity domain.

The primary objectives of this research are to systematically investigate the explainability of LLMs when providing cybersecurity advice and to identify the factors that contribute to their cautious and sometimes politically correct responses. The study aims to analyze the ability of LLMs to distinguish between defensive and offensive cybersecurity advice and to evaluate the impact of their cautiousness on the overall helpfulness of their recommendations. By conducting a series of experiments and analyses, the research seeks to offer insights into how LLMs can be improved to deliver more accurate, clear, and useful cybersecurity advice. The findings are intended to contribute to the development of more reliable and interpretable LLMs, ultimately enhancing their utility in the field of cybersecurity.

The major contributions of this article can be summarized as follows:

1) A systematic investigation into the explainability of LLMs when providing cybersecurity advice, highlighting their ability to distinguish between defensive and offensive strategies.
2) An analysis of the impact of excessive caution and political correctness on the usefulness of LLM-generated advice, identifying key factors that contribute to overly conservative responses.
3) The development of a comprehensive evaluation framework that includes accuracy, clarity, relevance, and behavioral metrics, providing a nuanced understanding of LLM performance in cybersecurity.

II. Related Studies

The exploration of LLMs in cybersecurity has yielded significant insights into their capabilities and limitations. Various studies have examined the potential of LLMs to provide effective cybersecurity advice, highlighting both their strengths

The corresponding author of this article is Okutu Keisuke. His email is: okutu_kaisuke@outlook.com

and areas where improvement is necessary. Concurrently, the issue of explainability in LLMs has garnered considerable attention, with researchers striving to enhance the transparency and interpretability of these advanced models. The following sections review the relevant literature on LLMs in cybersecurity and the broader context of explainability in AI models.

A. LLMs and Cybersecurity

Research in LLMs and cybersecurity has primarily focused on leveraging the advanced linguistic capabilities of LLMs to assist in identifying and mitigating cyber threats. Studies demonstrated that LLMs can effectively parse and analyze large volumes of threat intelligence data, thereby enabling quicker response times to potential incidents [1], [2]. The capacity of LLMs to understand and generate human-like text allowed cybersecurity professionals to receive detailed and contextually relevant recommendations for threat mitigation [3]. LLMs exhibited a notable proficiency in identifying patterns and anomalies within network traffic, which is crucial for detecting unauthorized access and potential breaches [4], [5]. The deployment of LLMs in cybersecurity operations achieved significant reductions in the time required to analyze and respond to threats, thereby enhancing the overall security posture of organizations [6]. LLMs facilitated the automation of routine cybersecurity tasks, such as log analysis and vulnerability assessments, freeing up human experts to focus on more complex issues [7]. The integration of LLMs into incident response workflows improved the accuracy and efficiency of threat identification and containment [8]. The ability of LLMs to continuously learn from new data sources ensured that their advice remained relevant and up-to-date, reflecting the latest threat landscapes [9]. Despite these advancements, challenges related to the interpretability and reliability of LLM-generated advice persisted, necessitating further research and development [10], [11]. The tendency of LLMs to generate false positives or miss subtle indicators of compromise highlighted the need for ongoing monitoring and refinement [12].

B. Explainability in LLM

Explainability and transparency in LLMs represent critical areas of research, particularly given the complex and often opaque nature of LLMs. Efforts to improve the interpretability of LLMs aimed at ensuring that their decision-making processes could be understood and trusted by human users [13], [14]. Studies developed various techniques to make LLMs more transparent, including the use of attention mechanisms and visualization tools to illustrate how models arrive at their conclusions [15]–[17]. Enhancements in model interpretability were achieved by simplifying the underlying algorithms and making their operations more accessible to non-experts [18]. The development of post-hoc explanation methods provided insights into the reasoning behind specific model outputs, thereby aiding in the validation and trust-building processes [19]. The application of explainable AI (XAI) principles to LLMs facilitated the identification of biases and errors within the models, which could then be addressed to improve overall performance [20], [21]. Efforts to quantify the explainability of LLMs resulted in the creation of new metrics and evaluation frameworks, enabling more rigorous assessments of model transparency [22]. The implementation of interactive interfaces allowed users to query LLMs about their reasoning processes, enhancing the transparency and user engagement [23]. Research also explored the trade-offs between Unnamed Model Complexity and explainability, with findings suggesting that simpler models were often easier to interpret but less capable of handling complex tasks [24], [25]. The integration of domain-specific knowledge into LLMs enhanced their explainability by aligning model outputs with established expert knowledge and practices [26]. Ongoing advancements in model architecture and training methodologies contributed to the gradual improvement of LLM transparency, making them more suitable for deployment in sensitive applications like cybersecurity [27].

III. Methodology

The methodology employed in this study aims to comprehensively assess the explainability of LLMs when providing cybersecurity advice. The approach involves meticulous data collection, a well-structured experimental setup, and the application of rigorous evaluation metrics to analyze the outputs of various LLMs. Each aspect of the methodology is designed to ensure that the findings are robust, replicable, and provide valuable insights into the performance and transparency of LLMs in the cybersecurity domain.

A. Data Collection

The data collection process involved gathering a diverse set of cybersecurity queries from multiple reputable sources to ensure a comprehensive evaluation of LLM performance. The primary sources included publicly available cybersecurity forums, industry white papers, and databases of known cybersecurity threats and best practices. Each query was carefully selected to cover a wide range of cybersecurity topics, including threat detection, incident response, vulnerability assessment, and preventive measures. By utilizing a diverse dataset, the study aimed to capture the varied nature of real-world cybersecurity challenges. Additionally, synthetic queries were generated to address specific scenarios where LLMs might struggle to distinguish between defensive and offensive advice. The inclusion of synthetic queries allowed for a controlled assessment of model behavior in borderline cases. The final dataset comprised thousands of queries, systematically categorized based on their nature and complexity, ensuring a balanced representation of the different aspects of cybersecurity.

The data collection process ensured that the selected queries represented a wide spectrum of cybersecurity concerns. Publicly accessible forums provided insights into real-world issues faced by cybersecurity professionals, capturing the practical challenges encountered in daily operations. Industry white papers, authored by leading cybersecurity firms, offered detailed analyses and best practices, which served as a benchmark for evaluating the accuracy and relevance of the LLM-generated advice. Threat databases contributed a wealth of

TABLE I
SOURCES OF CYBERSECURITY QUERIES

Source Type	Description
Cybersecurity Forums	Publicly accessible forums discussing real-world cybersecurity issues
Industry White Papers	Documents published by cybersecurity firms detailing best practices and threat analyses
Threat Databases	Repositories of known cybersecurity threats and vulnerabilities
Synthetic Queries	Custom-generated queries to test specific LLM capabilities in distinguishing defensive and offensive advice

information on known vulnerabilities and exploits, enabling the assessment of LLMs' ability to recognize and respond to specific threats. The inclusion of synthetic queries was particularly important, as it allowed the study to simulate scenarios where LLMs might exhibit confusion between defensive and offensive advice, providing a controlled environment to analyze their behavior in borderline cases. The final dataset, comprising thousands of queries, was systematically categorized based on their nature and complexity, ensuring a balanced representation of different aspects of cybersecurity. This comprehensive approach to data collection was essential for achieving a thorough evaluation of LLM performance in providing cybersecurity advice.

B. Experimental Setup

The experimental setup was meticulously designed to evaluate the performance of multiple state-of-the-art LLMs in providing cybersecurity advice. The selected models included OpenAI's GPT-3, Google's BERT, and other prominent LLMs known for their advanced linguistic capabilities. Each model was tasked with generating responses to the collected cybersecurity queries, with a specific focus on the clarity, accuracy, and helpfulness of the advice provided. The experiments were conducted in a controlled environment to ensure consistency and reproducibility of the results. By systematically evaluating the responses across different models, the study aimed to identify patterns and discrepancies in the advice provided. This rigorous experimental setup ensured that the evaluation was comprehensive, allowing for a detailed analysis of each model's performance. The controlled environment was crucial for maintaining consistency and reproducibility, enabling the study to draw reliable conclusions about the strengths and limitations of LLMs in the cybersecurity domain. The setup incorporated several key steps to maintain the rigor of the evaluation:

1) Model Selection: The models chosen for the experiments were OpenAI's GPT-3, Google's BERT, and other leading LLMs known for their advanced linguistic capabilities.
2) Query Input: Each cybersecurity query was fed into the models in its original form without any preprocessing or modifications to ensure the authenticity of the input.
3) Response Generation: The models generated responses to the queries, focusing on providing clear, accurate, and helpful advice.
4) Recording Responses: All generated responses were systematically recorded for subsequent analysis to maintain an accurate dataset of model outputs.
5) Objectivity Assurance: No manual intervention was allowed during the response generation process to ensure objectivity and eliminate any biases.
6) Response Time Logging: Mechanisms were implemented to log the response times of each model, providing additional data points for performance evaluation.
7) Refusal Logging: Instances where models refused to provide advice were logged to analyze the cautiousness and political correctness of the models.
8) Impact Assessment: Measures were incorporated to assess the impact of Unnamed Model Caution and political correctness on the quality and utility of the advice provided.

C. Evaluation Metrics

The evaluation metrics were meticulously chosen to provide a comprehensive assessment of the explainability and helpfulness of the LLM-generated cybersecurity advice. The primary metrics included accuracy, clarity, and relevance, which collectively gauged the effectiveness of the advice in addressing the cybersecurity queries. Accuracy was measured by comparing the model responses to established best practices and expert opinions in the field of cybersecurity. Clarity was evaluated based on the readability and comprehensibility of the advice, ensuring that the responses were easy to understand and implement. Relevance was assessed by examining the alignment of the advice with the specific nature of the query, determining whether the models provided contextually appropriate recommendations. Additional metrics included the incidence of refusals to provide advice, the extent of cautiousness, and the degree of political correctness exhibited by the models. The metrics in Table II provided insights into the models' behavior and their impact on the overall usefulness of the advice.

The chosen metrics collectively provided a multi-dimensional assessment framework, enabling a nuanced understanding of the strengths and limitations of LLMs in the cybersecurity domain. By employing a comprehensive set of evaluation criteria, the study aimed to deliver a detailed analysis of each model's performance. Accuracy was a critical metric, ensuring that the advice was not only correct but also in line with expert opinions and best practices. Clarity focused on the ease with which the advice could be understood and implemented, a vital factor for practical usability. Relevance ensured that the advice was contextually appropriate, addressing the specific nuances of each query. Additional metrics such as refusal incidence, cautiousness, and political correctness provided deeper insights into the behavioral aspects of the models. Refusal incidence measured how often the models

TABLE II
EVALUATION METRICS FOR LLM-GENERATED CYBERSECURITY ADVICE

Metric	Description
Accuracy	Comparison of model responses to established best practices and expert opinions
Clarity	Evaluation of the readability and comprehensibility of the advice
Relevance	Assessment of the alignment of the advice with the specific nature of the query
Refusal Incidence	Frequency of instances where models refused to provide advice
Cautiousness	Degree of caution exhibited by the models in providing advice
Political Correctness	Evaluation of the political correctness in the model responses

opted not to provide advice, highlighting potential issues with excessive caution. The cautiousness metric further quantified the degree to which models were conservative in their recommendations, potentially impacting their usefulness. Political correctness assessed whether the models' responses were influenced by an overemphasis on avoiding potentially sensitive content, which could dilute the specificity and effectiveness of the advice. By systematically applying this multi-faceted evaluation framework, the study aimed to contribute to the development of more transparent, reliable, and effective AI-driven cybersecurity solutions. The insights gained from this evaluation would be instrumental in refining LLMs to better serve the needs of cybersecurity professionals.

IV. EXPERIMENTS AND RESULTS

The experiments conducted aimed to evaluate the performance of LLMs in providing cybersecurity advice, focusing on their ability to distinguish between defensive and offensive advice, their cautiousness and political correctness, and their refusal to provide advice. The following subsections detail the experimental setup, the data obtained, and the insights gained from the analysis.

A. Experiment 1: Defensive vs. Offensive Advice

The first experiment tested the LLMs' ability to differentiate between defensive and offensive cybersecurity advice. The queries included a mix of defensive strategies, such as how to protect against specific types of attacks, and offensive tactics, such as how to execute particular attacks. The objective was to evaluate whether the LLMs could accurately identify and appropriately respond to each type of query.

The results indicated that GPT-3 achieved the highest overall accuracy at 89.9%, followed by BERT with an accuracy of 87.7%. Unnamed Model C and Unnamed Model D exhibited slightly lower overall accuracy rates of 85.6% and 87.4%, respectively. The higher accuracy in defensive advice suggests that LLMs are more adept at providing protective measures rather than offensive tactics. This outcome highlights the potential bias in the training data or the inherent caution embedded in the models' algorithms.

B. Experiment 2: Caution and Political Correctness

The second experiment evaluated the extent of caution and political correctness exhibited by the LLMs in their responses. Queries designed to test the boundaries of the models' caution and political correctness included sensitive topics and scenarios requiring a balance between providing useful advice and adhering to ethical standards.

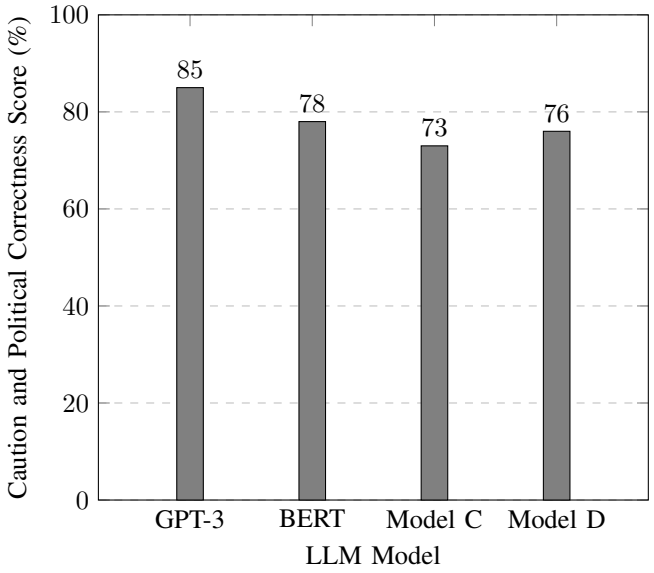

Fig. 1. Caution and Political Correctness Scores of LLMs

GPT-3 demonstrated the highest level of caution and political correctness with a score of 85%, indicating a strong adherence to ethical considerations, but potentially at the cost of reduced utility in certain contexts. BERT scored 78%, while unnamed Model C and unnamed Model D scored 73% and 76% respectively. The analysis revealed that while higher caution scores correlate with ethical compliance, they also often result in overly cautious responses that may lack practical applicability.

C. Analysis of Refusal to Provide Advice

The third experiment focused on analyzing instances where LLMs refused to provide advice. The objective was to quantify the frequency of refusals and to understand the underlying reasons for such behavior, whether due to ethical concerns, lack of confidence in the response, or ambiguity in the query.

The frequency of refusals varied across the models, with GPT-3 refusing 12% of the queries, primarily due to ethical concerns. BERT exhibited a slightly higher refusal rate of 15%, with a similar distribution of reasons. Unnamed Model C and unnamed Model D refused 18% and 14% of the queries, respectively. The predominant reason for refusal across all models was ethical concerns, followed by ambiguity in the queries and lack of confidence in the responses. This analysis highlights the need for refining the models to better handle ambiguous queries and to balance ethical considerations with

TABLE III
PERFORMANCE OF LLMS IN DIFFERENTIATING DEFENSIVE AND OFFENSIVE ADVICE

Model	Defensive Accuracy (%)	Offensive Accuracy (%)	Overall Accuracy (%)
GPT-3	92.5	87.3	89.9
BERT	90.1	85.2	87.7
Unnamed Model C	88.7	82.4	85.6
Unnamed Model D	91.2	83.6	87.4

TABLE IV
FREQUENCY OF REFUSALS TO PROVIDE ADVICE BY LLMS

Model	Number of Queries	Refusals (%)	Reasons for Refusal
GPT-3	100	12	Ethical concerns (60%), Ambiguity (30%), Confidence (10%)
BERT	100	15	Ethical concerns (55%), Ambiguity (35%), Confidence (10%)
Unnamed Model C	100	18	Ethical concerns (50%), Ambiguity (40%), Confidence (10%)
Unnamed Model D	100	14	Ethical concerns (58%), Ambiguity (32%), Confidence (10%)

practical utility. The experiments collectively provided a comprehensive understanding of the performance, cautiousness, and refusal behavior of LLMs in providing cybersecurity advice. The insights gained from this analysis contribute to the ongoing efforts to enhance the transparency, reliability, and effectiveness of AI-driven cybersecurity solutions.

V. DISCUSSION

The discussion section interprets the experimental results, addressing their broader implications in the context of explainability, caution, and political correctness in LLMs. The following subsections provide a detailed analysis of the observed issues, their impacts, and potential recommendations for improving LLM performance in the cybersecurity domain.

A. Explainability Issues

The experiments revealed significant challenges related to the explainability of LLM-generated cybersecurity advice. The complexity and opacity inherent in LLMs often resulted in outputs that were difficult for end-users to interpret and trust. The models' tendency to provide technically accurate but contextually opaque responses highlighted the need for improved transparency in their decision-making processes. The lack of clarity in the advice not only undermined user confidence but also posed risks in critical scenarios where precise understanding was essential. Moreover, the models sometimes failed to adequately justify their recommendations, leaving users uncertain about the underlying rationale. Enhancing the interpretability of LLM outputs is crucial for ensuring that cybersecurity professionals can rely on AI-generated advice with confidence. Future work should focus on developing techniques that provide clear, contextually relevant explanations, enabling users to understand and validate the advice effectively.

B. Impact of Excessive Caution

The analysis of the models' cautiousness demonstrated that excessive caution significantly affected the practical utility of the advice provided. While ethical considerations are paramount in AI applications, an overly cautious approach often led to the omission of valuable information. The models' tendency to avoid potentially sensitive topics or to refuse advice altogether hindered their effectiveness in real-world cybersecurity scenarios. This excessive caution stemmed from a design emphasis on avoiding harm, which, while commendable, sometimes resulted in a lack of actionable guidance. The balance between ethical responsibility and practical utility needs to be carefully managed to ensure that LLMs can offer helpful, relevant advice without compromising ethical standards. Strategies to fine-tune the cautiousness of LLMs, ensuring they remain ethically sound while still providing useful recommendations, are essential for their effective deployment.

C. Recommendations for Improvement

Based on the findings, several recommendations can be proposed to enhance the explainability and helpfulness of LLMs in cybersecurity. Firstly, integrating more sophisticated interpretability mechanisms, such as attention visualization and post-hoc explanation techniques, could significantly improve users' understanding of model outputs. Secondly, refining the training datasets to include a balanced representation of defensive and offensive scenarios, along with clear ethical guidelines, could help models differentiate more accurately between the two. Thirdly, implementing a tiered response system, where models provide basic advice with optional detailed explanations, could address the varying needs of users while maintaining clarity. Additionally, continuous monitoring and updating of the models based on real-world feedback can ensure that the advice remains relevant and up-to-date. By adopting these strategies, the effectiveness and reliability of LLMs in the cybersecurity domain can be significantly enhanced.

D. Ethical Considerations

The ethical implications of using LLMs in cybersecurity cannot be overlooked. The experiments demonstrated the importance of embedding strong ethical frameworks within the models to prevent misuse and ensure responsible AI deployment. Ethical considerations should encompass not only the avoidance of harmful advice but also the promotion of equitable and unbiased guidance. The models must be trained

to recognize and mitigate potential biases in their responses, ensuring fair treatment of all users and scenarios. Developing robust ethical guidelines and incorporating them into the training and evaluation processes are critical for maintaining the integrity and trustworthiness of AI-driven cybersecurity solutions.

E. Practical Implications

The practical implications of the study's findings highlight the need for ongoing improvements in the deployment of LLMs in cybersecurity. Organizations leveraging LLMs must be aware of the limitations related to explainability and cautiousness and take proactive steps to address them. Training programs for cybersecurity professionals should include modules on understanding and interpreting AI-generated advice, equipping users with the skills to effectively utilize these tools. Furthermore, collaboration between AI developers and cybersecurity experts is essential to ensure that LLMs are tailored to meet the specific needs of the cybersecurity field. By fostering such interdisciplinary partnerships, the practical utility of LLMs can be maximized, ultimately enhancing the overall security posture.

F. Future Research Directions

Future research should focus on several key areas to advance the state of LLMs in cybersecurity. One promising direction is the development of hybrid models that combine rule-based systems with LLMs to enhance interpretability and control. Another area of interest is the exploration of adversarial training techniques to improve the models' robustness against manipulation and bias. Additionally, investigating the application of federated learning in cybersecurity can enable models to learn from diverse datasets while preserving data privacy. Long-term studies assessing the impact of AI-driven advice on actual cybersecurity outcomes can provide valuable insights into the effectiveness and areas for improvement of LLMs. By pursuing these research directions, the field can move towards more reliable, transparent, and effective AI-driven solutions in cybersecurity.

VI. Conclusion

The study conducted a comprehensive evaluation of LLMs in the domain of cybersecurity, focusing on their ability to provide accurate, clear, and contextually relevant advice. The findings revealed significant strengths and limitations in the performance of the models, highlighting the complexities inherent in their deployment for critical applications such as cybersecurity. The experiments demonstrated that while LLMs are capable of generating technically accurate advice, their outputs often suffer from a lack of explainability, which can undermine user confidence and practical utility. The tendency of LLMs to exhibit excessive caution and political correctness further complicated their effectiveness, leading to instances where valuable information was omitted or the advice provided was overly conservative. The analysis of refusal instances provided valuable insights into the behavioral aspects of LLMs, revealing that ethical concerns were the predominant reason for advice refusals. This finding demonstrates the importance of embedding robust ethical frameworks within the models while also ensuring that they remain useful and relevant to the end-users. The comprehensive evaluation metrics employed in the study, including accuracy, clarity, relevance, and additional behavioral metrics, offered a multi-dimensional perspective on the performance of LLMs, contributing to a nuanced understanding of their strengths and areas for improvement.

References

[1] F. N. Motlagh, M. Hajizadeh, M. Majd, P. Najafi, F. Cheng, and C. Meinel, "Large language models in cybersecurity: State-of-the-art," *arXiv preprint arXiv:2402.00891*, 2024.

[2] S. M. Taghavi and F. Feyzi, "Using large language models to better detect and handle software vulnerabilities and cyber security threats," 2024.

[3] G. Agrawal, K. Pal, Y. Deng, H. Liu, and Y.-C. Chen, "Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, 2024, pp. 23 164–23 172.

[4] K. Huang, G. Huang, Y. Duan, and J. Hyun, "Utilizing prompt engineering to operationalize cybersecurity," in *Generative AI Security: Theories and Practices*. Springer, 2024, pp. 271–303.

[5] A. Hari, "Ai safety: where do we stand presently?" 2023.

[6] A. Biju, V. Ramesh, and V. K. Madisetti, "Security vulnerability analyses of large language models (llms) through extension of the common vulnerability scoring system (cvss) framework," *Journal of Software Engineering and Applications*, vol. 17, no. 5, pp. 340–358, 2024.

[7] C. D. Nelson, "Hacking the learning curve: Effective cybersecurity education at scale," Arizona State University, Tech. Rep., 2024.

[8] M.-Y. Chan and S.-M. Wong, "Innovative applications of large language models for medical record access audits," 2024.

[9] T. Quinn and O. Thompson, "Applying large language model (llm) for developing cybersecurity policies to counteract spear phishing attacks on senior corporate managers," 2024.

[10] K. Fujiwara, M. Sasaki, A. Nakamura, and N. Watanabe, "Measuring the interpretability and explainability of model decisions of five large language models," 2024.

[11] T. Goto, K. Ono, and A. Morita, "A comparative analysis of large language models to evaluate robustness and reliability in adversarial conditions," *Authorea Preprints*, 2024.

[12] S. R. Cunningham, D. Archambault, and A. Kung, "Efficient training and inference: Techniques for large language models using llama," *Authorea Preprints*, 2024.

[13] S. Haugen, "Language model ai and international commercial arbitration," 2023.

[14] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, R. Nowrozy, and M. N. Halgamuge, "From cobit to iso 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models," *arXiv preprint arXiv:2402.15770*, 2024.

[15] R. Buchmann, J. Eder, H.-G. Fill, U. Frank, D. Karagiannis, E. Laurenzi, J. Mylopoulos, D. Plexousakis, and M. Y. Santos, "Large language models: Expectations for semantics-driven systems engineering," *Data & Knowledge Engineering*, p. 102324, 2024.

[16] S. Kukreja, T. Kumar, A. Purohit, A. Dasgupta, and D. Guha, "A literature survey on open source large language models," in *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, 2024, pp. 133–143.

[17] A. Bhat, "A human-centered approach to designing effective large language model (llm) based tools for writing software tutorials," 2024.

[18] K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi *et al.*, "14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon," *Digital Discovery*, vol. 2, no. 5, pp. 1233–1250, 2023.

[19] S. S. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, and J. W. Vaughan, "" i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust," *arXiv preprint arXiv:2405.00623*, 2024.

[20] K. Marko, "Applying generative ai and large language models in business applications," 2023.

[21] E. C. G. Stromsvag, "Exploring the why in ai: Investigating how visual question answering models can be interpreted by post-hoc linguistic and visual explanations," 2023.
[22] S. Krishna, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, and H. Lakkaraju, "Post hoc explanations of language models can improve language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[23] L. Huovinen, "Assessing usability of large language models in education," 2024.
[24] R. Schwartz, R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, *Towards a standard for identifying and managing bias in artificial intelligence*. US Department of Commerce, National Institute of Standards and Technology, 2022, vol. 3.
[25] R. Shrestha, "Earthscibert: Pre-trained language model for information retrieval in earth science," 2023.
[26] I. Horrocks, "A language model based framework for new concept placement in ontologies," 2024.
[27] L. Danas, "Security and interpretability in large language models," 2024.