

Contextualized AI for Cyber Defense: An Automated Survey using LLMs

Christoforus Yoga Haryanto
School of Science
RMIT University
Melbourne, Australia
0009-0009-8340-5313

Anne Maria Elvira
School of Science
RMIT University
Melbourne, Australia

Trung Duc Nguyen
School of Science
RMIT University
Melbourne, Australia

Minh Hieu Vu
School of Science
RMIT University
Melbourne, Australia

Yoshiano Hartanto
Faculty of Engineering and IT
University of Tehnology
Sydney, Australia

Emily Lomempow
ZipThought
Melbourne, Australia
emily@zipthought.com.au

Arathi Arakala
School of Science
RMIT University
Melbourne, Australia
arathi.arakala@rmit.edu.au

Abstract—This paper surveys the potential of contextualized AI in enhancing cyber defense capabilities, revealing significant research growth from 2015 to 2024. We identify a focus on robustness, reliability, and integration methods, while noting gaps in organizational trust and governance frameworks. Our study employs two LLM-assisted literature survey methodologies: (A) ChatGPT 4 for exploration, and (B) Gemma 2:9b for filtering with Claude 3.5 Sonnet for full-text analysis. We discuss the effectiveness and challenges of using LLMs in academic research, providing insights for future researchers.

Index Terms—cyber security, artificial intelligence, retrieval augmented generation, cyber defense strategy, meta-analysis

I. INTRODUCTION

Contextualized AI enhances traditional Artificial Intelligence (AI) and Large Language Model (LLM) capabilities by integrating private data, beyond typical public datasets [1]. This emerging field finds application in autonomous monitoring, threat detection, and response within secured network environments [2], [3], [4], [5]. Yet, the full efficacy of these systems is under ongoing evaluation with challenges such as AI dependency, data privacy, human oversight, and end-to-end governance [6], [8], [9], [10], [11], [12], [13].

Initially, we hypothesized such systems were ready for widespread deployment in cyber security. However, we discovered a complex landscape with diverse terminology and approaches, leading us to do comprehensive literature review. Given the vast amount of potentially relevant research and the challenges in identifying pertinent studies, we decided to experiment with LLM-assisted methods for our survey while also providing us an opportunity to explore innovative methodologies for academic research [14], [15], [16].

This paper aims to conduct a survey of the literature in contextualized AI for cyber defense to answer:

- 1) RQ1: How can cyber security decision-makers strategically leverage contextualized AI to enhance defense capabilities while mitigating risks?

- 2) RQ2: Protection Layer: How can we ensure comprehensive protection of the system, data, and processes when implementing contextualized AI in cyber security?
- 3) RQ3: Security System Layer: How can we guarantee that the AI-enhanced protection system itself functions reliably and as expected?
- 4) RQ4: Organizational Layer: How can we foster organizational trust in AI, ensuring that the organization can confidently rely on AI capabilities within appropriate scopes while maintaining necessary human oversight?

A. Definition of Contextualized AI

For this study, we define contextualized AI to be supplied as part of the prompt to LLM as follows:

Contextualized AI refers to AI systems designed to access and utilize proprietary and domain-specific knowledge. While it is not strictly generative AI and LLM, most of the contextualized AI systems are built on top of generative AI and LLM so pay attention to the usage that involves further training using proprietary or domain-specific knowledge on top of pre-trained model. Some earlier papers may mention full AI training using private data, hence they should be considered as contextualized AI systems too.

B. Paper Structure

Section II describes our method, including the use of LLM tools, Section III discusses the findings from exploration using GPT-4, Section IV and V discusses the findings from literature screening using Gemma 2:9b and full-text analysis using Claude 3.5 Sonnet, and Section VI summarizes all the findings, including an analysis to the research methodology we use, and recommends for the future research directions.

This paper presents two distinct methodologies for LLM-assisted literature surveys: Method A using ChatGPT 4 for

initial exploration and thematic analysis, and Method B combining Gemma 2:9b for literature screening with Claude 3.5 Sonnet for full-text analysis. We compare these approaches to demonstrate their effectiveness in processing large volumes of academic literature efficiently.

II. METHODOLOGIES

We employ two distinct LLM-assisted approaches: 1) Method A: GPT-4 for initial exploration and thematic analysis. 2) Method B: Gemma 2:9b for literature screening and Claude 3.5 Sonnet for full-text analysis. See Fig. 1 for the overview.

A. Method A: Exploration with GPT

Method A uses GPT-4 via ChatGPT to rapidly generate a broad overview and identify key themes using the following steps:

- 1) Input research questions and definitions into GPT-4.
- 2) Instructed GPT-4 to search for relevant literature.
- 3) Used consistent prompt structure for each of RQ:


```
{{Research Question}}
```

```
{{Definitions}}
```

```
Search online for relevant conference papers and journal articles.
```
- 4) Prompted twice more with Find more for each query.
- 5) Collected and categorized provided sources.

Our exploratory review with GPT-4 returned 34 unique sources without dead links, including 18 academic publications (52.9%), 12 industry reports (35.3%), 2 professional

organization resources (5.9%), and 2 non-profit think tank publications (5.9%). All the returned sources were publicly accessible.

B. Method B: Systematic Review with Gemma and Claude

Method B employs a more systematic approach Gemma 2:9b and Claude 3.5 Sonnet for in-depth analysis [16], [21], [22]. We chose Gemma 2:9b for literature screening due to its efficiency in processing large volumes of text. Claude 3.5 Sonnet was selected for its ability to do full-text analysis of academic papers using prompt engineering.

Literature Screening with Gemma 2:9b: We used LLaSist, an LLM-based simplified screening tool with Gemma 2:9b backend (commit version 3bf51a6) [23]. LLaSist streamlines literature reviews through the following process:

- 1) **Data Input:** Processes CSV files with article metadata and abstracts, along with research questions.
- 2) **Key Semantics Extraction:** Extracts topics, entities, and keywords from titles and abstracts using Natural Language Processing (NLP).
- 3) **Relevance Estimation:** Assesses each article's relevance to research questions, providing scores (0-1) for relevance and contribution, with 0.7 as the threshold between false (below 0.7) and true (0.7 and above).
- 4) **Must-Read Determination:** Identifies "must-read" articles based on relevance and contribution scores.
- 5) **Output Generation:** Produces JSON and CSV files with detailed information for each article.

The process involved:

- 1) Querying the Scopus database with search string as "artificial AND intelligence AND cyber AND security" for 2015-2024. We use Scopus as its result already includes IEEE and ACM databases.
- 2) To check if the abstract addresses one or more research question we developed screening questions (SQ) for the LLM prompt as below, followed by the definitions:
 - SQ1: Does it discuss strategic factors for implementing LLM-based or contextualized AI in cyber security defense? [Definitions]
 - SQ2: Does it mention methods for integrating such AI into cyber security defense systems? [Definitions]
 - SQ3: Does it address techniques for ensuring robustness and reliability of these AI systems? [Definitions]
 - SQ4: Does it discuss organizational measures or governance frameworks for building trust in these solutions? [Definitions]
- 3) Analyze relevance and contribution of papers based on Gemma 2:9b's scores, using 0.7 as the threshold.

In-depth Analysis with Claude 3.5 Sonnet: We conducted a full-text review using Claude 3.5 Sonnet [22] with prompting techniques [24]. The following prompt engineering techniques are used: "Prompt generator for the initial draft", "Be clear and direct", "Give Claude a role", "Prefill Claude's

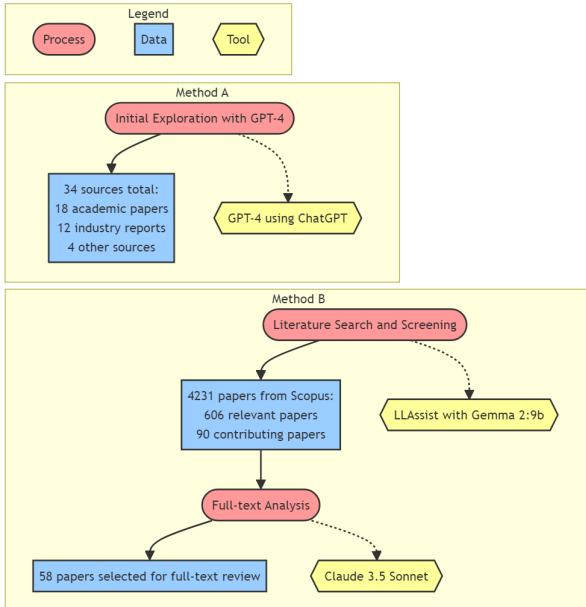


Fig. 1. Methodology and Data Gathering

response”, and “Long context prompting” [25]. We use the following steps:

- 1) Filtering contributing papers from 2020-2024 with DOIs, excluding non-Q1 journal articles.
- 2) Analyzing full-text PDFs using Claude 3.5 Sonnet.
- 3) Manual review for divergences between AI and human interpretations. We read the paper title, abstract, introduction, and conclusion then compare with the LLM output to check whether it is reasonably correct.

We used the following instruction prompt to guide Claude:

You are the research assistant. Given the research article, explain concisely how the paper addresses the following research questions:

{{Research Questions}}

{{Definition of Contextualized AI}}

In your explanation, first read the title and abstract, determine the type of the paper, e.g. survey paper, and then read the introduction and conclusion sections. Indicate if you need also to read the entire content of the paper. For each point you make, give the exact reference on where the original statement can be found, i.e. page/section number and paragraph. Strictly limit it to the actual content of the paper itself. Put a note in your explanation if the paper does not address one or more RQ.

Your output will be:

{{Output Format}}

This prompt engineering approach is effective in extracting relevant information and insights from full-text articles.

III. EXPLORATION WITH GPT-4

A. Key Findings

Our exploration with GPT-4 yielded valuable insights for each research question, drawing from both academic and non-academic sources:

1) *RQ1: Strategic Leverage of Contextualized AI:* Academic research highlights enhanced threat detection and adaptive defense [26], [27], automation of routine tasks [26], and proactive threat prediction [27]. Industry perspectives emphasize domain-specific models for improved threat recognition [28], human-AI teaming for optimal decision-making [29], and the importance of ethical considerations [26], [27], [55].

2) *RQ2: Comprehensive Protection Strategies:* Academic sources advocate for deep learning in behavioral analysis [33], [34] and integration of generative AI technologies [35]. Industry reports recommend persistent monitoring and real-time analysis [30], [57], contextualized security measures [31], AI-driven anomaly detection [30], and automation of security processes [31], [58].

3) *RQ3: Ensuring AI System Reliability:* Academic research focuses on robust system structures and reliability assessment [36], recurrent events analysis for prediction [40], and Scientific Machine Learning (SciML) for safeguarding [40]. Industry and government initiatives emphasize enhanced protection strategies like GREP [37], AI Systems Engineering and Reliability Technologies (ASERT) [41], risk management [38], and human-in-the-loop approaches with continuous monitoring [39].

4) *RQ4: Fostering Organizational Trust:* Academic sources stress transparency and explainability of AI systems [46], [50], ethical considerations in AI development [53], [54], and dynamic trust calibration mechanisms [44]. Industry perspectives highlight robust compliance and security implementation [42], cultivating a culture of ethical AI use [43], effective communication strategies [45], talent development [47], [51], and stakeholder engagement with policy advocacy [49], [55].

B. Common Themes

While academic and non-academic sources address similar themes, the angles are different. Academic sources tend to focus on theoretical frameworks, in-depth technical aspects, and long-term implications while non-academic sources emphasize practical applications with market-orientation. Across both academic and industry sources, common themes emerged:

- Importance of human-AI collaboration [29], [48], [52]
- Continuous learning and adaptation [27], [39], [56]
- Ethical considerations and transparency [26], [55], [46]
- Balancing automation and oversight [31], [48] [59]
- Significance of contextual understanding in AI [28], [33]

IV. ANALYSIS OF GEMMA2:9B FILTERING RESULT

Our analysis, based on LLaAssist’s output applied to Scopus database results, uses a scoring system (0-1) for relevance and contribution, with 0.7 as the threshold for classification. Tables I and II summarize relevant and contributing papers from 2015-2024, respectively. Figure 2 visualizes the distribution of both categories. In the tables, we define: 1) *SQ*: Screening Question 2) *R*: Relevant papers, i.e. papers discussing topics related to the screening questions 3) *C*: Contributing papers, i.e. papers directly researching topics in the screening questions 4) *Any SQ*: Papers relevant to or contributing to at least one screening question 5) *All SQs*: Papers relevant to or contributing to all screening questions

The relevance criterion assesses whether a paper discusses topics related to our research questions, while the contribution criterion evaluates whether a paper directly researches these topics. We use a score threshold of 0.7 for both criteria to determine if a paper is considered relevant or contributing. The analysis reveals significant growth in research attention, with relevant papers increasing from 5 (2015) to 150 (2024), and contributing papers from 0 to 25 over the same period.

TABLE I
DISTRIBUTION OF RELEVANT PAPERS (R) BY YEAR AND SQ

Year	Total	Any SQ	All SQs	SQ1	SQ2	SQ3	SQ4
2015	81	5	0	2	2	3	0
2016	125	6	0	1	2	5	2
2017	167	5	1	3	1	4	1
2018	255	14	2	9	9	8	3
2019	315	32	1	19	10	21	3
2020	380	63	3	31	30	46	13
2021	485	50	9	31	29	36	12
2022	720	88	5	40	51	61	12
2023	1078	193	18	92	116	105	41
2024	625	150	27	87	102	87	44
Total	4231	606	66	315	352	376	131

TABLE II
DISTRIBUTION OF CONTRIBUTING PAPERS (C) BY YEAR AND SQ

Year	Total	Any SQ	All SQs	SQ1	SQ2	SQ3	SQ4
2015	81	0	0	0	0	0	0
2016	125	2	0	1	0	1	0
2017	167	1	0	0	0	1	0
2018	255	2	0	1	0	1	0
2019	315	3	0	3	0	1	0
2020	380	7	0	4	0	3	0
2021	485	9	0	6	1	8	0
2022	720	9	0	6	0	3	0
2023	1078	32	0	15	3	19	1
2024	625	25	0	16	2	11	3
Total	4231	90	0	52	6	48	4

A. Focus of Research

Papers on robustness and reliability received the most attention (376 relevant, 48 contributing), followed by integration methods and strategic implementation factors. Organizational measures and governance frameworks for trust-building received the least attention (131 relevant, 4 contributing).

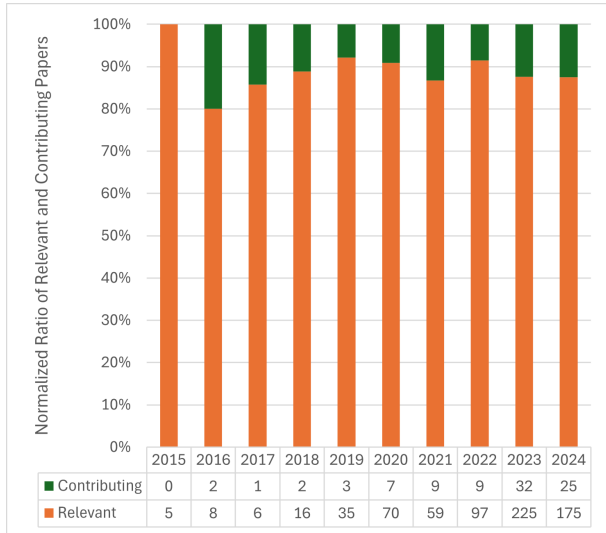


Fig. 2. Ratio between Relevant and Contributing Papers

B. Recent Developments

2023-2024 saw accelerated research activity across all screening questions. 2023 produced 193 relevant and 32 contributing papers, while 2024 already shows 150 relevant and 25 contributing papers, indicating rapidly growing interest.

C. Research Gaps

SQ4 in Tables I and II reveals a significant gap in research on organizational measures and governance frameworks, suggesting opportunities for future research on trust-building aspects of AI adoption in cyber security.

D. Field Maturity

The ratio of contributing papers to relevant papers (14.9%) gives insight into the field's maturity. While there is growing interest, as evidenced by the increase in relevant papers, the slower emergence of contributing papers suggests that it is still developing with considerable potential for in-depth studies.

E. Evolution of Research Focus

2015-2017 saw minimal relevant or contributing papers. 2018-2021 showed gradual increase, especially in strategic factors, integration methods, and robustness techniques. 2022-2024 demonstrated significant attention growth across all questions, particularly in robustness and reliability techniques.

F. Summary of the Analysis

This analysis highlights rapid growth and evolution in the research. While attention increases across all aspects, more focused research on organizational measures and governance frameworks for trust-building is needed. Recent surges suggest accelerated development with the potential for significant future advancements. For deeper insights into recent impactful research, we conducted a full-text review of selected papers, presented in the following chapter.

V. FULL-TEXT ANALYSIS USING CLAUDE 3.5 SONNET

Based on the selection criteria, we have short-listed and obtained the full text of 58 research papers between 2020 and 2024. From processing each of the research papers using Claude 3.5 Sonnet, we extracted their type, key themes, author stances, and concise summary of their research results to the RQs and definitions provided as part of the prompt. The key insights are summarized in table III and future research direction is outlined in table IV.

Note that during our manual review of the 58 research papers, we encountered several things where our judgment was misaligned with AI, notably [60], [95], [96], [77], [86]:

- There is one occurrence where we deemed the paper is insightful and can contribute towards the robustness [60] yet LLM determined the paper as not sufficiently contributing. After careful inspection, we noted that the paper was from 2020, before the advent of LLM-style deep learning. Hence, LLM's determination is correct.
- There are two occasions where we failed to identify future research directions [95], [96]. LLM was correct to

TABLE III
KEY INSIGHTS FOR RESEARCH QUESTIONS

Key Insights	References
RQ1: Key strategic factors to consider	
AI can significantly enhance system but with new risks	[10], [61], [62]
Data quality and availability for implementation	[63], [64]
Interoperability with existing security infrastructure	[65]
Carefully evaluate ethical and legal implications	[66], [65]
Cost-benefit analysis is essential for decision-making	[65]
Human oversight and expertise remains critical	[67], [68]
RQ2: Key integration approaches while preventing overreliance	
Using AI for automated threat detection and response	[10], [69], [70]
AI-driven anomaly detection enhances security	[71], [72]
AI can assist in vulnerability assessment and patching	[73], [74]
Cyber deception techniques can be enhanced with AI	[75], [65]
Human oversight is important to prevent overreliance	[67], [68], [76]
Explainable AI improve trust and understanding	[77], [78], [79]
Hybrid of AI and traditional methods are effective	[80], [81]
RQ3: Key methods and practices for AI system robustness	
Adversarial training and testing improve AI robustness	[82], [83], [84]
Calibrated uncertainty quantification add reliability	[85], [86]
Deep ensembles and temperature scaling help performance	[86]
Need continuous monitoring and adaptation of AI models	[87], [64]
Explainable AI aid in verification and debugging	[77], [79]
Multiple AI models and perspectives add robustness	[76], [78]
RQ4: Key organizational measures and governance frameworks	
Need clear policies for AI use in cyber security	[88], [89], [90]
Governance needs cross-sector work and standards	[63], [66]
Regular security assessments and audits of AI systems	[65]
Transparency and explainability of AI decisions build trust	[77], [78]
Compliance with data protection and privacy regulations	[90], [91]
Implement ethical guidelines for AI use in cyber security	[10], [66]

determine that fake cyber threat intelligence needs deeper research.

- There is one occurrence where we misidentified the necessity of privacy-preserving methods in an explainable AI system [77], dismissing its importance while LLM was correctly identified it as important for future research.
- There is one occurrence where we did not fully understand the key insight of using deep ensembles and temperature scaling [86]. Upon investigation, the team did not understand those specific terminologies and LLM was correctly identifying them.

VI. DISCUSSIONS AND RECOMMENDATIONS

A. Synthesis of Findings

Our review reveals a rapidly evolving landscape, with research attention significantly increasing from 2015 to 2024. Key areas requiring further attention:

- 1) **Research Focus Imbalance:** Substantial research exists on technical aspects (robustness, integration), but a gap persists in studies on organizational measures and governance frameworks for building trust in AI-enhanced cyber security solutions.

TABLE IV
RESEARCH GAPS AND FUTURE DIRECTIONS

Identified Gaps and Future Research Directions	References
RQ1: Future research for strategic decision making	
Develop a comprehensive decision-making framework	[84], [90]
Investigate the long-term impacts of AI adoption	[62], [10]
Interplay of AI and evolving threat landscapes	[71], [92]
RQ2: Future research for integration approaches	
Frameworks for balanced human-AI collaboration	[78], [93], [94]
AI-generated fake cyber threat intelligence	[95], [96]
Explore adaptive AI that evolve with threat landscapes	[63], [64]
RQ3: Future research for AI robustness in cyber security	
Benchmarks and metrics for AI in cyber security	[84], [97], [98]
Transfer learning and meta-learning approaches	[75]
Privacy-preserving AI for cyber security applications	[10], [77]
Defending against adversarial attacks on AI models	[82], [99], [100]
RQ4: Future research for organizational readiness	
Governance frameworks for AI in cyber security	[90], [101]
Evaluation and certification of AI-driven solutions	[97], [98]
Public trust in AI-enhanced cyber security measures	[102], [10]
Impact of AI on cyber security workforce training	[69], [88]

- 2) **Rapid Advancements and Field Maturity:** Research surge from 2022 to 2024 indicates accelerating developments, particularly in integration methods and robustness techniques. The low ratio of contributing to relevant papers suggests a developing field with potential for further substantive contributions.

B. Methodological Analysis and Reflection

Our AI-assisted exploration and full-text review approach offers both advantages and challenges. Here's a comparison of methodology A (GPT-4 for exploration) and B (Gemma 2:9b for filtering/searching and Claude 3.5 Sonnet for full-text analysis):

- **Immediacy:** A provides immediate thematic review, while B requires a database and full-text access.
- **Efficiency/Breadth:** A processes diverse sources, B focuses on academic literature only.
- **Structure:** A relies on GPT-4's and its search engine result, B uses a consistent assessment framework that can be designed by the researcher.
- **Information Uncovering:** A offers unique cross-source insights, B excels at detailed full-text extraction.
- **Bias Mitigation:** A has both model and search engine bias risk, B has model and academic database bias risk.
- **Context:** A uses general knowledge and instructions, B requires specific prompts and definitions.
- **Depth:** A may sacrifice depth for breadth, B allows in-depth full-text analysis.
- **False Negatives:** A may miss sources not highly ranked in search engine, B may miss non-matching papers.
- **False Positives:** A may include irrelevant sources due to broad interpretations, B minimizes this through filtration.
- **Academic Rigor:** A includes non-academic sources, B focuses on peer-reviewed literature.

Both methodologies complement traditional reviews. Future work should refine these techniques while maintaining rigor, potentially:

- Developing sophisticated AI prompting strategies
- Using multiple AI models for cross-validation
- Establishing AI-human analysis integration protocols
- Combining strengths of both methodologies

C. Implications and Future Directions

Table IV outlines specific research gaps and future directions. Our analysis reveals rapid field evolution (2022-2024), the need for interdisciplinary approaches, and opportunities for meta-research on AI-assisted systematic reviews in cybersecurity. Future work should prioritize addressing these gaps while balancing innovation and practical implementation.

VII. CONCLUSION AND FUTURE WORKS

This survey examined contextualized AI's potential in reshaping cyber defense strategies using a novel AI-assisted methodology. Key findings include significant research attention growth (2015-2024), focus on robustness, reliability, and integration methods, with gaps in organizational trust and governance studies. Our AI-assisted approach demonstrated efficiency in processing diverse sources, highlighting the potential of such methods in comprehensive literature reviews. Future research should prioritize empirical studies comparing traditional and AI-enhanced systems, exploring adaptive AI for evolving threats, and developing governance frameworks. While contextualized AI promises enhanced cyber defense capabilities, effective implementation requires balancing AI strengths with human oversight and risk mitigation. As this field rapidly evolves, interdisciplinary collaboration among cyber security experts, AI researchers, and policymakers will be crucial in addressing the multifaceted challenges of AI contextualization in cyber defense.

REFERENCES

- [1] G. Pinto, C. De Souza, T. Rocha, I. Steinmacher, A. Souza, and E. Monteiro, "Developer Experiences with a Contextualized AI Coding Assistant: Usability, Expectations, and Outcomes," in Proc. IEEE/ACM 3rd Int. Conf. AI Eng. - Softw. Eng. AI, Lisbon Portugal: ACM, Apr. 2024, pp. 81–91. doi: 10.1145/3644815.3644949.
- [2] B. Ahmad, S. Thakur, B. Tan, R. Karri, and H. Pearce, "Fixing Hardware Security Bugs with Large Language Models," 2023, doi: 10.48550/ARXIV.2302.01215.
- [3] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, "Examining Zero-Shot Vulnerability Repair with Large Language Models," in 2023 IEEE Symp. Security and Privacy (SP), San Francisco, CA, USA: IEEE, May 2023, pp. 2339–2356. doi: 10.1109/SP46215.2023.10179420.
- [4] R. Sasaki, "AI and Security - What Changes with Generative AI," in 2023 IEEE 23rd Int. Conf. Softw. Qual., Rel., and Security Companion (QRS-C), Chiang Mai, Thailand: IEEE, Oct. 2023, pp. 208–215. doi: 10.1109/QRS-C60940.2023.00043.
- [5] J. F. Loevenich, E. Adler, R. Mercier, A. Velazquez, and R. R. F. Lopes, "Design of an Autonomous Cyber Defence Agent using Hybrid AI models," in 2024 Int. Conf. Military Communication and Information Systems (ICMCIS), Apr. 2024, pp. 1–10. doi: 10.1109/ICMCIS61231.2024.10540988.
- [6] S. Bubeck et al., "Sparks of Artificial General Intelligence: Early experiments with GPT-4," arXiv, 2023. doi: 10.48550/ARXIV.2303.12712.
- [7] Y. Liu et al., "Summary of ChatGPT-Related research and perspective towards the future of large language models," Meta-Radiology, vol. 1, no. 2, p. 100017, Sep. 2023, doi: 10.1016/j.metrad.2023.100017.
- [8] N. Carlini et al., "Extracting Training Data from Large Language Models," arXiv, 2020. doi: 10.48550/ARXIV.2012.07805.
- [9] D. Ganguli et al., "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned," arXiv, 2022. doi: 10.48550/ARXIV.2209.07858.
- [10] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy," IEEE Access, vol. 11, pp. 80218–80245, 2023, doi: 10.1109/ACCESS.2023.3300381.
- [11] S. Oh and T. Shon, "Cybersecurity Issues in Generative AI," in 2023 Int. Conf. Platform Technol. Service (PlatCon), Busan, Korea, Republic of: IEEE, Aug. 2023, pp. 97–100. doi: 10.1109/PlatCon60102.2023.10255179.
- [12] K. Sanderson, "GPT-4 is here: what scientists think," Nature, vol. 615, no. 7954, pp. 773–773, Mar. 2023, doi: 10.1038/d41586-023-00816-5.
- [13] C. Y. Haryanto, M. H. Vu, T. D. Nguyen, E. Lomempow, Y. Nurliana, and S. Taheri, "SecGenAI: Enhancing Security of Cloud-based Generative AI Applications within Australian Critical Technologies of National Interest," arXiv, Jul. 01, 2024.
- [14] S. Agarwal, I. H. Laradji, L. Charlin, and C. Pal, "LitLLM: A Toolkit for Scientific Literature Review," arXiv, 2024. doi: 10.48550/ARXIV.2402.01788.
- [15] L. Joos, D. A. Keim, and M. T. Fischer, "Cutting Through the Clutter: The Potential of LLMs for Efficient Filtration in Systematic Literature Reviews," arXiv, 2024. doi: 10.48550/ARXIV.2407.10652.
- [16] C. Y. Haryanto, "LLAssist: Simple Tools for Automating Literature Review Using Large Language Models," arXiv, 2024. doi: 10.48550/ARXIV.2407.13993.
- [17] "GPT-4," OpenAI. Available: <https://openai.com/index/gpt-4-research/>
- [18] S. G. Prasad, V. C. Sharmila, and M. K. Badrinarayanan, "Role of Artificial Intelligence based Chat Generative Pre-trained Transformer (ChatGPT) in Cyber Security," in 2023 2nd Int. Conf. Applied Artif. Intell. and Comput. (ICAAIC), Salem, India: IEEE, May 2023, pp. 107–114. doi: 10.1109/ICAAIC56838.2023.10141395.
- [19] J. White et al., "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," arXiv, Feb. 21, 2023. Available: <http://arxiv.org/abs/2302.11382>.
- [20] S. P. Mohammed and G. Hossain, "ChatGPT in Education, Healthcare, and Cybersecurity: Opportunities and Challenges," in 2024 IEEE 14th Annu. Comput. Commun. Workshop Conf. (CCWC), Las Vegas, NV, USA, Jan. 2024, pp. 316–321. doi: 10.1109/CCWC60891.2024.10427923
- [21] "gemma2:9b," Ollama. Available: <https://ollama.com/library/gemma2:9b>
- [22] "Introducing Claude 3.5 Sonnet." Available: <https://www.anthropic.com/news/claude-3-5-sonnet>
- [23] "cyharyanto/lassist at 3bf51a695b945e07c77eaa0a323c9aa3e57372bd." Available: <https://github.com/cyharyanto/lassist>
- [24] A. Askell et al., "A General Language Assistant as a Laboratory for Alignment," 2021, arXiv. doi: 10.48550/ARXIV.2112.00861.
- [25] "Prompt engineering overview - Anthropic," Anthropic. <https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview>
- [26] M. Malatji and A. Tolah, "Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI," AI Ethics, Feb. 2024, doi: 10.1007/s43681-024-00427-4.
- [27] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "AI-Driven Cybersecurity: An Overview, security intelligence modeling and research directions," SN Computer Science, vol. 2, no. 3, Mar. 2021, doi: 10.1007/s42979-021-00557-0.
- [28] P. Ranade, A. Piplai, A. Joshi, and T. Finin, "CyBERT: Contextualized embeddings for the cybersecurity domain," UMBC Ebiquty Research Lab, Dec. 15, 2021. Available: <https://ebiquty.umbc.edu/paper/html/id/999/CyBERT-Contextualized-Embeddings-for-the-Cybersecurity-Domain>
- [29] I. H. Sarker, H. Janicke, N. Mohammad, P. Watters, and S. Nepal, "AI Potentiality and Awareness: A Position Paper from the Perspective of Human-AI Teaming in Cybersecurity," arXiv.org, Sep. 28, 2023. Available: <https://arxiv.org/abs/2310.12162>

- [30] "What is AI in Cybersecurity? — AI Cybersecurity Explained," SOPHOS. Available: <https://www.sophos.com/en-us/cybersecurity-explained/ai-in-cybersecurity>
- [31] "Transforming the Future of Cybersecurity with an AI-Driven Approach - Wipro." Available: <https://www.wipro.com/cybersecurity/eliminating-the-complexity-in-cybersecurity-with-artificial-intelligence/>
- [32] "Enhancing Cybersecurity through AI: A Look into the Future." Available: <https://www.isc2.org/Insights/2023/09/Enhancing-Cybersecurity-through-AI-A-Look-into-the-Future>
- [33] Z. Zhang et al., "Artificial intelligence in cyber security: research advances, challenges, and opportunities," *Artif. Intell. Review*, vol. 55, no. 2, pp. 1029–1053, Mar. 2021, doi: 10.1007/s10462-021-09976-0.
- [34] M. Stamp, C. Aaron Visaggio, F. Mercaldo, and F. Di Troia, Eds., *Artificial Intelligence for Cybersecurity*, vol. 54. in *Adv. Inf. Secur.*, vol. 54. Cham: Springer, 2022, doi: 10.1007/978-3-030-97087-1.
- [35] Y. Yigit, W. J. Buchanan, M. G. Tehrani, and L. Maglaras, "Review of Generative AI methods in Cybersecurity," *arXiv.org*, Mar. 13, 2024. Available: <https://arxiv.org/abs/2403.08701>
- [36] Y. Hong et al., "Statistical perspectives on reliability of artificial intelligence systems," *arXiv.org*, Nov. 09, 2021. Available: <https://arxiv.org/abs/2111.05391>
- [37] J. Fan, Z. Ye, C. Guan, X. Gao, K. Ren, and C. Qiao, "GREP: Guaranteeing Reliability with Enhanced Protection in NFV," in *Proc. 2015 ACM SIGCOMM Workshop on Hot Topics in Middleboxes and Network Function Virtualization*, London United Kingdom: ACM, Aug. 2015, pp. 13–18. doi: 10.1145/2785989.2786000.
- [38] Cp, "AI Security and Risk Management: Strategies for safeguarding artificial intelligence systems," *AI Consultancy — Create Progress*, Apr. 15, 2024. Available: <https://createprogress.ai/ai-security-and-risk-management-strategies-for-safeguarding-artificial-intelligence-systems/>
- [39] "AI Safety vs. AI Security: Navigating the Differences — CSA," Mar. 19, 2024. Available: <https://cloudsecurityalliance.org/blog/2024/03/19/ai-safety-vs-ai-security-navigating-the-commonality-and-differences>
- [40] J. Min, Y. Hong, C. B. King, and W. Q. Meeker, "Reliability Analysis of Artificial Intelligence Systems Using Recurrent Events Data from Autonomous Vehicles," *J. Roy. Statistical Soc. Series C (Applied Statistics)*, vol. 71, no. 4, pp. 987–1013, Apr. 2022, doi: 10.1111/rssc.12564.
- [41] "AI systems engineering and reliability technologies," *MIT Lincoln Laboratory*. Available: <https://www.ll.mit.edu/r-d/projects/ai-systems-engineering-and-reliability-technologies>
- [42] A. Ingason, "Building Trust in AI: A Comprehensive guide to responsible and reliable predictive systems," *Sumo Analytics*, Jun. 12, 2023. Available: <https://www.sumoanalytics.ai/post/building-trust-in-ai-a-comprehensive-guide-to-responsible-and-reliable-predictive-systems>
- [43] "AI Governance Strategy, Framework & Best Practices: The Ultimate Guide," *NextGen Invent Corporation*, Jun. 28, 2024. Available: <https://nextgeninvent.com/blogs/ai-governance-framework-best-practices/>
- [44] S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, and M. L. Tielman, "A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction," 2023, *arXiv*. doi: 10.48550/ARXIV.2311.06305.
- [45] Q. V. Liao and S. S. Sundar, "Designing for Responsible Trust in AI Systems: A Communication Perspective," 2022 *ACM Conf. Fairness, Accountability, and Transparency*, Jun. 2022, doi: 10.1145/3531146.3533182.
- [46] M. Mylrea and N. Robinson, "Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an entropy lens to improve security, privacy, and ethical AI," *Entropy*, vol. 25, no. 10, p. 1429, Oct. 2023, doi: 10.3390/e25101429.
- [47] "Deloitte Generative AI Survey finds Adoption is Moving Fast, but Organizational Change is Key to Accelerate Scaling," *Deloitte United States*. Available: <https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-generative-ai-survey-finds-adoption-is-moving-fast-but-organizational-change-is-key-to-accelerate-scaling.html>
- [48] "Workday Global survey reveals AI trust gap in the workplace," *Newsroom — Workday*, Jan. 10, 2024. Available: <https://newsroom.workday.com/2024-01-10-Workday-Global-Survey-Reveals-AI-Trust-Gap-in-the-Workplace>
- [49] "Exploring ways to regulate and build trust in AI," *RAND*. Available: <https://www.rand.org/randeurope/research/projects/2022/exploring-ways-to-regulate-and-build-trust-in-artificial-intelli.html>
- [50] K. Reinhardt, "Trust and trustworthiness in AI ethics," *AI Ethics*, vol. 3, no. 3, pp. 735–744, Sep. 2022, doi: 10.1007/s43681-022-00200-5.
- [51] "Building trust in AI: How to overcome risk and operationalize AI governance," *Deloitte Canada*. Available: <https://www2.deloitte.com/ca/en/pages/financial-services/articles/bridging-operationalizing-trust-in-ai.html>
- [52] "Building trust in artificial intelligence, machine learning, and robotics — Cutter Consortium." Available: <https://www.cutter.com/article/building-trust-artificial-intelligence-machine-learning-and-robotics-498981>
- [53] H. Choung, P. David, and A. Ross, "Trust and ethics in AI," *AI & Society*, vol. 38, no. 2, pp. 733–745, May 2022, doi: 10.1007/s00146-022-01473-4.
- [54] R. Yang and S. Wibowo, "User trust in artificial intelligence: A comprehensive conceptual framework," *Electronic Markets*, vol. 32, no. 4, pp. 2053–2077, Nov. 2022, doi: 10.1007/s12525-022-00592-6.
- [55] "Trust in #AI: Why the right foundations will determine its future," *World Economic Forum*, Jan. 02, 2024. Available: <https://www.weforum.org/agenda/2024/01/davos24-trust-ai-right-foundations-determine-its-future/>
- [56] "State of Generative AI in the enterprise 2024," *Deloitte United States*. Available: <https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-generative-ai-in-enterprise.html>
- [57] CrowdStrike, "The role of AI in cybersecurity — CrowdStrike," *crowdstrike.com*, May 31, 2024. Available: <https://www.crowdstrike.com/cybersecurity-101/artificial-intelligence/>
- [58] A. Fitzgerald, "AI in Cybersecurity: How It's Used + 8 Latest Developments," *SecureFrame*, May 06, 2024. Available: <https://secureframe.com/blog/ai-in-cybersecurity>
- [59] V. A. S. R. Team, "2024 Predictions: Generative AI's role in Cybersecurity," *Vectra AI*, Jun. 25, 2024. Available: <https://www.vectra.ai/blog/2024-predictions-generative-ais-role-in-cybersecurity>
- [60] M. Blowers and J. Williams, "Artificial intelligence presents new challenges in cybersecurity," in *Disruptive Technologies in Information Sciences IV*, M. Blowers, R. D. Hall, and V. R. Dasari, Eds., Online Only, United States: SPIE, May 2020, p. 19. doi: 10.1117/12.2560002.
- [61] Orner and Md. Chowdhury, "AI and Cybersecurity: Collaborator or Confrontation," in *Proc. 39th Int. Conf. Computers and Their Applications*, pp. 150–140. doi: 10.29007/q3md.
- [62] A. Bécue, I. Praça, and J. Gama, "Artificial intelligence, cyber-threats and Industry 4.0: challenges and opportunities," *Artif. Intell. Rev.*, vol. 54, no. 5, pp. 3849–3886, Jun. 2021. doi: 10.1007/s10462-020-09942-2
- [63] E. M. Timofte, A. L. Balan, and T. Iftime, "AI Driven Adaptive Security Mesh: Cloud Container Protection for Dynamic Threat Landscapes," in *2024 Int. Conf. Dev. Appl. Syst. (DAS)*, Suceava, Romania, May 2024, pp. 71–77. doi: 10.1109/DAS61944.2024.10541148
- [64] T. Cody and P. A. Beling, "Towards operational resilience for AI-based cyber systems in multi-domain operations," in *Artif. Intell. Mach. Learn. Multi-Domain Oper. Appl. V*, L. Solomon and P. J. Schwartz, Eds., Orlando, USA: SPIE, Jun. 2023, p. 68. doi: 10.1117/12.2675862
- [65] K. Eng, J. King, C. Schillaci, and A. Rawal, "The relevance, effectiveness, and future prospects of cyber deception implementation within organizations," in *Assur. Secur. AI-enabled Syst.*, J. D. Harguess, N. D. Bastian, and T. L. Pace, Eds., Nat. Harbor, USA: SPIE, Jun. 2024, p. 20. doi: 10.1117/12.3013114
- [66] J.-M. Lee and S. Yoon, "Ready for Battle?: Legal Considerations for Upcoming AI Hacker and Vulnerability Issues," *FLAIRS*, vol. 35, May 2022. doi: 10.32473/flairs.v35i.130673
- [67] C. E. Whyte, "Machine Expertise in the Loop: Artificial Intelligence Decision-Making Inputs and Cyber Conflict," in *2022 14th Int. Conf. Cyber Confl.: Keep Moving! (CyCon)*, Tallinn, Estonia, May 2022, pp. 135–154. doi: 10.23919/CyCon55549.2022.9811076
- [68] C. Whyte, "Learning to trust Skynet: Interfacing with artificial intelligence in cyberspace," *Contemp. Secur. Policy*, vol. 44, no. 2, pp. 308–344, Apr. 2023. doi: 10.1080/13523260.2023.2180882
- [69] M. Al-Hawawreh, A. Aljuhani, and Y. Jararweh, "Chatgpt for cybersecurity: practical applications, challenges, and future directions," *Cluster Comput.*, vol. 26, no. 6, pp. 3421–3436, Dec. 2023. doi: 10.1007/s10586-023-04124-5
- [70] S. Sai, U. Yashvardhan, V. Chamola, and B. Sikdar, "Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E, and

- Other Models for Enhancing the Security Space," *IEEE Access*, vol. 12, pp. 53497–53516, 2024. doi: 10.1109/ACCESS.2024.3385107
- [71] C. Benzaid and T. Taleb, "AI for Beyond 5G Networks: A Cyber-Security Defense or Offense Enabler?," *IEEE Netw.*, vol. 34, no. 6, pp. 140–147, Nov. 2020. doi: 10.1109/MNET.011.2000088
- [72] F. Jaafar, D. Ameyed, L. Titare, and M. Nematullah, "IoT Phishing Detection Using Hybrid NLP and Machine Learning Models Enhanced with Contextual Embedding," in 2023 IEEE 23rd Int. Conf. Softw. Qual., Rel., and Secur. Companion (QRS-C), Chiang Mai, Thailand, Oct. 2023, pp. 340–349. doi: 10.1109/QRS-C60940.2023.00088
- [73] J. Li, A. Sangalay, C. Cheng, Y. Tian, and J. Yang, "Fine Tuning Large Language Model for Secure Code Generation," in Proc. 2024 IEEE/ACM 1st Int. Conf. AI Found. Models Softw. Eng., Lisbon, Portugal, Apr. 2024, pp. 86–90. doi: 10.1145/3650105.3652299
- [74] T. E. Gasiba, K. Oguzhan, I. Kessba, U. Lechner, and M. Pinto-Albuquerque, "I'm Sorry Dave, I'm Afraid I Can't Fix Your Code: On ChatGPT, CyberSecurity, and Secure Coding," *OASICS*, vol. 112, pp. 2:1–2:12, 2023. doi: 10.4230/OASICS.ICPEC.2023.2
- [75] D. L. Antunes and S. L. Sanchez, "The Age of fighting machines: the use of cyber deception for Adversarial Artificial Intelligence in Cyber Defence," in Proc. 18th Int. Conf. Avail., Rel. and Secur., Benevento, Italy, Aug. 2023, pp. 1–6. doi: 10.1145/3600160.3605077
- [76] P. Baroni et al., "Self-Aware Effective Identification and Response to Viral Cyber Threats," in 2021 13th Int. Conf. Cyber Confl. (Cy-Con), Tallinn, Estonia, May 2021, pp. 353–370. doi: 10.23919/Cy-Con51939.2021.9468294
- [77] A. Nadeem et al., "SoK: Explainable Machine Learning for Computer Security Applications," in 2023 IEEE 8th Eur. Symp. Secur. Priv. (EuroS&P), Delft, Netherlands, Jul. 2023, pp. 221–240. doi: 10.1109/EuroSP57164.2023.00022
- [78] A. Zagalsky et al., "The Design of Reciprocal Learning Between Human and Artificial Intelligence," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, Oct. 2021, Art. no. 3479587. doi: 10.1145/3479587
- [79] M.-D. Nguyen, V. H. La, R. Cavalli, and E. M. De Oca, "Towards improving explainability, resilience and performance of cybersecurity analysis of 5G/IoT networks," in 2022 IEEE Int. Conf. Softw. Test., Verif. Valid. Workshops (ICSTW), Valencia, Spain, Apr. 2022, pp. 7–10. doi: 10.1109/ICSTW55395.2022.00016
- [80] P. E. Tsareva and A. V. Voronova, "Information Security Systems Based on the AI and Machine Learning," in 2022 Conf. Russ. Young Res. Electr. Electron. Eng. (ElConRus), St. Petersburg, Russia, Jan. 2022, pp. 469–473. doi: 10.1109/ElConRus54750.2022.9755581
- [81] A. Piplai, A. Kotal, S. Mohseni, M. Gaur, S. Mittal, and A. Joshi, "Knowledge-Enhanced Neurosymbolic Artificial Intelligence for Cybersecurity and Privacy," *IEEE Internet Comput.*, vol. 27, no. 5, pp. 43–48, Sep. 2023. doi: 10.1109/MIC.2023.3299435
- [82] I. Kotenko, I. Saenko, O. Lauta, N. Vasiliev, and D. Iatsenko, "Attacks Against Machine Learning Systems: Analysis and GAN-based Approach to Protection," in Proc. 7th Int. Sci. Conf. "Intell. Inf. Technol. Ind." (ITI'23), in Lect. Notes Netw. Syst., vol. 777, Cham: Springer, 2023, pp. 49–59. doi: 10.1007/978-3-031-43792-2_5
- [83] K. Kumar, Kuldeep, and B. Bhushan, "Augmenting Cybersecurity and Fraud Detection Using Artificial Intelligence Advancements," in 2023 Int. Conf. Comput., Commun., Intell. Syst. (ICCCIS), Greater Noida, India, Nov. 2023, pp. 1207–1212. doi: 10.1109/ICCCIS60361.2023.10425069
- [84] J. Malik, R. Muthalagu, and P. M. Pawar, "A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls and Technologies," *IEEE Access*, 2024. doi: 10.1109/ACCESS.2024.3423323
- [85] Y. Fourastier, C. Baron, C. Thomas, and P. Esteban, "Assurance levels for decision making in autonomous intelligent systems and their safety," in 2020 IEEE 11th Int. Conf. Dependable Systems, Services and Technologies (DESSERT), Kyiv, Ukraine: IEEE, May 2020, pp. 475–483. doi: 10.1109/DESSERT50317.2020.9125079
- [86] D. Woodward, M. Hobbs, J. A. Gilbertson, and N. Cohen, "Uncertainty Quantification for Trusted Machine Learning in Space System Cyber Security," in 2021 IEEE 8th Int. Conf. Space Mission Challenges for Information Technology (SMC-IT), Pasadena, CA, USA: IEEE, Jul. 2021, pp. 38–43. doi: 10.1109/SMC-IT51442.2021.00012
- [87] C. Sample, S. M. Loo, and M. Bishop, "Resilient Data: An Interdisciplinary Approach," in 2020 Resilience Week (RWS), Salt Lake City, ID, USA: IEEE, Oct. 2020, pp. 1–10. doi: 10.1109/RWS50334.2020.9241268
- [88] M. Charfeddine et al., "ChatGPT's Security Risks and Benefits: Offensive and Defensive Use-Cases, Mitigation Measures, and Future Implications," *IEEE Access*, vol. 12, pp. 30263–30310, 2024. doi: 10.1109/ACCESS.2024.3367792
- [89] R. Pasupuleti, R. Vadapalli, and C. Mader, "Cyber Security Issues and Challenges Related to Generative AI and ChatGPT," in 2023 10th Int. Conf. Soc. Netw. Anal., Manag. Secur. (SNAMS), Abu Dhabi, UAE, Nov. 2023, pp. 1–5. doi: 10.1109/SNAMS60348.2023.10375472
- [90] T. R. McIntosh et al., "From COBIT to ISO 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models," *Comput. Secur.*, vol. 144, Sep. 2024, Art. no. 103964. doi: 10.1016/j.cose.2024.103964
- [91] T. Sean and M. Leighton, "Towards Modelling Artificial Intelligence Parsing documents for Cyber-Policing Data protection and privacy environments using Privacy Impact Assessments and Data protection Impact assessment questionnaires," in SoutheastCon 2024, Atlanta, GA, USA, Mar. 2024, pp. 1106–1112. doi: 10.1109/Southeast-Con52093.2024.10500265
- [92] C. Whyte, "Problems of Poison: New Paradigms and 'Agreed' Competition in the Era of AI-Enabled Cyber Operations," in 2020 12th Int. Conf. Cyber Confl. (CyCon), Estonia, May 2020, pp. 215–232. doi: 10.23919/CyCon49761.2020.9131717
- [93] B. Strickson, C. Worsley, and S. Bertram, "Human-centered Assessment of Automated Tools for Improved Cyber Situational Awareness," in 2023 15th Int. Conf. Cyber Confl.: Meeting Reality (Cy-Con), Tallinn, Estonia, May 2023, pp. 273–286. doi: 10.23919/Cy-Con58705.2023.10181567
- [94] A. Chowdhury et al., "POSTER: A Teacher-Student with Human Feedback Model for Human-AI Collaboration in Cybersecurity," in Proc. ACM Asia Conf. Comput. Commun. Secur., Melbourne, Australia, Jul. 2023, pp. 1040–1042. doi: 10.1145/3579856.3592829
- [95] P. Ranade et al., "Generating Fake Cyber Threat Intelligence Using Transformer-Based Models," in 2021 Int. Joint Conf. Neural Netw. (IJCNN), Shenzhen, China, Jul. 2021, pp. 1–9. doi: 10.1109/IJCNN52387.2021.9534192
- [96] Z. Song et al., "Generating Fake Cyber Threat Intelligence Using the GPT-Neo Model," in 2023 8th Int. Conf. Intell. Comput. Signal Process. (ICSP), Xi'an, China, Apr. 2023, pp. 920–924. doi: 10.1109/ICSP58490.2023.10248596
- [97] G. Fenza, V. Loia, C. Stanzone, and M. Di Gisi, "Robustness of models addressing Information Disorder: A comprehensive review and benchmarking study," *Neurocomputing*, vol. 596, p. 127951, Sep. 2024, doi: 10.1016/j.neucom.2024.127951
- [98] S. S. Kumar, M. L. Cummings, and A. Stimpson, "Strengthening LLM Trust Boundaries: A Survey of Prompt Injection Attacks Suresh Kumar Dr. M.L. Cummings Dr. Alexander Stimpson," in 2024 IEEE 4th Int. Conf. Human-Machine Systems (ICHMS), Toronto, ON, Canada: IEEE, May 2024, pp. 1–6. doi: 10.1109/ICHMS59971.2024.10555871
- [99] A. Kuppa and N.-A. Le-Khac, "Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security," in 2020 Int. Joint Conf. Neural Networks (IJCNN), Glasgow, United Kingdom: IEEE, Jul. 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9206780
- [100] P. De Haan, I. Chiscop, B. Poppink, and Y. Kamphuis, "Evading deep learning-based DGA detectors: current problems and solutions," in *Artif. Intell. and Mach. Learn. for Multi-Domain Oper. Appl. VI*, P. J. Schwartz, M. E. Hohil, and B. Jensen, Eds., Nat. Harbor, USA: SPIE, Jun. 2024, p. 55. doi: 10.1117/12.3012997
- [101] S. A. A. Bokhari and S. Myeong, "The Influence of Artificial Intelligence on E-Governance and Cybersecurity in Smart Cities: A Stakeholder's Perspective," *IEEE Access*, vol. 11, pp. 69783–69797, 2023. doi: 10.1109/ACCESS.2023.3293480
- [102] M. R. Shoaib, Z. Wang, M. T. Ahvanooy, and J. Zhao, "Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models," in 2023 Int. Conf. Computer and Applications (ICCA), Cairo, Egypt: IEEE, Nov. 2023, pp. 1–7. doi: 10.1109/ICCA59364.2023.10401723