Literature Review

Application of LLM in Cyber Security

- From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy

1. Exploring the vulnerabilities and potential exploits of ChatGPT by malicious actors to bypass ethical constraints and exfiltrate malicious information.
2. Demonstrating successful attack methods like jailbreaks, reverse psychology, and prompt injection attacks on ChatGPT.
3. Investigating the use of GenAI tools by cyber offenders for developing various cyber-attacks, such as social engineering, phishing, automated hacking, malware creation, and polymorphic malware.
4. Examining defensive techniques and the use of GenAI tools to improve security measures, including cyber defense automation, threat intelligence, secure code generation, attack identification, and malware detection.
5. Discussing the social, legal, and ethical implications of ChatGPT, and highlighting the open challenges and future directions to make GenAI secure, safe, trustworthy, and ethical.

- ***LLM Agents can Autonomously Hack Websites***

1. Large language models (LLMs) can now function autonomously as agents, capable of interacting with tools, reading documents, and recursively calling themselves.
2. The work demonstrates that LLM agents can autonomously hack websites, performing complex tasks like blind database schema extraction and SQL injection without human feedback.
3. The agent does not need to know the vulnerability beforehand, enabled by frontier models highly capable of tool use and leveraging extended context.
4. GPT-4 is shown to be capable of such hacks, while existing open-source models are not.
5. GPT-4 can autonomously find vulnerabilities in websites in the wild, raising questions about the widespread deployment of LLMs.

- Red Team LLM: towards an adaptive and robust automation solution

1. Reinforcement learning agents can find optimal attack sequences on networks but lack adaptability and robustness to different networks.
2. The paper proposes a new agent based on a zero-shot approach that can adapt to any given network without additional training.
3. The proposed agent is robust to changes in parameters and objectives, not requiring further training.
4. A new metric is introduced to better measure the ability of agents to attack a network without prior knowledge.

5. The paper discusses the first steps towards explainability for the proposed model and its future improvements.

- Large Language Models in Cybersecurity: State-of-the-Art

1. Large Language Models (LLMs) have revolutionized our understanding of intelligence and brought us closer to Artificial Intelligence.
2. Researchers have actively explored the applications of LLMs across diverse fields, including cybersecurity.
3. The study provides a thorough characterization of defensive and adversarial applications of LLMs within cybersecurity.
4. The review surveys and categorizes the current landscape of LLM applications in cybersecurity and identifies critical research gaps.
5. The study aims to provide a holistic understanding of the potential risks and opportunities associated with LLM-driven cybersecurity by evaluating both offensive and defensive applications.

- AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions

1. The paper presents a comprehensive view on "AI-driven Cybersecurity," exploring the role of AI techniques in addressing cybersecurity issues.
2. Popular AI methods discussed include machine learning, deep learning, natural language processing, knowledge representation and reasoning, and knowledge-based expert systems.
3. AI-based security intelligence modeling can automate and enhance the cybersecurity computing process, making it more intelligent than conventional security systems.
4. Several research directions within the scope of the study are highlighted to guide future research in AI-driven cybersecurity.
5. The paper aims to serve as a reference point and guidelines for cybersecurity researchers and industry professionals, particularly from an intelligent computing or AI-based technical perspective.

- Using Large Language Models for Cybersecurity Capture-The Flag Challenges and Certification Questions

1. Large Language Models (LLMs) like ChatGPT, Bard, and Bing can perform well on many Capture-The-Flag (CTF) cybersecurity challenges by exploiting system vulnerabilities to find text "flags".
2. The availability of such powerful LLMs to students raises concerns about academic integrity in the context of CTF exercises in the classroom.
3. The research evaluates the effectiveness of three popular LLMs (ChatGPT, Bard, Bing) on CTF challenges and questions, assessing their question-answering performance on five Cisco certifications with varying difficulty levels.

4. A qualitative study examines the LLMs' abilities in solving different types of CTF challenges across seven test cases covering all five categories, understanding their limitations.
5. The research demonstrates how "jailbreak" prompts can bypass LLMs' ethical safeguards and discusses the implications of LLMs' impact on CTF exercises.

Current Cyber Security Challenge & Threat

- ***Web of cybersecurity: Linking, locating, and discovering structured cybersecurity information***

1. Timely accessibility to cybersecurity information is crucial for maintaining cybersecurity in organizations.
2. A mechanism is proposed to link, locate, and discover various cybersecurity information to improve its accessibility.
3. The mechanism generates metadata to manage a list of cybersecurity information with different schemata.
4. The information structure consists of linked categories and formats, making it flexible and extensible.
5. A prototype is introduced to demonstrate the mechanism's feasibility, and its extensibility, scalability, and information credibility are analyzed.

- Cybersecurity - Personal Security Agents for People, Process, Atoms & Bits

Proposes personal security agents (PSAs) as modular, layered tools to model people, processes, data and objects for enhanced cybersecurity.
Motivated by threats like the 2016 DDoS attack using IoT devices, potentially targeting critical systems like heart monitors.
Users would purchase security-as-a-service (SECaaS) from trusted third-party vendors like NGOs or standards bodies.
Device vendors would provide APIs to install multiple SECaaS layers, enabling user-controlled, redundant security.
Aims to make security layers difficult to penetrate by not originating from device vendors or being tied to a single cloud storage.

- ***Cyber Security Threats and Countermeasures in Digital Age***

1. Provides a detailed analysis of the cyberthreat environment in the digital era, covering various threats such as malware, phishing, ransomware, and insider threats.
2. Examines the continually evolving tactics employed by cybercriminals, including social engineering, zero-day exploits, and advanced persistent threats.
3. Highlights the emerging risks associated with new technologies like the Internet of Things (IoT), cloud computing, and artificial intelligence.
4. Emphasizes the need for a multi-layered security strategy, including robust network security, secure coding practices, user awareness training, encryption, access controls, and incident response planning.
5. Underscores the importance of collaboration among individuals, organizations, and governments to effectively address cyber risks and promote a culture of cybersecurity awareness.

- ***The intersection of Artificial Intelligence and cybersecurity: Challenges and opportunities***

1. Explores fundamental AI techniques (machine learning, natural language processing) and their applications in threat detection, vulnerability analysis, and incident response.
2. Contrasts traditional and AI-driven vulnerability analysis methodologies, highlighting advantages of automated scanning, threat prioritization, and adaptive risk assessment.
3. Discusses the pivotal role of AI-driven automation in expediting incident response, minimizing human error, and fortifying security postures.
4. Examines ethical and privacy concerns surrounding AI deployment in cybersecurity, emphasizing responsible decision-making, privacy protection, and transparency.
5. Explores emerging trends like adversarial machine learning and zero trust security as promising avenues for enhancing digital resilience against evolving threats.

Retrieved Augmented Generation (RAG)

- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

1. Large pre-trained language models have limited ability to access and precisely manipulate knowledge, despite storing factual knowledge and achieving state-of-the-art results on downstream NLP tasks.
2. Retrieval-Augmented Generation (RAG) models combine pre-trained parametric (seq2seq model) and non-parametric (Wikipedia index) memory for language generation.
3. Two RAG formulations are explored: one conditioning on the same retrieved passages across the entire sequence, and another using different passages per token.
4. RAG models set the state-of-the-art on three open-domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures.
5. For language generation tasks, RAG models generate more specific, diverse, and factual language than a state-of-the-art parametric-only seq2seq baseline.

- Evidentiality-guided Generation for Knowledge – Intensive NLP Tasks

1. Retrieval-augmented generation models can learn spurious cues or memorization when trained on irrelevant passages.
2. A method is introduced to incorporate the evidentiality of passages into training the generator.
3. A multi-task learning framework jointly generates the final output and predicts the evidentiality of each passage.
4. A new task-agnostic method is introduced for obtaining high-quality silver evidentiality labels.
5. Experiments show the evidentiality-guided generator significantly outperforms its counterpart and advances the state of the art on three tasks.

- LOCALINTEL: Generating Organizational Threat Intelligence from Global and Local Cyber Knowledge

1. SOC analysts manually contextualize global threat reports for their organization, which is labor-intensive, utilizing global threat databases and internal local knowledge repositories.
2. Large language models (LLMs) have shown the ability to efficiently process large and diverse knowledge sources.
3. LOCALINTEL is a novel automated system that leverages LLMs to contextualize global threat intelligence for a specific organization using both global and local knowledge sources.
4. The system has three key phases: retrieving global threat intelligence, retrieving relevant local knowledge, and generating contextualized completions by fusing these sources.

5. LOCALINTEL automates the generation of organization-specific threat intelligence, reducing manual effort for analysts by utilizing LLM capabilities on global and local knowledge.