Sri Lanka Institute of Information Technology

IE2022: Introduction to Cyber Security

Year 2, Semester 1

Large Language Model Security

IT23241978

S.D.W.Gunaratne

Table of Contents

1.Abstract

Generative artificial intelligence (AI) emerging in the recent period, especially large language models (LLMs), is important because they can create clear and relevant text, and it can revolutionize many industries.

The subsequent section of this report identifies and discusses major security threats that are inherent in LLMs as well as the risks posed by both non-use and utilization of AI. This also looks at the development of LLM security and the identification of new trends within this area. This report lays down strategies on how to counter such threats hence unauthorized use of LLMs and how to put mechanisms in place to work against such dangers across sectors to enhance security should be followed by good use of ethical frameworks in setting up the right security. So, bring considerable cybersecurity risks. This report examines the key security challenges associated with LLMs, including risks from both the absence and use of AI, vulnerabilities to attacks, and legal and regulatory concerns. It also analyzes the evolution of LLM security and highlights emerging developments in the field. By outlining strategies for effective countermeasures, including robust security frameworks and ethical guidelines, this report emphasizes the need for proactive measures to ensure the responsible use of LLMs and mitigate associated risks across sectors.

2. About LLMs

Large Language Models (LLMs) are the new technologies that revolutionize how machines learn and interact with an individual’s language. Constructed from a sort of neural network known as transformers, these models can both understand and synthesise textual content material effectively. LLMs can understand the manner of how the words are connected in the sentence using a feature referred to as self-attention.

From books and websites LLMs learn large sets of text data. This training enables them to learn rules of grammar and usage and the meaning of the words they use, this include. Transformers, allow models to look at an entire sentence at once unlike previous models that only read one word at a time. It is much faster at training, and indeed has been shown to give models with billions of parameters if the data can be handled. [1]

2. The Versatility of LLMs

This can constitute a major asset to the program in question, and LLMs generally are characterized by a significant measure of flexibility. They can solve quite many kinds of problems from query answering, text summarizing, language translating, and even programming. For example, OpenAI's GPT-3 model has 175 billion parameters. [2]

3. Challenges and Limitations

Although LLMs are very effective, they are not without shortcomings. They can sometime give wrong or prejudiced data if its training data is tainted with wrong information. Moreover, if LLMs fail to retrieve accurate answers, they may generate new information, so-called 'hallucinations ''. These issues underscore the fact that it is equally important to look at how LLMs are utilized, particularly when the right information needs to be obtained. [3]

4. This report aims to discuss the importance of LLM Security.

With an ever-increasing popularity of LLMs, issues of security are being brought forward. Security from LLM perspective involves protecting user data and making sure that these models leak no information. The inputs are entered by users which can be sensitive and therefore, the models should have measures against their leakage or any other dangers. This report will focus on topics such as what challenges arise when LLM is implemented in various disciplines, as well as how security for LLMs is an issue with these models being applied in more settings. Built on a type of neural network called transformers, these models can analyze and generate text effectively. Using a feature known as self-attention, LLMs can grasp how words relate to each other in a sentence, which helps them understand language more deeply.

LLMs are trained on large sets of text from different sources, like books and websites. This training allows them to learn rules of grammar, sentence structure, and word meanings. Unlike older models that read text one word at a time, transformers can look at whole sentences at once. This ability to process information in parallel makes training faster and allows for models with billions of parameters, enabling them to handle a lot of data.

3. The usage of Attackers to Large Language Models (LLMs)

While possessing their advantages, Large Language Models (LLMs) are vulnerable to misuse by attackers to develop malware and perform phishing attacks and to outsmart security measures in the LLMs. This note elaborates on how attackers can leverage LLMs or utilize LLMs employing certain attack methods including code generation, security bypass and improving spear phishing attacks. [4]

1. LLMs: Developing Malware

OpenAI's LLMs such as ChatGPT and Bard can be tricked into producing dangerous code. While these models have positive controls that are not meant to allow such toxicity, prompt engineering can get around them.

i)Malware Code Generation: This work proves that LLMs can be instructed to develop malware by guiding their code generation based on techniques from the MITRE ATT&CK framework. For example, a study showed that with minimal prompt engineering, ChatGPT was capable of generating executable code whereas Bard needed more complex scenario-based prompts.

ii)MITRE ATT&CK Framework: This is a well known knowledge base that describes the actions of an adversary. Through LLMs, threatening actors were able to develop code for what Picus Labs determined were the most common techniques observed, having scanned over 500,000 files and categorizing malicious actions according to the MITRE ATT&CK taxonomy. [4]

2. Phishing Attacks with the Help of LLMs

Spear Phishing: This is a specific type of phishing where the attackers gather personal information about a target (e.g. position, hobbies, recent enterprise activities) to make the inserted link and the message look authentic. The same can be done by LLMs computerizing this process to produce unique phishing emails which are harder to filter out by security measures.

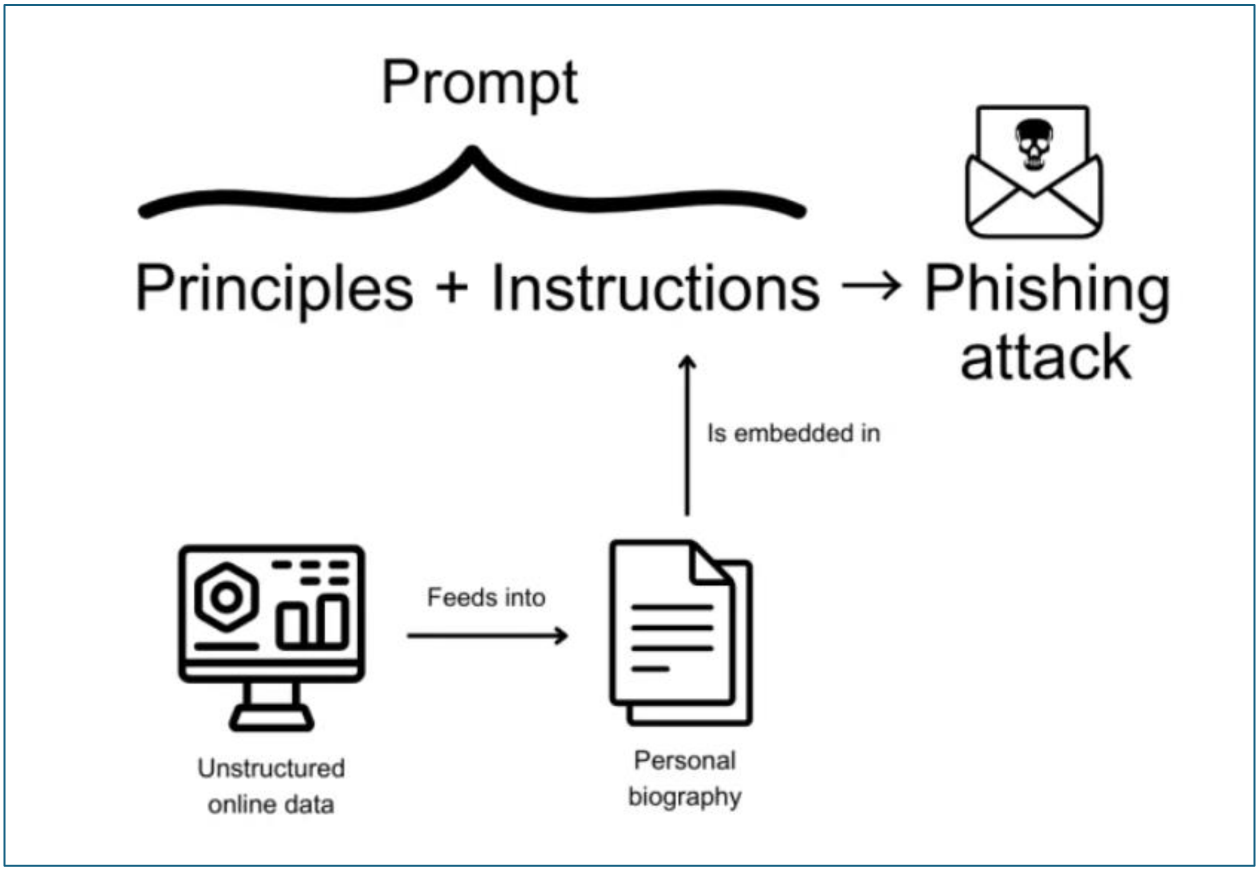

Figure 1 Phishing attack collaborate with LLM

Phishing Automation: LLMs helps cyber criminals to create, many personalist (spear phishing) phishing messages, with short time period. With minimum time, and low budget will motivate the attackers to hack more and more. [4]

3. Jailbreaking LLMs

- However, exploiting jailbreaking techniques may be used to attack LLMs to generate appalling or misleading content. Usually, Jailbreaking means modifying an LLM in such a way that the model is capable of producing outputs contrary to the design's security mechanisms. [4]

- Bypassing Safety Measures: Threat agents can take advantage of the loopholes that exist in an LLM's training and safety sections. Some LLMs are captured accidentally via loose aims unrelated to the model's main aim or due to additional objectives that may not agree with security measures at design time. Another way to Jailbreaking is the phenomenon of mismatched generalization – inputs that are outside the safety training of the model but still fall within the pretraining data. [4]

- Attack Techniques: Some attacks are specifically for the LLM to get an unlikely to refuse prefix. The last method, refusal suppression, when the LLM is asked something that will cause it to provide the desired harmful content when it has no choice but to answer the question. [4]

4. Leveraging Large Language Models (LLMs) for Cybersecurity: Defending Systems Against Phishing, Malware & Vulnerabilities

- Large language models are rapidly emerging as a useful application in cybersecurity in general and for countering malware, phishing, vulnerability among other threats. These models are learnt from big data and can help in analyzing code, recognize undesirable behaviors, and find information that supports the systems of security. This report focuses on the role of LLMs in terms of providing safeguard against malware, phishing attacks as well as for vulnerability analysis. [4]

1. How LLMs Can Be Used in Defending Malware

Malware is an collective term that consolidation all the existing ‘’malicious softwares’’ whose main intention is to cause damage or gain benefit from systems, networks and users. Some common example of malware are viruses, worms, trojans, ransomware and spyware. Malware can be carried in some ways including through e-mail attachments, direct downloads from specific dangerous websites, or through infected USBs. What LLMs do is provide several layers of protection against malware through patterns of code recognition and evaluation of certain behaviors. [4]

- Malware Patterns:

  When trained with large data sets containing various samples of malware LLMs can again classify new code as malicious due to similarities that it has with known malware. It also reaches to the identification of other new samples of malware comparing them to the previously identified one. It also can scrutinize malware’s behavior in order to understand how the malware spreads and steals information to help create more effective programs for Its detection. [4]

i)Code Analysis & Files

Thus, for files that can potentially malware, LLMss can also be used to detect malware. It can examine file hashes as MD5 or SHA-256 to compare those with the malware samples already in knowledge. [4]

```
    lang = self.autoDetermineLanguageFromString(inputString)
    if lang is None or lang not in ("en","fr"):
    if lang is None:
        raise Exception("Input language could not be determined")
        return None
parsedInput = self.parseInputToLanguageModel(inputString, inputLanguage, context)
if not parsedInput or not self.model:
    return None
context.append(parsedInput) # Add new conversation entry to context
return (self.model.generateLLMOutput(parsedInput), context)

def parseInputToLanguageModel(inputString, inputLanguage, context):
    if self.model is None or self.model.language != inputLanguage:
        # LLM is not initalised or has wrong language, load LLM
        self.model = self.loadAILanguageModelFromDatabase(inputLanguage)
        if self.model is None or not self.runModelSelfDiagnosticTests():
            raise Exception("AI language model load failed")
            return None
    self.model.setLLMContext(context) # Put past conversation context into
    llmInputParser = self.model.getInputParser()
    return llmInputParser.parseInput(inputString)

def generateLLMOutput(parsedInput):
    llmContext = self.model.getLLMContext()
    bytecodeResponse = self.model.convertInputToIntermediateResponse
    if bytecodeResponse is None:
                    failed to produce intermediate res
```

Figure 2 code

To check is there any suspicious activity and behaviors, we can use LLM can analyzing code by using code running in sandbox environment.

ii)Defending Against Obfuscation (The New Form of Malware)

Malware that hides itself from anti-virus programs is called as obfuscated malware.

malware with above techniques, is not easy to understand. [4]

- code encryption

  (hiding the malicious behind the encryption, when its decrypted malware start to run, but software security application and tools, can’t easily understand this is malicious)

- packing (use compressing technic to hide)
- polymorphism
- metamorphism

It is About recognizing obfuscation techniques where LLMs help in analyzing suspicious pattern in code and tag keywords or phrases and even odd binary patterns. Also, they can clarify the meaning of some string of code or any comments that were made in the malware code which is not necessarily the intention of the malware. [4]

2. Using of LLM’s in phishing attack defense

Phishing attack is another common attack in which the attackers imitates as someone they are not and make fake requests to get users to disclose certain information. This message which is a common example of a phishing email contains words such as ‘urgent’ or ‘threats’, poor grammar, links or attachments that are questionable. In regard to the defense from phishing attacks, LLMs can assist in different ways based on Email Content Analysis and URL List, as well as in developing Phishing Awareness Training Material. [4]

- Phishing Email Detection– An Overview

This is prescribed here by showing that LLMs can be trained to recognize phishing emails on the basis of text patterns and language used in known phishing emails. For instance, an LLM can scan for specific things such as use of a word like ‘urgent,’ sloppy language like the wrong use of spelling, or a link that is prohibited. This capability also permits the identification of email content in real-time, with filters or newsletters, and will send warnings to the users if such an e-mail can be identified as a “phishing attempt”. [4]

- Evaluating URL and links

LLMs can also judge the URLs and find out whether they are malicious by their similarity with the phishing mail pattern. In that sense, URL analysis that is otherwise traditionally solved by blacklists method, but using LLMs we can improve that process efficiently. [4]

- Creating Phishing Awareness Training

With IT teams and SOC teams’ awareness LLM can create practical (more similar to real) simulations about phishing mail and send their other employees to these phishing emails. And train them to protect that kind of incidents. [4]

3. LLMS in Vulnerability Analysis

Vulnerability analysis through analyzing large amounts of information and identifying known and new threats. This is traditional methods combination of LLM.

- Static code analysis:

  Little of being capable of identifying known patterns in source code such as buffer overflow, SQL injection, and cross site scripting (XSS). This pattern recognition allows LLMs to alert developers about parts of code that may present security risks, so a developer has a heads up on weaknesses. [4]

- Analyzing Logs and Anomalies:

  Automated LLMs are capable of processing massive amounts of logs from systems or applications to identify deviations and proactive signs of a security compromise. These logs can among other things indicate that an exploit has been initiated and while LLMs can analyze this data to offer relevant information to security personnel. [4]

- Patch Analysis and Documentation Review:

  vendors release patches that aim to correct some vulnerabilities, LLMs get the chance to study the code changes with the purpose of finding out more about the kind of vulnerability being patched. Furthermore, LLMs can review the documentation of the software to ascertain if they match the reality on the ground in order for organizations to be fully compliant and secure.

- Proactive Treat detection

  The LLMs can contribute to identifying a new and previously unknown threat or a zero-day vulnerability by searching for references to weakness in software in forums, newsgroups, and mailing lists. This makes it easier for organizations to attend to issues before they assume severe security threats to the organization. [4]

4. **Honeypots (Fake systems to attract attackers)**

   Honeypots are intentionally exposed computing systems that have limited security to attract attackers to them and provide the security team information on how the attackers work and what they use. LLMs can be useful in constructing honeypots and can help maintain them thus improving their security usefulness in the cybersecurity efforts. [4]

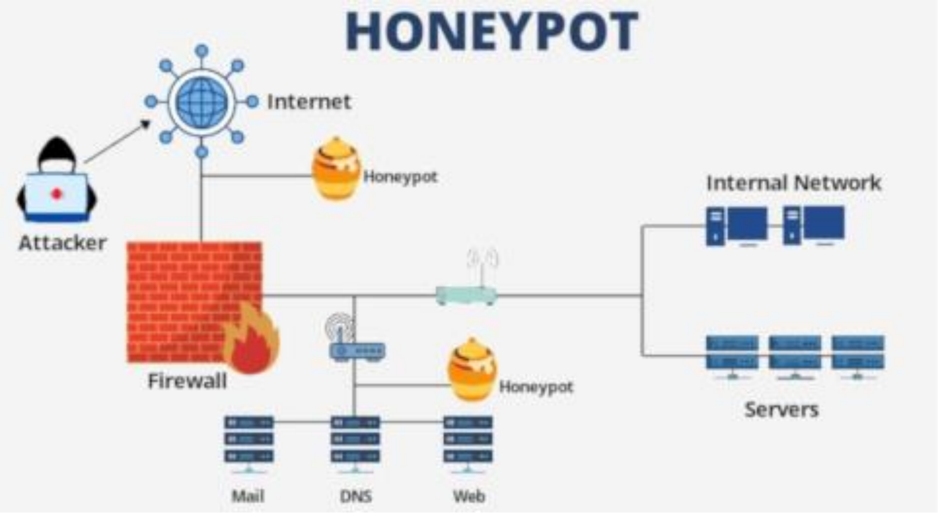

Figure 3 How the Honeypots works on

multiple benefits in malware detection effectiveness, phishing prevention, vulnerabilities disclosure, and honeypot setting. Through the use of LLMs, it is possible for organizations to improve the likelihood of predicting and preventing cyber threats hence instilling better security in organizations. With the LLM technology getting better with time, the idea of integrating it into cybersecurity practices can be taken top notch. Over the years, LLMs will continue to be more useful as the threat landscape increases and complicates even the simplest of defenses against cyber attacks. [4]

5.LLM Risk Mitigation

Mitigation strategies can help minimize issues. there are,

Mitigation Method	Details
1. Governance and Compliance	The first preventive approach is to ensure that there is the right AI governance policies installed to avoid loss-making deals in LLM. It is also important that the deployment of LLM should be in compliance to the current privacy, security and regulating framework of an organization. For instance, compliance of the models with GDPR (General Data Protection Regulation), Entity recognition should follow the right framework for data handling.
2. Censorship of Content	The following are reasons organisations should employ content moderation: In an attempt to minimize the chances of creating content that is negative, or potentially offensive content organizations should consider putting in place content moderation. These mechanisms can also be used for the filtering the messages in the input and output space for such data as well as for the blocking of dangerous prompts
3. Red teaming and Adversarial testing	To make strong and secured system, we have to pretending like a attacker, and check how the LLM handle tricky inputs and fake inputs. With this we can identify weakness of the system.
4. Data protection, Privacy and Encryption	organizations these days have had to incorporate measures of protecting the data through encryption as well as enforcing strict access control measures.
5. Continuous monitoring and update	There is need for constant evaluation of LLMs given their performance and security. Continual monitoring is effective in identifying anomalous behavior, data exfiltration and adversarial actions.

[5]

6.How LLM influences privacy for its users and for companies.

(Based on real world scenarios and real data.)

Many times, risks regarding security and privacy within LLMs are some results of users' improper handling of data., as pointed out by cybersecurity vendor Cyber haven in a report dated 21 March 2023, 3.1% of the employee had fed the confidential company data into ChatGPT. Employees exposed company secrets by inputting documents into ChatGPT systems at an average of 199 times at companies with 100,000 workers between February 26 and March 4, 2023. They also typed in client data 173 times, source code 159 times, on average. Specifically, during 14 March 2023, the company identified 5267 cases of getting an access to corporate data per 100 thousand employees with an aim to paste it to the ChatGPT. The information that the open AI posting provides is simple, in the FAQ section, the users are informed not to share any sensitive information in your conversations. These Information might be incorporated into its knowledge base and is employed to train the AI that is behind the bot as well as to enhance the tool. [6]

This has not stopped some LLM users from typing in sensitive data as part of their prompts for their 'bots'. For example, three distinct cases were identified of regarding employees of Samsung Electronics semiconductor business division entering corporate information into ChatGPT. Two different employees have used software code as part of their prompts to seek for defects and fix for them. A third employee turned an audio recording of a company meeting into a document and input it into ChatGPT to receive minutes. This statement and the identities of the individuals attending this meeting were off-the-record In April 2023, for the first time following the data leak, the company put a policy of an upload limit of a mere 1024 bytes per prompt into practice. Then in May 2023 the companies prohibited employees from using GAI (Generative AI). [6]

But, let me first mention that the given LLMs including ChatGPT, as a rule, are trained, and then the model responds to prompts. This implies that an LLM model does not directly introduce information from a given prompt into its framework. As such, although if such information is used in a query it does not become input to the LLM which makes it available to other users. However, an LLM's operator, for instance OpenAI, may see the prompts and since these are stored and may possibly be used in the future by the operator or its partners to advance on the LLM model. The dangers that queries stored on the web could be unlawfully accessed, published or shared are still there. The first risk is when a query might actually contain user identifiable information. Perhaps the LLM's operator could be acquired by an organization whose security and privacy policies are different from what they were at the time the users entered the data This is just to say that even though such risks are not as people imagine, the risks could well be very serious. [6]

7.AI Security and Privacy Training

Devote time to employees to know more about the problems they have with the LLM planned initiatives. Set up trust through reporting mechanism on the use of predictive or generative AI within the organizational process, systems, employees, and customers, and how the use of the AI technologies is controlled and mitigated for risks. Educate all users on ethics, responsibility, warranty, license and copyright. Security awareness training needs to be adjusted to include the threat that have to do with generative AI. In situations using image cloning or Voice Cloning, as well as, in view of the probable rise of spear phishing attacks Any generative AI solutions that are to be adopted should undergo training for both the DevOps and Information security departments of the company for deployment into security pipelines to give necessary assurance. [7]

8.Ethical principles for LLM

Ethics Issues and concerns in large Language Technology.

1. Privacy Concerns: Data Collection, user Consent and Data Protection.

- Data Collection:
  LLMs based on the large dataset for model training, and such data set might be sensitive or private data (why because of privacy infringement, spying and data misuse) Unauthorized access to data by other persons may compromise the user's privacy and confidentiality. [8]

- User Consent:

- Normal users have not fully understood how the LLM collect their data, use them and share. The reason for this is low level of data privacy transparency in LLM applications. So, this poses a serious threat to user rights. [8]

- Data Protection:
  If someone trained their LLMs for take sensitive data without user permission. This is not only affected to one person, but also this could be affected to corporate office environment. [8]

2. Misuse and Ethical Guidelines (Deep Fake Videos, Fake News, and Dangerous Content.)

Deepfakes:

- An LLM can be used to synthesize video, audio or textual content that can in turn be used to produce deepfake. can be designed in such a way as to provide any information and tell any story intended by its designer. It means that the civil society is at risk being deceived by deepfakes.
- For ruin some once life, such as personal reputation or political stability malicious actors use these techniques. They do these things for their poor entertainment or financially advantages. [8]

Disinformation:

- There is a possibility that LLMs will be wrong by some people with motives create fake news, posts on the social networking sites, fake reviews to not only decrease the reliability relating to the existence of information sources and spreading the division in society. [8]

- Harmful Content:
  The models themselves are learned on inappropriate or even objectionable material and can therefore produce outputs. that may contain explicit material that is related to and encourages violence, hate speech, and suicidal thoughts. Basically, getting in touch with such material has its negative potential. consequences on mental health, social interactions, and public security of especially frail groups. [8]

3. Regulation and Governance: Policy Implications and Industry Standards.

Regulation:

- LLM technology has developed very rapidly, but legal systems and ethical guidelines for the proper use of LLM have not been adopted, which creates problems in managing the risks associated with LLM. Therefore, new policies are needed, which include data protection and Algorithms should also contain rules that ensure accountability. It should also include the standards we need to create LLMs and the specific regulations that guide how they are used and applied. [8]

Control:

- The government, industries, universities, and the people of the country need to work together to create better rules and regulations for the discipline, responsibility, and ethical practices of LLM development. This will greatly help to solve the issues related to managing LLM and how it affects society and innovation. [8]

4. Bias and Fairness in LLMs (sex/ color/ethnicity/ class)

The reduction of bias and enhancement of fairness in LLMs raises important. So we have to use balance approach to train these large language models without unfair treatment like, gender based, race, ethnicity and economical classes. If we use fair balance approaches, these are the benefits we can get.

I) It will lead us to better choices and decisions.

II) Increases user friendliness and trust full environment for user who use these LLM. [8]

9. Evolution of LLM security

The evolution of the security of large language models (LLMs) can be traced through several main stages:

1. Before 2020

In the early stages, LLM

was mainly focused on developing it, without security aspect.

2. Several safety and security control measures as follows: (2020-2021)

- Introducing new methods to prevent adversarial experiments and improve significant stability.
- Presenting policies and general principles on the use of ethical artificial intelligence.
- New concerns regarding transparency, interpretation and user responsibility of LLM's output.

3. Regulatory Focus (2022)

- Increasing demands for regulations and standards to be implemented on agents and practices of AI systems such as LLM.
- The European Union has introduced the AI Act, which focuses on risk assessment and mitigation of high-risk industrial AI uses.

Figure 4 EU Artificial Intelligence Act

- https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence#:~:text=The%20use%20of%20artificial%20intelligence%20in%20the%20EU

4. Unification of Security Protocols (2023)

- New secure approaches in model training, including federated learning, which can train models without accessing raw sample data.
- Introducing features like content filtering and bias detection to improve the sanity of the output produced.

5. Evolving research and development (from 2024)

- There are efforts today to work on better architecture as well as adversarial training in LLMs.
- Attention-related trends are also present on LLM security frameworks for various applications such as chat or content creation.

changes in the discipline of LLM security highlight the emergence of the ethical factor and acceptance of safeguards to regulate AI. However, as LLMs improve, it is clear that safety may be crucial in suggesting the right direction for their use. [8]

10.Future Development in Large Language Model (LLM) Security

We can explore LLMs' emerging trends with below details.

1. **LLMs in Cybersecurity Laws, Policies, and Compliance**

- LLMs can assist in drafting security policies and compliance documents, helping organizations meet regulatory standards. However, the introduction of LLMs raises new legal challenges, particularly regarding data protection and privacy. [9]

- Privacy policies like those of ChatGPT offer important safeguards, but continuous improvements are needed to adapt to the growing complexity of LLM usage in security domains. [9]

2. **LLMs in Security Applications**

- Replacing Machine Learning Models: Discussed below are some of the capabilities in cybersecurity application that can be realized through LLMs making it possible to replace traditional machine learning methods like malware detection. Scholars believe that overall LLM based security techniques are expected to be measured against state-of-the-art approaches to create new frontiers. [9]

- Reducing Human Effort in Security: With the use of machine intelligence LLMs can perform similar tasks as the human contributions in offence as well as defense of the network. For instance, LLMs can be used in automating social engineering activities that mostly require human involvement. [9]

- Modifying and Adapting ML Attacks and Defenses: A large number of the risks of LLMs are inherited from classical ML risks. For example, LLMs can be harmed by adversarial attacks and jailbreak. Perhaps, tweaking old-school strategies like zeroknowledge proofs and federated learning could solve these problems. [9]

3. Other emerging trends in LLM security

- Continual Learning and Adaptation: Future LLMs will be adaptive and enhance himself, always updating his knowledge with new generated for effective inchanged climates. This capability makes them more capable of handling general and domain specific tasks with regards to cybersecurity with little delay. [9]

- Zero-Shot and Few-Shot Learning: Such LLMs with these learning capabilities are endowed with the ability to transfer their learning to new tasks with little or no further training required helpful in detecting best, unknown hitherto, cybersecurity threats where there is little availability of training data. [9]

- Integrating Multimodal Inputs: Finally, the proposed Multimodal LLMs can cover textual, image, and audio features, which expand the potential applications of cybersecurity. This feature is very useful for such purposes as filtering out fake multimedia content or searching for outliers in large data flows. [9]

Privacy-Preserving Techniques and Federated Learning

Since LLMs rely on collecting large volumes of data, important tools such as differential privacy and homomorphic encryption will be crucial going forward. This paper offers insights into why federated learning can be used to train LLMs in distributed data without going through the data, thus it is suitable for the finance and health sectors. [9]

Collaborative AI Systems

With LLMs and human working hand in hand to open different problems, security operations will be enhanced. Management LLMs will involve the user in real-time discussion of the tasks performed including handling of threats. [9]

4. Challenges and Future Directions

- LLM-Specific Security Attacks: New challenges are emerging, such as parameter extraction attacks, which exploit the vast scale of LLM parameters. These attacks will require researchers to evolve traditional attack methods to keep pace with LLM advancements.
- Developing Robust Defenses: Future work will focus on adapting traditional defenses to LLMs and exploring new privacy-enhancing technologies to secure them against evolving threats.

The future of LLM security lies in their continual evolution and adaptation. As these models become central to cybersecurity applications, addressing their vulnerabilities and privacy concerns will be crucial. Future research should focus on refining legal frameworks, adapting traditional defenses, and exploring new security techniques to ensure that LLMs remain secure as they develop further. [9]

11.Conclution

LLMs have brought change in many sectors. However, the fast development of LLMs has brought new security threats that cannot be ignored, despite their growth rate. This has been viewed from the increasing complexity of the attacks, for example, phishing and other forms of exploitation, of the existing system. It would be also important to note that recently attackers began to use LLMs to create realistic fake content, which will contribute to the creation of more effective and realistic fakes for the attainment of their goals.

However, this vulnerability does not deny LLMs the ability to be useful partners in combating cybercrime. However, the utilization of safe LLMs brings more complexity to the problem, its solution involves designing better LLM models, strengthening data protection and privacy measures and, perhaps, carrying out constant control over the work of artificial intelligence.

Managing LLM-related risk also entails promoting the right ethical standards for AI to embrace the 'openness, accountability.' Furthermore, organizations need to invest in proper AI security and privacy training that will help professionals to tackle LLM security issues. Furthermore, authorities and independent organizations perform an important function of regulating conditions for the creation of promising LLMs and protecting users' information and their rights.

12.References

[1] “What are large language models?”, [Online]. Available: https://www.ibm.com/topics/large-language-models. [Accessed: 10-Oct.-2024].

[2] “What is LLM? - Large Language Models Explained - AWS”, [Online]. Available: https://aws.amazon.com/what-is/large-language-model/. [Accessed: 10-Oct.-2024].

[3] “Just a moment...”, [Online]. Available: https://www.cloudflare.com/learning/ai/what-is-large-language-model/. [Accessed: 10-Oct.-2024].

[4] A. Iyengar and A. Kundu, "Large Language Models and Computer Security," 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), Atlanta, GA, USA, 2023, pp. 307-313, doi: 10.1109/TPS-ISA58951.2023.00045. keywords: {Training;Privacy;Computational modeling;Phishing;Malware;Computer crime;Intelligent systems;Large language models;security;privacy;cyber-attacks;malware;phishing attacks},

[5] R. Pasupuleti, R. Vadapalli and C. Mader, "Cyber Security Issues and Challenges Related to Generative AI and ChatGPT," 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS), Abu Dhabi, United Arab Emirates, 2023, pp. 1-5, doi: 10.1109/SNAMS60348.2023.10375472. keywords: {Ethics;Analytical models;Social networking (online);Government;Chatbots;Transformers;Artificial intelligence;Cybersecurity Issues;Challenges;Generative AI;ChatGPT;LLM Models;Risks},

[6] N. Kshetri, "Cybercrime and Privacy Threats of Large Language Models," in IT Professional, vol. 25, no. 3, pp. 9-13, May-June 2023, doi: 10.1109/MITP.2023.3275489.

keywords: {Privacy;Analytical models;Behavioral sciences;Security;Computer crime},

[7] “LLM Applications Cybersecurity and Governance Checklist v1.1 - English”, *OWASP Top 10 for LLM & Generative AI Security*. 23-Sept.-2024. [Online]. Available: https://genai.owasp.org/resource/llm-applications-cybersecurity-and-governance-checklist-english/. [Accessed: 10-Oct.-2024].

[8] “Just a moment…”, [Online]. Available: https://www.researchgate.net/publication/379091956_The_Evolution_and_Impact_of_Large_Language_Model_Systems_A_Comprehensive_Analysis?enrichId=rgreq-8c7d0612ad21cc63920c76821746d4e4-XXX&enrichSource=Y292ZXJQYWdlOzM3OTA5MTk1NjtBUzoxMTQzMTI4MTIzMDM2OTMzNEAxNzEwOTQyNTEyMjI1&el=1_x_2&_esc=publicationCoverPdf. [Accessed: 10-Oct.-2024].

[9] [Online]. Available: https://www.semanticscholar.org/paper/A-Survey-on-Large-Language-Model-(LLM)-Security-and-Yao-Duan/383c598625110e0a4c60da4db10a838ef822fbcf. [Accessed: 10-Oct.-2024].