HIGH SCHOOL EDITION
Journal of Student Research

# Navigating Data Risks: The Impact of Large Language Models on Privacy and Security

Lara Kawle[1] and Sheetal Dhavale[#]

[1]Woodbridge Academy Magnet School, USA
[#]Advisor

## ABSTRACT

Throughout the 21st century, the increasing presence of Artificial Intelligence (AI) in everyday applications underscores its crucial role in enhancing community comfort and convenience. And with the world's increased use in technology, there is no doubt that Artificial Intelligence will remain as a persistent center of focus on the stage of technological advancement. However, it is crucial to understand how the inner workings of AI components could be manipulated for malicious use by threat actors. Large Language Models (LLMs) are a type of Artificial Intelligence solution that process data, can recognize patterns and generate output text. Through large and extensive training, LLMs have the ability to produce natural language text as a response – but at what cost? The use of LLMs is prone to security risks and data breaches, as manipulated data that is fed to these systems could ultimately lead to incorrect, false, or unintended outputs. Overall, the vulnerabilities in the input data can compromise the integrity of information produced by the model and introduce unforeseen privacy attacks. This research paper summarizes the latest findings on the security risks that are associated with LLMs. In addition, the paper explores the inner workings of LLMs, the advantages and limitations of the use in this model, and recommendations to address the risks.

## Introduction

Large data sets are used by LLMs to power deep learning algorithms, which can produce, synthesize, and analyze text. LLMs mark a new generation of technology with its ability to utilize hundreds of billions of parameters to generate natural language that humans can understand. With this capability, LLM's variety of applications, such as text generation, content summarization, AI assistants, code generation, and sentiment analysis, proves it to be a worthy tool for companies and organizations to serve their clients/customers. LLMs consist of neural networks that are enhanced and fixed at each level during training. The training process involves adjusting billions of parameters—numerical values within the model that are fine-tuned to capture linguistic patterns and relationships. This extensive training enables LLMs to perform a wide array of language tasks with high proficiency. The scale of these models, both in terms of data and computational resources, contributes to their ability to understand and generate language in a way that closely mimics human communication, making LLMs a powerful automation tool in numerous applications across different domains (IBM, 2023).

A key difference between a traditional computer program and a LLM program is the ability of a LLM to produce output for an unpredicted query. Using the learnings from past inputs, training data, and probabilistic analysis, the large language model will respond to the provided input and give a relevant result. In other words, while traditional computer programs are only able to take an input and simply produce its corresponding outputs, LLMs can not only complete that function, but also learn how to manipulate other inputs to produce different outputs.

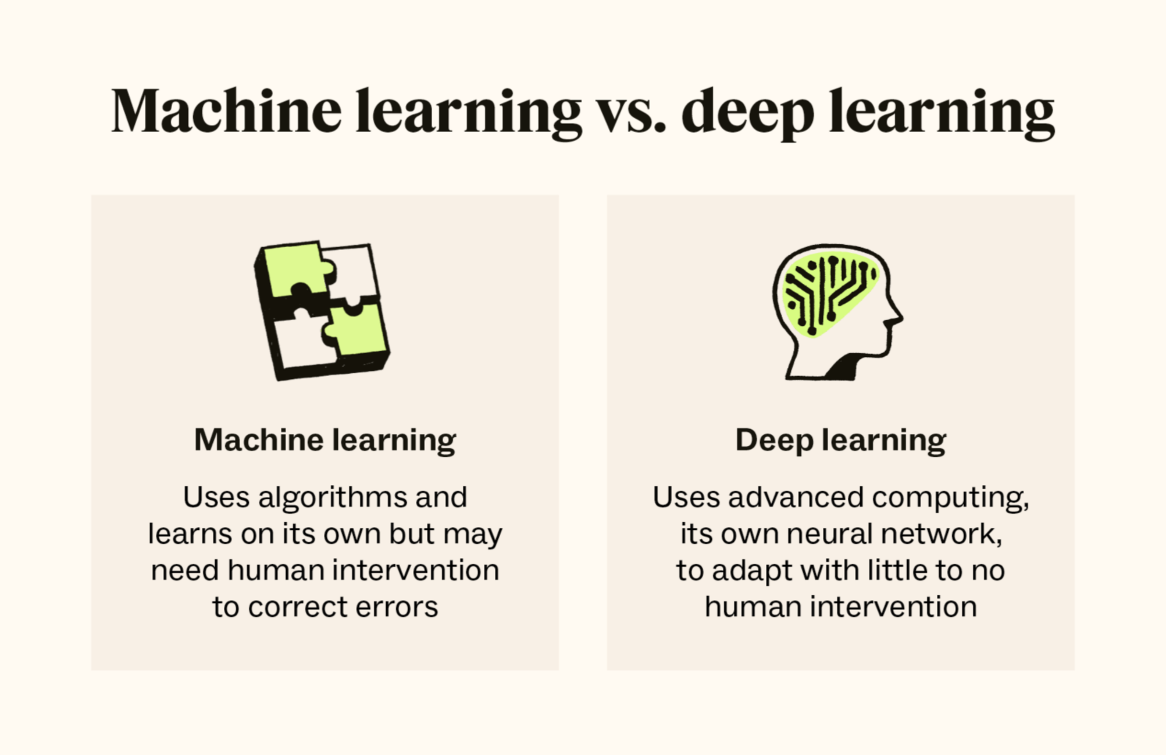
In addition, LLMs are a type of machine learning technique that parses the input user text by recognizing complex patterns in it and makes associations to natural language. This type of machine learning is also known as "deep learning", or a method in artificial intelligence that utilizes multilayered neural networks to replicate complex-level decision making.

## Deep Learning

The process of deep learning is based on the fundamental workings of a neural network. It is a form of machine learning that requires the computer to learn from experience. In simpler terms, it helps to train the computer to think like a human. Deep learning is modeled like the human brain – the same way that the brain has numerous interconnected neurons firing back at each other, deep learning neural networks contain artificial neurons that work together to fulfill a specific purpose for the computer. These artificial neurons are also known as “nodes”. Nodes can be utilized to calculate complex mathematical procedures.

A deep learning network contains the following components: an input layer, a hidden layer, and an output later. The input layer of a neural network is made up of nodes that data can be inputted into. The hidden layer is the layer that helps to pass the data farther into the neural network. This layer can also help to analyze problems from different perspectives to get the most appropriate output. Lastly, the output layer is made up of nodes that output data. It produces the final results of the network and is typically composed of a fully connected layer. If there are two nodes in the output layer, the model will output either “yes” or “no” outputs. On the other hand, if there are a greater number of nodes present, there will be a larger variety of different outputs (Amazon, 2023).

A few different types of deep learning neural networks include Convolutional Neural Networks, Recurrent Neural Networks, and Generative Adversarial Networks. Convolutional Neural Networks are neural networks that have the ability to detect patterns and specific details in pictures and videos, as well as other image classification applications. Recurrent Neural Networks are neural networks that convert sequential data into specific outputs. This type of network is typically found in natural language and speech recognition applications. Lastly, Generative Adversarial Networks are neural networks that take original data and produce a new output (CerboAI, 2024).



Uses algorithms and learns on its own but may need human intervention to correct errors

Uses advanced computing, its own neural network, to adapt with little to no human intervention

**Figure 1**. This diagram depicts the difference between machine learning and deep learning. This figure is from (Turing).

## Anatomy of Neural Networks

A neural network, as described before, is a model that is designed to mimic the functions of the human brain. In simple terms, similarly to how human brains contain neurons, neural networks consist of interconnected nodes that work to solve complex problems.

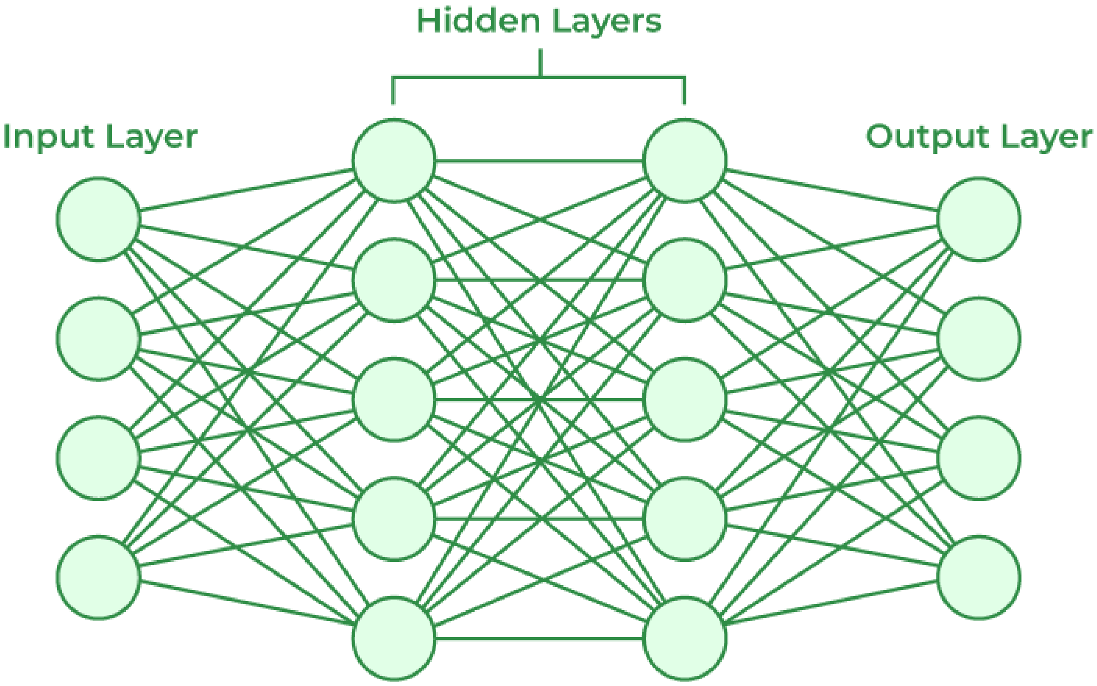


**Figure 2**. This diagram models a neural network. This figure is from (GeeksforGeeks, 2024).

In a neural network, data is first received in the input region. This data will move through a series of nodes through the hidden layers and is analyzed all throughout. Each node contains its own sphere of knowledge, where it attributes details and characteristics to the data based on its own background. Lastly, the output system will then produce the output. Artificial neural networks are known to be adaptive, in which it is able to modify itself as it learns more from its environment through several runs of itself.

In terms of LLMs, they use a special type of neural network known as the “Transformer Model”. Transformer models use a mathematical technique called attention, or self-attention, to detect ways that elements in a sequence may relate to each other. This feature helps LLMs to interpret human language. Even if a person were to use slang language or incomplete sentences to convey a message to a LLM, it is still able to understand enough to associate words and ideas to meaning (Cloudflare). With the help of word embeddings, transformers can process text as numerical representations to understand the context behind words or phrases.

## Benefits of Large Language Models

The addition of LLMs in software provides numerous benefits, especially with its ability to be versatile and yet still produce relatively accurate results. A few benefits that come with applying LLMs include its efficiency in automating various tasks, its scalability, its customization, its accessibility, and its trait of being able to continuously learn. With its feature for efficiency, LLMs can accelerate completing what may be known as tedious tasks, such as generating code, helping with complex decision making, and more. Not only that, its scalability allows for it to perform complex inquiries, which can be incredibly helpful for the public with extensive and long documents. These excellent features are not limited to one group of people; its large accessibility to various crowds allows for a wide range of users that can participate effectively in software acquisition.

In addition, because LLMs are a type of machine learning model, LLMs have the ability to improve over time as they are exposed to more data. They learn from their past runs, determining ways to fine-tune their responses to become more proficient in addressing software acquisition. LLMs are transforming software with their speed, ability to handle complex tasks, and ongoing improvements. Their efficiency and wide accessibility benefit diverse users. As they continue to evolve, they will play an increasingly important role in advancing technology (Robert and Schmidt, 2024).

## Limitations of Large Language Models

While LLMs offer impressive advantages, it's also crucial to address their limitations. Despite their strengths, these models face challenges such as handling incorrect results, overstepping in ethical use, and managing its dependence on context. Not only that, LLMs can be prone to produce incorrect results and is a highly resource intensive process. Understanding these limitations helps provide a balanced perspective on their role in software development.

When LLMs produce incorrect results, or also commonly known as hallucinations, they can have varied effects on the software. Whether it be a small bug or a shutdown in the program, its vulnerability to software defects is an example of the limits that LLMs have. As LLMs become more fine-tuned through its learning process, the model would have to be consistently tested and undergo multiple validation tests to ensure accuracy in its outputs. In addition, as described before, LLMs are able to understand sentiments in human language based on the data that is fed to the model. However, when this data is manipulated to contain bias or strong opinions that skew from the original data, the model may produce unwanted outcomes. To ensure fairness and appropriateness, human oversight should be used first before the input provided by a Large Language Model.

Most parameters of a LLM are matrix weights used in the training and inference. Thus, even though the usage of matrix operations such as graphics processing units (GPUs), tensor processing units (TPUs), and other specialized AI chips have made training these gigantic models possible, it can be problematic in terms of its resource consumption. Additionally, because the LLM needs to be trained using large datasets, there will be high monetary costs as a result due to the need for powerful computing platforms.

Furthermore, the use of LLMs also introduces significant concerns related to data security and privacy. Especially in the 21st century, leaking of confidential data can have dangerous consequences. When interacting with LLMs, there's the potential risk of disclosing sensitive or private information, which could be compromised if not managed carefully (Robert and Schmidt, 2024).

## Difference Between Natural Language Processing and Large Language Models

Natural Language Processing, by definition, is a branch of artificial intelligence that uses machine learning to understand and produce human language. Outwardly, Natural Language Processing may seem like LLMs, as Natural Language Processing also provides human language processing capabilities. However, the underlying technology is different between the two types of language processing.

The main difference between Natural language processing and LLMs is that Natural Language Processing is achieved by following specific rules like nuances of grammar, syntax, and context of the language used in the input text. Many of these functions and resources are manually set to complete their tasks. On the other hand, LLMs do their own learning, as it is a type of machine learning, to understand patterns and utilize that data to perform complex decision making. Thus, it makes LLMs a more advanced tool to predict outcomes. Overall, Natural Language Processing focuses more on producing accurate outputs, while LLMs are more likely to provide biased or incorrect results.

In terms of other differences, Natural Language Processing and LLMs differ in their application cases. For example, Natural Language processing is typically used in settings where information extraction, sentiment analysis, or machine translation is required. On the other hand, LLMs may be used in settings like content creation, chatbots, and virtual assistants. Understanding these distinctions helps in choosing the right technology for the desired outcome and leveraging each tool's strengths effectively (Timbó, 2024).

## Navigating Data Risks

As the adoption of LLMs grows, it's important to acknowledge and address the associated risks. While these models offer powerful capabilities, such as enhanced automation and advanced decision-making, they also come with potential pitfalls that can impact both users and organizations. Understanding these risks and solutions—ranging from data privacy concerns to the challenges of managing biased outputs— can help to mitigate the issue. The following paragraphs will describe a few key vulnerabilities with the usage of LLMs as well as solutions to each one of them.

As described earlier in previous sections, it is possible for someone to attack LLMs by manipulating inputs for legitimate prompts, which may cause the model to produce unwanted bias or information. This type of cyberattack is also known as a prompt injection attack. This type of attack can cause models to ignore any guidelines it was prompted to follow and may even pose greater security risks when outputting sensitive information. Prompt injection vulnerabilities pose a significant challenge for AI security researchers due to the lack of a definitive solution. These vulnerabilities exploit a key aspect of generative AI systems—their ability to process and respond to natural-language instructions from users. Detecting malicious inputs reliably is complex and restricting user inputs could fundamentally alter the functionality of LLMs. To mitigate this issue, a few solutions include restricting LLMs to a minimum level of access, ensuring a human review process with the model, and segregating vulnerable data with untrusted content (Forrest and Kosinski, 2024).

In addition, since LLMs use the process of deep learning, these models need extensive training and fine tuning to get reliable results. While the use of extensive training may help the model to produce more accurate results, it can also become vulnerable to training data poisoning. Training data poisoning is described as a malicious contamination of data that causes the model to induce biases and errors. This type of cyberattack focuses on affecting the training phase of the model. There are two types of data poisoning attacks: targeted attacks and non targeted attacks. Targeted attacks work to prevent a model's normal functionality without degrading its overall performance. On the other hand, non targeted attacks focus on entirely degrading the model's overall performance. As a solution, LLMs can use training data from verified sources only, introduce a manual review process in the training-feedback loop of the model, and incorporate red team testing, or a type of testing that simulates real-world attacks for organizations to improve their security on their model (Nightfall AI).

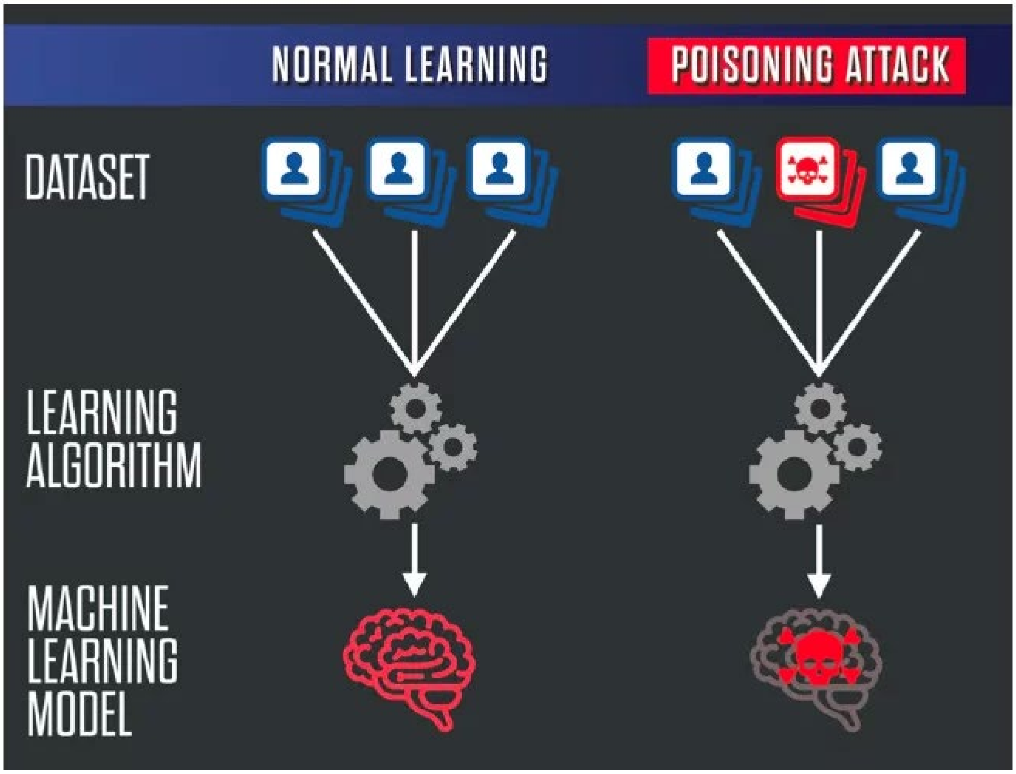


**Figure 3**. This diagram shows the difference between normal learning and a poisoning attack. This figure is from (Information Matters, 2020).

Lastly, one of the biggest issues LLMs face today is its ability to retain sensitive information. This “sensitive information” can include personal identifiable information (PII), financial details, health resources, confidential business data, and legal documents. If private details like these are exposed, companies could encounter serious legal issues, particularly under regulations such as the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR) in Europe. A few solutions to this issue include that data from external sources should be validated, appropriate access control reviews should be enforced, and strict access controls to external or third-party data should be implemented (Menon and Sheth, 2024).

## Recommendations

To enhance the AI skills of employees involved in designing and implementing LLM solutions, it is essential to enforce comprehensive training programs. Engaging Subject Matter Experts (SMEs) throughout the design and development life cycle is also crucial, as they provide critical insights and help identify gaps in the implementations. Adequate capital and funding are necessary to build and maintain the robust infrastructure required to support LLM solutions, ensuring both business availability and system resilience. This includes budgeting for red team and quality assurance testing to thoroughly evaluate the solutions. Given the rapid advancements in AI technology and the potential for new vulnerabilities to emerge, organizations must possess the technical agility to swiftly adopt and integrate remedial solutions to mitigate these risks. Furthermore, it is important for companies to review and integrate guidance from trusted organizations such as NIST and OWASP to adhere to best practices and recommendations when designing systems that incorporate LLM models. This comprehensive approach will help ensure that the solutions are both effective and secure.

## Conclusion

This paper helped to demonstrate the several vulnerabilities that come with LLMs and what can be done to prevent them from causing detrimental consequences in the future. As LLMs continue to advance and integrate into various applications, navigating the associated data risks remains a critical challenge. The potential for privacy breaches and security vulnerabilities underscores the need for robust safeguards and thoughtful regulation. Organizations must adopt stringent measures to protect sensitive information and ensure compliance with data protection laws. By balancing innovation with vigilance, the benefits of LLMs can be harvested while mitigating their risks, ultimately fostering a secure and trustworthy digital environment.

## Acknowledgments

I would like to thank my advisor for supporting me throughout this research journey.

## References

Amazon. (2023). *What is Deep Learning? Deep Learning Explained - AWS*. Amazon Web Services, Inc. https://aws.amazon.com/what-is/deep-learning/

CerboAI. (2024, May 2). *CerboAI's Guide: Understanding CNN/RNN/GAN/Transformer and Other Architectures*. Medium; Medium. https://medium.com/@CerboAI/cerboais-guide-understanding-cnn-rnn-gan-transformer-and-other-architectures-2ded10988eee

*Data Poisoning: The Essential Guide | Nightfall AI Security 101*. (n.d.). Www.nightfall.ai. https://www.nightfall.ai/ai-security-101/data-poisoning

Forrest, A., & Kosinski, M. (2024, April 11). *What Is a Prompt Injection Attack? | IBM*. Www.ibm.com. https://www.ibm.com/topics/prompt-injection

Grieve, P. (2018). *A simple way to understand machine learning vs deep learning - Zendesk*. Zendesk. https://www.zendesk.com/blog/machine-learning-and-deep-learning/

harkiran78. (2020, June 24). *Artificial Neural Networks and its Applications*. GeeksforGeeks. https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/

IBM. (2023). *What Are Large Language models? | IBM*. Www.ibm.com. https://www.ibm.com/topics/large-language-models

Information Matters. (2020, July 31). *IBM TechXchange: IBM Automation & AI Day*. Information Matters - AI in the UK. https://informationmatters.net/data-poisoning-ai/

Robert, J., & Schmidt, D. (2024, January 22). *10 Benefits and 10 Challenges of Applying Large Language Models to DoD Software Acquisition*. Insights.sei.cmu.edu. https://insights.sei.cmu.edu/blog/10-benefits-and-10-challenges-of-applying-large-language-models-to-dod-software-acquisition/

Sheth, J., & Menon, R. (2024, May 30). *Perils of AI: LLM applications and sensitive information handling | Globant Blog*. Globant Blog. https://stayrelevant.globant.com/en/technology/cybersecurity/sensitive-information-disclosure-in-llm-applications/

Timbó, R. (2024, July 2). *NLP vs. LLM*. Revelo.com. https://www.revelo.com/blog/nlp-vs-llm#:~:text=Natural%20language%20processing%20(NLP)%20refers

*What is a large language model (LLM)?* (n.d.). Cloudflare. https://www.cloudflare.com/learning/ai/what-is-large-language-model/