Cyber Security Experts Association of Nigeria (CSEAN)
Society for Multidisciplinary & Advanced Research Techniques (SMART)
West Midlands Open University
SMART Scientific Projects & Research Consortium (SMART SPaRC)
Sekinah-Hope Foundation for Female STEM Education
ICT University Foundations USA
Academic Innovations City University Foundations

Proceedings of the Cyber Secure Nigeria Conference – 2024

AI's Dark Secret: An Understudy into ChatGPT's Privacy, Ethical, and Security Vulnerabilities.

[1]Adefunke Bolatito, [2]Ruth Sam-Mbok & [3]Nasir Baba Ahmed
University of Ibadan, Ibadan, Oyo State, Nigeria[1]
University of Jos, Plateau State, Nigeria[2] Octagon Cybersecurity, Abuja, Nigeria[3]
E-mails: adefunkebolatito5@gmail.com[1] Samruthnen@gmail.com[2] nasir4u07@yahoo.com[3]
Phone Nos: +2348125815350[1] +234903443822[2] +2349076185107[3]

ABSTRACT

This research paper explores the capabilities and implications of ChatGPT, a cutting-edge AI chatbot that leverages topic modelling and reinforcement learning to produce human-like responses. While ChatGPT has vast potential across various sectors, including customer service, education, mental health treatment, and content creation, we must examine its security, privacy, and ethical ramifications. By tracing the evolution from GPT-1 to GPT-4 and examining the model's characteristics, limitations, and applications, this case study sheds light on the potential risks associated with integrating ChatGPT into our daily lives. With a focus on security, privacy, and ethical concerns, we identify the challenges these concerns pose for widespread adoption. Finally, we pinpoint unresolved issues in these areas, emphasising the need for collaborative efforts to develop secure and ethically responsible large language models.

Keywords: ChatGPT, Large Language Models, Privacy, Security, Ethics.

Proceedings Citation Format
Adefunke, B., Sam-Mbok, R. & Nasir, B. A. (2024): Healthcare Cyber Infrastructure Risk Assessment: A Use-Case Approach. Proceedings of the Cyber Secure Nigeria Conference held at The Ballroom Center, Central Business District, Federal Capital Territory, Abuja, Nigeria - 25th - 26th September, 2024. Pp 121-130. https://cybersecurenigeria.org/conference-proceedings/volume-1-2024/ dx.doi.org/10.22624/AIMS/CSEAN-SMART2024P11

1. INTRODUCTION

In December 2022, OpenAI unveiled ChatGPT, a transformative chat platform that has left an indelible mark on the world. This AI tool prides itself on exceptional in-context learning and natural response generation capabilities, distinguishing it from its predecessors like GPT (Rashford et al., 2018) and BERT (Devin et al., 2018).

Unlike these earlier models, which excelled in specific tasks with carefully crafted inputs, ChatGPT shines with its remarkable versatility and adaptability. ChatGPT's capabilities are vast and varied, including free-form conversations, diverse tasks, and responses in multiple styles. It can generate codes, summarise texts, and explain complex concepts. ChatGPT delivers lengthy, human-like responses while acknowledging its knowledge boundaries and declining inappropriate requests. With a staggering 100 million users and 1.6 billion visits as of June 2023, ChatGPT has emerged as the most prominent AI application, captivating global attention (Xiaodong et al., 2024). The advent of ChatGPT has brought about a transformative shift in human-AI interaction, but it also raises critical ethical concerns and risks. The powerful text reasoning and generation capabilities of ChatGPT have led to a plethora of issues, including academic dishonesty, copyright infringement, and privacy vulnerabilities.

Students have exploited ChatGPT to complete their homework, blurring the lines between original work and plagiarism. The unattributed use of ChatGPT-generated content has also raised questions about ownership and accountability for accuracy. Furthermore, the potential for data leakage or malicious attacks, such as jailbreaking attacks, compromises personal information and other AI models. While numerous research efforts have been undertaken to address these issues, a cohesive summary and comparative analysis of proposed solutions are lacking.

This paper highlights the critical security, privacy, and ethical concerns surrounding ChatGPT, the latest and most renowned AI technique. The primary objective is to raise awareness about these issues and explore potential solutions. The key contributions of this case study are:

- A comprehensive overview of the evolution from GPT-1 to GPT-4, including a detailed comparison of model size, data size, and performance. We discuss the features, limitations, and advanced applications of these Large Language Models (LLMs), highlighting ChatGPT's promising applications.
- An examination of the security threats posed by ChatGPT, including its potential use in generating attack codes and phishing websites, thereby enhancing adversarial attack capabilities. We also discuss the unintended consequences of inaccurate information, human misuse, and potential social security risks arising from excessive reliance on ChatGPT.
- An analysis of OpenAI's privacy policy and current privacy laws regarding personal data protection, emphasising ChatGPT's privacy violations.

2. BACKGROUND

In this section, we briefly discuss the technical path from GPT-1 to GPT-4 and its features and limitations.

2.1 From GPT-1 to GPT-4

In 2018, OpenAI revolutionised the field of natural language processing (NLP) by introducing the generative pre-trained transformer (GPT), a powerful large language model that demonstrated exceptional performance across a wide range of complex language tasks. GPT's capabilities positioned it as a strong competitor to other prominent models like BERT, which was proposed by Google in the same year (Devin et al., 2018).

Before GPT and BERT emerged, NLP had seen significant advancements with the development of effective algorithms and applications in areas such as machine translation (Ranathunga et al., 2023), voice recognition (Çayır & Navruz, 2021) and summary generation (Liu & Lapata, 2019). However, these applications relied heavily on extensive annotated data, resulting in time-consuming and costly model training as the models grew. Furthermore, their generalisation to other tasks became a challenge despite their proficiency, limiting their ability to perform diverse tasks like humans. To address these challenges, OpenAI developed GPT, a methodology that could be trained without labelled data and possessed superior generalisation capabilities across multiple tasks.

One of the most significant advantages of GPT was its ability to train without relying on vast amounts of annotated data. This is achieved through a two-step process that revolutionises how we approach natural language processing. The first step, unsupervised pre-training, leveraged rich text materials to train the model, using 12 transformer blocks as decoders. Each block was designed to predict the next word in a sentence based on the preceding words, requiring only raw text materials. This approach eliminated the need for extensive labelled datasets, making it a game-changer in the field. The second step, supervised fine-tuning, focused on a specific task, such as sentiment classification. This phase requires a significantly smaller annotated dataset, making it a more efficient and cost-effective approach. Users could easily adapt GPT to various tasks by modifying the input format for fine-tuning, further increasing its versatility.

GPT's innovative approach, which employed unsupervised pre-training and supervised fine-tuning, significantly reduced the need for vast labelled datasets while maintaining the model's ability to generalise across various tasks at a reasonable fine-tuning cost. This approach presented a remarkable solution to the challenges posed by traditional NLP models.

However, GPT-1 still required annotations during the fine-tuning phase, which led to the development of GPT-2 in 2019 (Rashford et al., 2019). GPT-2 revolutionised the training process by shifting from supervised learning to unsupervised learning, enabling the model to acquire knowledge from a vast dataset with rich materials. This pre-trained knowledge was a foundation for tackling complex tasks, making supervised learning a mere application of this preexisting knowledge. To achieve this, GPT-2's architecture remained unchanged from GPT-1, with critical modifications including increased layers and dataset size. These enhancements equipped the model with the knowledge to handle a wide range of problems proficiently.

The progression from GPT-1 to GPT-2 demonstrated the potential for improving large language models' (LLMs) generalisation capabilities by increasing parameter size and training dataset. Building on this concept, GPT-3 was introduced in 2020 as a significantly more powerful LLM (Brown et al., 2020). With many parameters and an extensive training dataset, GPT-3 achieved state-of-the-art performance across numerous NLP tasks. In addition to its exceptional gesture, GPT-3 introduced a novel training paradigm called in-context learning, departing from the conventional approach of predicting outputs solely based on queries. Instead, the model is trained to predict outputs by considering both queries and their corresponding examples, enabling it to acquire comprehensive knowledge from texts. This approach empowers GPT-3 with remarkable generation capabilities, delivering exceptional performance in diverse NLP tasks, often comparable to human-level performance.

Despite GPT-3's impressive performance, it still falls short of true intelligence. Surprisingly, smaller models like T5 (Roberts et al., 2020) have outperformed GPT-3 in certain tasks, suggesting that its vast training datasets and complex parameters do not guarantee superior performance(Vallecillo Rodríguez et al., 2024). One hypothesis is that GPT-3 struggles with understanding and responding accurately to user queries despite its rich foundational knowledge. To address this, OpenAI refined GPT-3's performance through advanced training methodologies, including code-based training (Chen et al., 2021) and instruction tuning (Ouyang et al., 2022), to enhance its reasoning capabilities and responsiveness to instructions. The updated GPT-3 demonstrates improved, reasonable responses, complex reasoning, and greater generalisation power, even in unseen tasks. It is speculated that GPT-3's remarkable capabilities were latent (Fu et al., 2022), requiring specific training techniques to unlock its full potential. OpenAI further improved GPT-3 using reinforcement learning from human feedback (RLHF) to align machine-generated answers with human common knowledge, leading to the development of ChatGPT, which quickly gained widespread attention and interest globally.

Just four months after the launch of ChatGPT, OpenAI announced the release of GPT-4 (OpenAI, 2023), boasting enhanced generation capabilities and significant advancements. GPT-4 empowers users to engage in creative and collaborative endeavours like personalised writing and song composition, learning specific writing styles to generate tailored works. It also exhibits improved reasoning abilities, delivering more accurate and nuanced outcomes for complex questions. GPT-4 surpasses ChatGPT's performance in text benchmarks and simulates exams with higher scores. Additionally, it supports visual inputs, allowing users to input text and visual content, and can mimic or emulate the style of inputs in its responses. GPT-4 demonstrates superiority in handling multimodal inputs, showcasing improved reliability and alignment. OpenAI has implemented RLHF techniques to enhance safety, making GPT-4 less likely to respond to illegal requests and more inclined to generate factual and appropriate replies. These advancements have generated significant attention and interest worldwide, solidifying GPT-4 as a groundbreaking model in the field of AI.

2.1.1 Features and Limitations

ChatGPT has made significant strides in overcoming the limitations of its predecessor, GPT-3. One notable improvement is its ability to humbly admit when it is unsure or does not know the answer to a question, especially regarding events beyond its knowledge scope. For instance, when asked about future events, it wisely responds that it cannot predict the future. Additionally, ChatGPT generates longer, more neutral responses that align better with human common sense, thanks to its RLHF training, which prioritises responses based on real human preferences. It has also become more discerning, declining to respond to inappropriate or unsuitable queries.

However, despite these advancements, ChatGPT still has some limitations. Sometimes, it may provide incorrect or unrelated answers, which can be frustrating. Its responses might contain inaccurate facts or biased perspectives rooted in specific regional domains. Another challenge is the cost of retraining the model, which limits its knowledge to datasets before 2021, leaving it without contemporary training. Lastly, ChatGPT still lacks emotional expression, responding in a dispassionate voice that fails to convey emotions.

Addressing these aspects would bring ChatGPT closer to achieving more human-like conversational abilities, allowing it to better understand and connect with users on a deeper level.

2.1.2 Application

ChatGPT has emerged as a versatile and powerful tool, revolutionising how we interact with technology. Its exceptional fluency and rapid response generation have captivated millions of users, making it an ideal chat robot or artificial assistant, similar to Siri or Google Assistant. Whether seeking to understand complex concepts, engage in theoretical discussions, or receive personalised advice such as "a meal plan for a 25-year-old girl working out her glutes," ChatGPT delivers natural, accurate, and helpful responses. This has positioned it as a viable alternative to traditional search engines like Bing or Google. Moreover, ChatGPT's capabilities extend beyond conversational applications. Researchers have leveraged their generation ability to aid in various fields, such as climate research (Biswas, 2021) and public health (Biswas, 2023). Studies have shown that ChatGPT can improve people's understanding of climate change, increase the accuracy of climate predictions, and even assist in screening patient health conditions, providing early detection for diseases with notable precision.

Another significant application of ChatGPT lies in its prowess as a code generator. By describing a task, such as image classification, ChatGPT can generate complete code with clear explanations. This demonstrates ChatGPT's proficiency in comprehending and articulating artificial machine languages. Additionally, ChatGPT serves as a valuable code debugger, offering comprehensive correction plans for problematic code, making it an indispensable tool for programmers.

2.2 ChatGPT Vs CyberSecurity

The security risks associated with ChatGPT are escalating at an alarming rate. Its advanced intelligence amplifies existing security threats and introduces novel risks, posing a dual challenge to users and the public. As ChatGPT's capabilities continue to evolve, adversaries are becoming increasingly sophisticated in exploiting vulnerabilities, making it essential to implement robust protective measures and maintain heightened vigilance to mitigate these emerging threats. The exponential growth of security risks necessitates a proactive and adaptive approach to ensure the safe and responsible use of ChatGPT.

2.2.1 Malware Package Execution

While ChatGPT can reject inappropriate requests, such as generating malware code, sophisticated hackers can still find ways to manipulate the system for malicious purposes. This includes using ChatGPT to generate malware codes or provide guidance on identifying vulnerabilities. Although ChatGPT's capabilities are still limited, it can potentially lower the barrier for individuals with limited technical skills to engage in hacking activities, increasing the efficiency of generating attack variants. However, it is important to note that Large Language Models (LLMs) like ChatGPT are not yet advanced enough to surpass human hackers entirely. They require adjustments to function properly and can make mistakes when handling complex projects, limiting their effectiveness. Nevertheless, the potential risks associated with ChatGPT's misuse warrant continued vigilance and research into mitigation strategies.

2.2.2 AI Package Hallucinations

Hackers have discovered a new way to spread malicious packages by exploiting the code generation capabilities of ChatGPT, a technique known as AI package hallucination. This cunning method begins with a simple query to ChatGPT, seeking package recommendations to solve a coding problem. ChatGPT responds with a list of suggested packages, including some that do not exist. The hacker then publishes a malicious package with the same name as one of the non-existent packages. When an unsuspecting user asks ChatGPT the same question, the AI recommends the malicious package, which the user installs and executes.

This insidious technique takes advantage of ChatGPT's imperfections to create customised attacks that can evade traditional detection methods. The malicious packages can be designed to use obfuscation techniques and function as trojans, making them extremely difficult to detect. As a result, this emerging threat highlights the urgent need for advanced security measures to combat AI-facilitated attacks and protect users from these sophisticated hacking techniques.

2.2.3 Prompt Injection and Evasion

Prompt injections (Lu et al., 2023) and evasion attacks pose significant threats to Large Language Models (LLMs) like ChatGPT. Prompt injections involve crafting clever prompts to bypass filters or manipulate the model into ignoring previous instructions or performing unintended actions. This can lead to sensitive information disclosure, restricted response revelation, or misleading the model into taking unwanted actions.

Similar to prompt injections, evasion attacks (Kaoloudi & Li, 2020) aim to deceive the model by introducing carefully crafted input data, causing incorrect or unexpected predictions without requiring access to the model's internal workings. In the context of language models, evasion attacks involve constructing input texts that exploit the model's weaknesses, producing biased or unintended responses. ChatGPT is an LLM, so it is vulnerable to evasion attacks, which is a concerning risk, even though no practical attacks have been applied to it yet. These attacks can have severe consequences, highlighting the need for robust security measures to protect against such manipulations.

2.2.4 Training Data Poisoning

Training data poisoning attacks (Huang et al., 2020) pose a significant threat to the field of AI, particularly in the context of Large Language Models (LLMs). By contaminating the training data, attackers can manipulate LLMs to produce erroneous outputs and make unreliable decisions. This vulnerability is especially concerning in LLMs, as they can be fine-tuned to behave maliciously during inference. Attackers can exploit vulnerabilities by manipulating the training data or fine-tuning procedures, introducing backdoors or vulnerabilities that compromise the security and effectiveness of the models.

Despite being black-box models, LLMs are still susceptible to attacks, where an attacker can infiltrate the training data pipeline and inject malicious data. The lack of robust data sanitisation methodologies, training data integrity checks, and audits makes LLMs vulnerable to malicious manipulations. As a result, malicious insiders can compromise the fine-tuning process, introducing backdoors or vulnerabilities into the LLM, which can have severe consequences for the model's security and effectiveness.

3. ChatGPT Vs PRIVACY

As we explore the capabilities of ChatGPT, we cannot help but wonder: What secrets has it uncovered in the vast expanse of the internet? Personal information, sensitive data, and private details - all fair game in the AI's quest for knowledge. However, as we marvel at its insights, we must also confront the uncomfortable truth: ChatGPT may be harbouring secrets that are not its to keep. In this section, we will explore the privacy implications of ChatGPT's training data and ask the tough questions: What does it know, and how can we ensure it does not spill the beans?

3.1 Privacy Policy and Laws

OpenAI's privacy policy is a crucial document that outlines how user data is collected, processed, shared, and deleted. The policy states that various forms of personal information, including account details, user content, and social media data, are collected when users create accounts to access ChatGPT services. Additionally, OpenAI automatically collects log data, usage data, device information, cookies, and analytics through its services. The policy also indicates that certain personal information may be shared with third-party entities, such as cloud vendors, web analytics service providers, government authorities, and industry peers, for business operations and legal compliance. Users may not be notified of such disclosures.

While OpenAI is responsible for managing and protecting user data, users have certain rights, including access to their personal information, the ability to update or delete it, and the right to restrict how OpenAI processes their data. However, the regulation of OpenAI's handling of personal information depends on privacy laws in different countries, such as the Nigeria Data Protection Regulation (NDPR) in Nigeria, the General Data Protection Regulation (GDPR) in Europe, and the California Consumer Privacy Act (CCPA) in the US. Although OpenAI claims to comply with these laws, some users may still be concerned about storing and handling their personal information. For instance, disabling chat history may not be enough to alleviate all privacy concerns related to ChatGPT.

3.2 Privacy Risks

ChatGPT's data handling practices raise significant privacy concerns, particularly regarding GDPR compliance. One major issue is privacy leakage due to public data exploitation. ChatGPT's training process involves scraping data from various sources, including websites, posts, books, and articles, which may contain personal data. The large and growing dataset (over 570 GB) increases the likelihood of including personal data without proper consent from individuals. This practice may violate privacy laws like the above-mentioned, and as LLMs become more widespread, the potential for privacy violations and their impact on individuals will only grow.

Another major issue is privacy leakage due to personal input exploitation in ChatGPT. The model's reinforcement learning component, which relies on user prompts to improve responses, raises concerns about the management and security of user data. This has led to investigations and bans in several countries, including Italy, which initially banned ChatGPT due to GDPR violations. Although new user controls have been added, concerns persist, and other countries like Canada, Germany, Sweden, and France are investigating the model.

The risk of privacy leaks is further increased by the challenge of ensuring the absolute security of personal data stored on OpenAI's cloud or third-party servers, particularly given the frequency of cybersecurity incidents. Moreover, ChatGPT's ability to infer sensitive information from user inputs highlights the limitations of its approach to avoiding sensitive information and the potential for privacy leaks. Beyond the privacy concerns related to public data and user inputs, the issue of privacy leakage from Large Language Models (LLMs) is currently under investigation. Unlike traditional deep learning models, LLMs like ChatGPT are more challenging to attack due to their limited model parameter accessibility and API-based usage. However, vulnerabilities have been identified, such as multi-step jailbreaking privacy attacks on New Bing, which can extract personal information. Probing attacks can also be used to detect personal data leakage. It is essential to identify these vulnerabilities to understand the potential risks and develop robust privacy preservation solutions to mitigate them, ensuring the protection of user data.

4. ChatGPT Vs ETHICS

AI technology has positively and negatively impacted human security, privacy, and dignity. On the positive side, AI has been utilised to enhance privacy through techniques like federated learning (Kairouz et al., 2021) and machine unlearning (Wu et al., 2022) and has improved various aspects of life, such as automobile technology. However, AI also poses significant risks, including adversarial attacks like poisoning (Biggio et al.,2012), backdoors, membership inference, and model inversion attacks, which can lead to information leakage and compromise privacy. Additionally, accidents involving AI-controlled systems can jeopardise human physical security and well-being. To ensure the responsible and ethical deployment of AI, it is essential to balance its benefits and risks and continue researching and implementing measures to mitigate and address these challenges.

AI technology has far-reaching implications for society, raising important questions about fairness, impartiality, accountability, and transparency. One major concern is the potential for AI models to perpetuate and amplify biases present in their training data, leading to harmful behaviours and discriminatory practices. Furthermore, the complexity of AI models makes it difficult to understand their decision-making processes, hindering our ability to control their behaviour and ensure they align with ethical principles. Additionally, the growing use of AI algorithms for training models has significant environmental implications, including increased electricity consumption, carbon emissions, and electronic waste. As the demand for computational resources continues to rise, addressing these issues and developing sustainable AI practices that prioritise transparency, accountability, and environmental responsibility is essential.

4.1 Legal and Ethical Challenges

The advent of ChatGPT has raised significant legal challenges, primarily due to the lack of regulations governing content created by non-human entities. The copyright ownership of ChatGPT-generated texts is a contentious issue, as it can produce original content, making it difficult to determine ownership. This raises questions about using ChatGPT's responses for academic purposes and whether it should be credited as a co-author. Moreover, there is uncertainty about accountability if ChatGPT-generated content is misused for malicious purposes. Governments are collaborating to develop comprehensive regulations addressing legal concerns, including copyright issues.

Establishing a balanced framework for attribution, ownership, and responsible use of AI-generated content is crucial. Additionally, ChatGPT's writing proficiency has raised concerns among professional writers, who argue that it poses a threat to their livelihoods and raises issues of fairness. The technology gap between high-income and low-income countries may also widen, exacerbating existing disparities.

5. CONCLUSION AND RECOMMENDATIONS

This paper delves into the intricacies of ChatGPT, a paradigmatic Large Language Model (LLM), scrutinising its technological underpinnings, features, limitations, and applications, particularly emphasising the pressing concerns of security, privacy, and ethics. Despite its remarkable capabilities, ChatGPT grapples with a multitude of challenges, including:

- The propensity for inaccurate responses and the concomitant difficulties in plagiarism detection
- The hallucination problem, which precipitates the dissemination of misinformation and potentially deleterious consequences
- Security vulnerabilities, such as prompt injection and data poisoning, which can culminate in erroneous decision-making
- Privacy leakage, which is a formidable concern necessitating stringent compliance with privacy laws and the development of efficacious detection measures
- Ethical considerations, including the perils of bias and manipulation, which can have far-reaching consequences for marginalised groups
- Plagiarism and copyright violations underscore the need for sophisticated AI-written text detectors and watermarking techniques

To surmount these hurdles, joint efforts are required to:

- Implement rigorous input validation and filtering mechanisms. This means setting clear rules for what users can input, like not allowing harmful words or phrases. We also need to use filters to block any harmful content that might get through.
- Enhance the accuracy of ChatGPT and integrate human oversight to ensure the veracity and appropriateness of AI-generated content. This entails improving the AI model's understanding of what it's being asked to do and having humans review the content to ensure its accurate and appropriate. This way, we can catch any mistakes or confusing information.
- Identify and rectify security vulnerabilities quickly. This will mean regular system checks for any holes or gaps that hackers could use to get in. If we find any weaknesses, we must address them immediately to keep the system safe.
- Develop and deploy privacy leakage detection measures, like encryption, to keep user information safe. We also need to create tools to detect potential privacy breaches so we can stop them before they happen.
- Address ethical concerns by cultivating diverse teams and identifying and mitigating bias in AI systems. This can be done by having diverse teams working on AI development and testing the systems regularly.
- Create and refine AI-written text detectors and watermarking techniques to safeguard intellectual property and ensure proper attribution. This means developing tools to identify AI-generated content and adding watermarks to AI-generated text so we know who created it.

REFERENCES

1. B. Biggio, B. Nelson, P. Laskov(2012). Poisoning attacks against support vector machines. https://arxiv.org/pdf/1206.6389.pdf
2. S.S. Biswas. Potential use of Chat GPT in global warming. Annals of Biomedical Engineering, 51 (6) (2021), pp. 1126–1127. Google Scholar.
3. S.S. Biswas (2023). Role of Chat GPT in public health. Annals of Biomedical Engineering, 51 (5) (2023), pp. 868-869. https://link.springer.com/article/10.1007/s10439-023-03172-7
4. T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33 (2020), pp. 1877-1901. Google Scholar.
5. A.N. Çayır, T.S. Navruz (2021). Effect of dataset size on deep learning in voice recognition. Proceedings of the 2021 3rd International Congress on Human-Computer Interaction, Optimisation and Robotic Applications (HORA), IEEE (2021), pp. 1-5. Google Scholar.
6. M. Chen, J. Tworek, H. Jun, Q. Yuan, H.P.d.O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. (2021). Evaluating large language models trained on code. https://arxiv.org/pdf/2107.03374.pdf
7. J. Devlin, M.W. Chang, K. Lee, K.Toutanova.(2018). BERT: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/pdf/1810.04805.pdf
8. Y. Fu, H. Peng, T. Khot (2022). How does GPT obtain its ability? Tracing emergent abilities of language models to their sources. https://rb.gy/5ez0w
9. W.R. Huang, J. Geiping, L. Fowl, G. Taylor, T. Goldstein Metapoison: Practical general-purpose clean-label data poisoning. Advances in Neural Information Processing Systems, 33 (2020), pp. 12080–12091. https://doi.org/10.1364/oe.390297
10. P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14 (1-2) (2021), pp. 1–210. Google Scholar.
11. N. Kaloudi, J. Li (2020). The AI-based cyber threat landscape: A survey. ACM Computing Surveys (CSUR), 53 (1) (2020), pp. 1-34. Google Scholar.
12. Y. Liu, M. Lapata (2019). Text summarisation with pre-trained encoders. https://arxiv.org/pdf/1908.08345v2.pdf
13. Lu, H. Zhang, Y. Zhang, X. Wang, D. Yang (2023). Bounding the capabilities of large language models in open text generation with prompt constraints.https://arxiv.org/pdf/2302.09185.pdf
14. OpenAI (2023). GPT-4 technical report. https://arxiv.org/pdf/2303.08774.pdf
15. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. (2022). Training language models to follow instructions with human feedback. https://arxiv.org/pdf/2203.02155.pdf.
16. Radford, K. Narasimhan, T. Salimans, I. Sutskever (2018). *Improving language understanding by generative pre-training.* https://api.semanticscholar.org/CorpusID:49313245
17. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever (2019). *Language models are unsupervised multi-task learners.* Google Scholar.
18. S. Ranathunga, E.S.A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, R. Kaur. Neural machine translation for low-resource languages: A survey. ACM Computing Surveys, 55 (11) (2023), pp. 1–37. https://dl.acm.org/doi/10.1145/3567592
19. Roberts, C. Raffel, N. Shazeer (2020). *How much knowledge can you pack into the parameters of a language model?* https://arxiv.org/pdf/2002.08910.pdf
20. G. Wu, M. Hashemi, C. Srinivasa. *PUMA: Performance unchanged model augmentation for training data removal.* Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Menlo Park (2022), pp. 8675–8682. Google Scholar.
21. Xiaodong Wu, Ran Duan, Jianbing Ni. (2024). Unveiling security, privacy, and ethical concerns of ChatGPT. Journal of Information and Intelligence, Volume 2, Issue 2. https://doi.org/10.1016/j.jiixd.2023.10.007