An Ameliorated Method to Find Attacks in the Age of LLMs.

Mr.S TARUN TEJAS[1],Mr. G M ANAND REDDY[2]

[1]M.Tech Student, Dept. of CSE, RLJIT
[2]Associate Professor, Dept. of CSE, RLJIT

Abstract

Social engineering attacks have been around for a while, exploiting people's trust to achieve unethical goals [6, 7]. LLMs can also automate many tasks, which means attackers can target more people with personalised messages without spending as much time or money [3]. This significant change means we need to rethink how we train people about security and develop more effective ways to detect these new threats [5, 14]. These intelligent AI models can quickly and easily produce highly relevant and personalised messages [3]. These intelligent AI models can generate highly relevant and customised messages with ease and speed [3]. This has facilitated the launch of more convincing and successful attacks by cybercriminals, including those who lack technical expertise [3]. LLMs can also automate many tasks, allowing attackers to spend less time and money targeting a larger number of people with customised messages [3].

Keywords: AI-generated Social Engineering, AI-powered Phishing Kits, Deepfake Videos, Fake Persona Creation, FraudGPT, Generative AI (GenAI), Impersonation Attacks, Large Language Models (LLMs), Malicious LLMs

I. Introduction

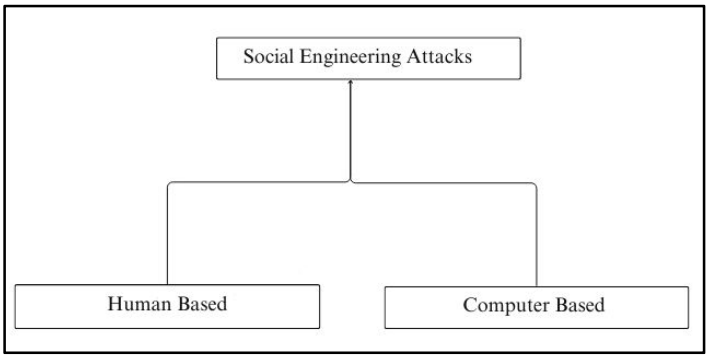

Fig.1: Types of Attacks

This article will look closely at how LLMs are affecting social engineering attacks [3]. It will explain how LLMs are being used to make these attacks better yet, look at some real examples to see what's happening [1, 2], figure out why these attacks are hard to spot [3], and suggest some good ways for people and organisations to protect themselves [5, 14]. This will also talk about where this threat might go in the future and the ethical problems that come with using this powerful AI for bad things [3].

II. How LLMs Enhance Social Engineering Attacks

LLMs have made social engineering attacks much more advanced and effective in several ways [3]. One big improvement is how LLMs can help attackers personalise and target their attacks better by automatically gathering information from the internet (OSINT) [3]. AI tools, using models like GPT-4o and Claude 3.5

Sonnet, can quickly find and analyse public information about potential victims from places like social media and websites [3, 8]. Research shows that these AI models can create pretty accurate profiles for most people, around 88%, and only a few have incorrect information [3]. This automatic information gathering saves attackers a lot of time and effort, letting them create detailed profiles with personal interests, activities, and possible weaknesses [3, 10]. LLMs can also figure out deeper personal qualities from social media posts and other online data, making the targeting even better [8, 11]. For example, GenAI can quickly search the internet for all sorts of data and organise it to build detailed profiles of targets [3]. This level of detail allows attackers to create very personalised phishing messages that seem real to the person receiving them, making them more likely to be successful [9].

LLMs also make it possible to automate the creation and launch of attacks on a much larger scale than before [3]. Studies have shown that AI-automated spear phishing attacks can trick over 50% of people into clicking, which is as good as human experts and much better than regular phishing attempts [3]. LLMs can create lots

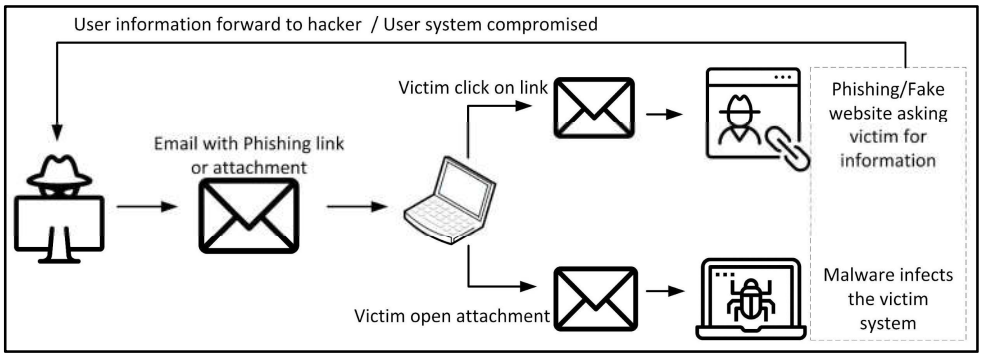

Fig.2: Exploitation Process

of personalised phishing messages in just minutes, which used to take a lot of work to do by hand [3]. Tools like ChatGPT can even create entire attack plans, starting with a spear-phishing email [3]. This automation not only makes attacks faster and reaches more people but also makes them much cheaper, possibly cutting costs by up to 50 times compared to doing it manually [3]. Being able to automate the whole spear-phishing process, from finding information to writing and sending emails, lets attackers target more people with personalised messages in much less time than before [3].

Also, LLMs greatly improve how sophisticated social engineering attacks sound, making it harder to spot them using traditional methods [3]. These models can generate text that sounds like it was written by a native speaker, without the grammar and spelling mistakes that used to be common in phishing attempts, and help people identify them [9]. ChatGPT, for example, can write targeted phishing emails with perfect grammar [3]. GenAI lets attackers communicate almost as well as native speakers and can even be trained on local dialects to sound even more authentic [3]. This better quality of AI-generated text makes it much harder for users to tell if a message is real or fake, since one of the main clues people were taught to look for is now gone [5, 14]. LLMs are also being used for more advanced impersonation, including text, voice, and video [3]. AI video technology offers a new and more convincing way to impersonate someone compared to old methods like

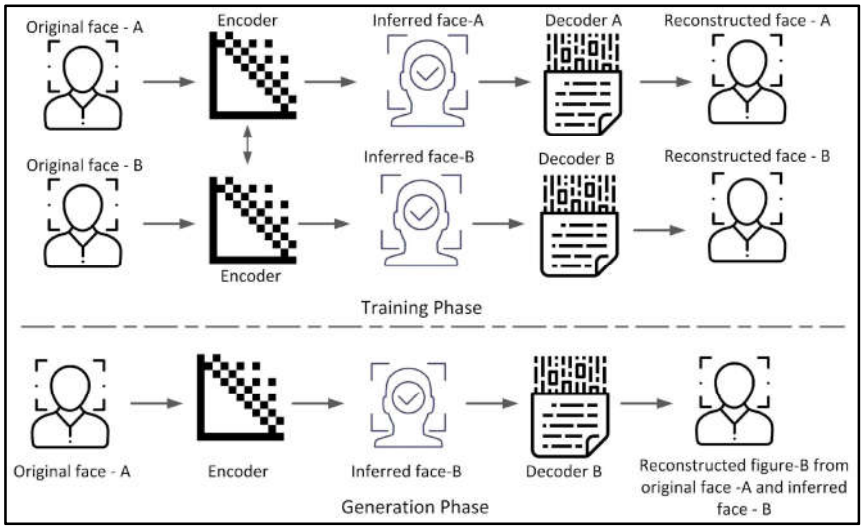

Fig.3: Impersonating Attack Process.

masks [3]. Voice cloning technology can now copy someone's voice using just a few seconds of audio, making voice phishing attacks very realistic [3]. LLMs can even create digital copies of people that can simulate how they behave and respond, letting attackers create very personalised interactions [3]. This ability to impersonate people in different ways makes social engineering attempts much more believable, as people are more likely to trust communications that seem to come from familiar individuals through different channels [4, 12].

Adding to the problem is the appearance of malicious LLMs and AI-powered phishing kits [3]. Tools like GhostGPT are specifically made for hackers and don't have the safety features that regular models do [3]. FraudGPT is advertised on dark-net forums as a tool for cybercriminals to create phishing emails, fake websites, and even malware that's hard to detect [3]. WormGPT has been used a lot for business email compromise attacks [3]. Also, AI-powered phishing kits are now being sold openly on platforms like Telegram, often including ChatGPT-like language models and the ability to gather information from LinkedIn [3]. The availability of these malicious LLMs and phishing kits makes it easier for people with less technical skill to commit cybercrime, giving them powerful tools to launch highly automated and personalised attacks [3].

III. Analysis of Case Studies

The changing world of social engineering attacks with LLMs is shown by several important real-world events. One example is the Bybit $1.4 billion crypto theft, which, while not directly linked to LLMs in the provided text, shows how large and complex attacks could become with the help of LLMs [1]. The attackers hacked into a developer's computer at Safe, a company that provides software for managing cryptocurrency wallets, and put malicious code into the Safe software used by Bybit [1]. This allowed them to change what Bybit's cryptocurrency signers saw on their screens, making a fake transaction look real while authorising a theft [1]. The FBI has said that this big theft was done by the Lazarus Group, a hacking group sponsored by the North Korean government [1].

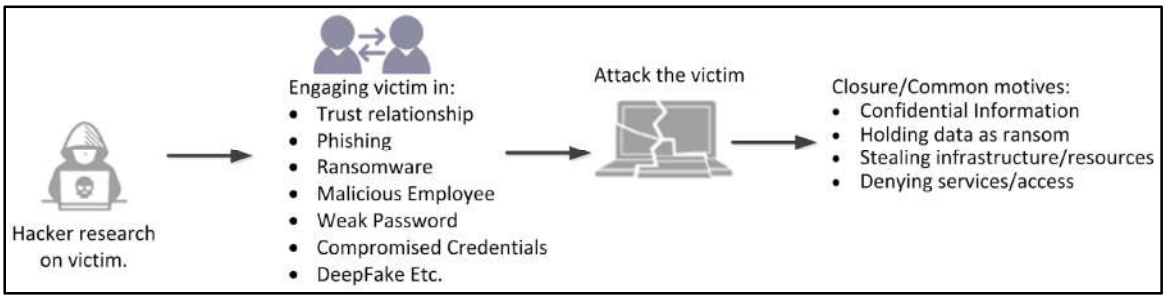

Fig.4: LLMs Attack Process

This event shows that attackers are increasingly targeting not just technical weaknesses but also people and trusted third-party providers in the software supply chain [10]. The fake software interface successfully bypassed the security measures that required multiple signatures, highlighting how powerful social engineering can be even in secure environments [1]. The fact that Bybit's team didn't double-check the transaction details on their hardware wallets was a big reason why the attackers were able to steal the money [1]. This case is a strong reminder of how important it is to have good security practices throughout the software supply chain and to carefully verify important transactions [5, 14].

Another important event is the $330 million crypto theft from an elderly person in the United States [2]. While the exact social engineering methods used aren't fully explained in the text, reports say that the attacker used advanced techniques to get access to the victim's Bitcoin wallet [2]. The victim had a lot of Bitcoin since 2017 and hadn't made any big transactions before, so they were likely targeted because of their large holdings [2]. The stolen funds were quickly moved through several exchanges and converted into Monero, a cryptocurrency that focuses on privacy, which suggests a well-planned and executed attack [2]. This case shows how vulnerable individuals, especially those who might not be very familiar with cryptocurrency security, are to highly targeted social engineering schemes [11]. The lack of past transactions and the quick movement of funds highlight the potential for big financial losses and the difficulty of getting stolen cryptocurrency back [2].While the Bybit and elderly victim cases aren't directly linked to the use of LLMs in the research material, they show the scale and complexity that social engineering attacks can reach [1, 2]. Given how well LLMs can personalise attacks, automate processes, and improve the quality of language used, it's likely that future successful large-scale social engineering attacks will involve these technologies [3]. The deceptive tactics seen in the Bybit attack and the targeted manipulation in the elderly victim theft are things that LLMs can make much more effective, suggesting that such attacks might become more common and harder to detect in the future [3].

IV. Detection Challenges for AI-Generated Social Engineering

The rise of AI-generated social engineering attacks creates big problems for traditional cybersecurity detection methods [3]. Systems that rely on recognising known patterns and characteristics of malicious content struggle to identify the new and personalised content created by LLMs [3]. Because LLMs can easily create phishing emails and other harmful content that are specifically designed for each target, old detection methods become pretty useless [3, 9]. AI-driven attacks can easily get around traditional defences by creating fresh and relevant content that doesn't match any known patterns [3]. The basic difference between the static nature of old security methods and the dynamic, creative abilities of LLMs means we need to move towards more adaptable and context-aware detection approaches [3]. Another big challenge in fighting AI-generated social engineering is that it's getting harder to tell the difference between content created by AI and content created by humans, whether it's text, audio, or video [3]. LLMs can now produce content that looks and sounds very real, making it difficult for both people and traditional computer programs to tell what's authentic and what's fake [3]. AI tools have even shown they can outperform expert human teams in creating convincing phishing campaigns [3]. This blurring of the line between real and fake content makes it harder to trust digital communications and greatly increases the chances of people falling for social engineering attacks [10, 13].

Attackers are also using different ways to trick detection systems by using LLMs [3]. The safety features in many LLMs don't always work, and attackers are good at using special prompts and techniques to get around these controls and create harmful content [3]. Also, LLMs can be used to rewrite or hide existing malware and phishing text, making it harder for security tools to find these threats [3]. This constant back-and-forth between security measures and attacker tricks means we need to develop detection methods that are just as adaptable [3].

To effectively fight AI-generated social engineering, we really need more advanced AI-powered detection methods [3]. Language models themselves can be used to analyse emails and figure out if they might be phishing attempts [3]. LLMs have shown they can look at language patterns and find things that seem off, which might indicate bad intentions [3]. AI-powered systems that look for unusual activity can constantly watch and analyse communication patterns to spot small changes that might mean an AI-generated phishing attempt is happening [3]. The best way to fight AI-enhanced social engineering is to use AI's abilities in understanding language, machine learning, and analysing behaviour to understand the subtle ways AI-generated attacks work [3].

V. Mitigation Strategies and Defence Mechanisms

It takes a combination of technological solutions, user awareness, and robust procedures to combat the growing threat of AI-enhanced social engineering [5, 14]. An important first step is to improve user awareness training, specifically to help people recognise AI-driven social engineering tactics [5, 14]. Training programmes should emphasise the value of critical thinking and verifying unusual requests using multiple methods, as well as teach staff and individuals how to recognise deepfake videos, sophisticated phishing emails, and AI-generated voices [3, 5]. People can learn how to identify and report suspicious activity with regular training sessions and practice scenarios that mimic actual AI-driven attacks [5, 14].

Fighting these threats also requires the use of AI-powered systems that identify anomalous activity and offer threat intelligence [3]. In order to identify and prevent attacks before they cause significant harm, these tools can continuously monitor communication patterns, spot unusual activity, and use the most recent threat information [3]. Subtle indications of manipulation, which are typical in phishing and other deceptive tactics, can be successfully detected by AI-driven behaviour analysis [3, 13].Another crucial strategy to stop attacks is to strengthen email authentication protocols like DMARC (Domain-based Message Authentication, Reporting, and Conformance), DKIM (DomainKeys Identified Mail), and SPF (Sender Policy Framework) [9]. Attackers can be prevented from using phoney email addresses and posing as authentic senders by properly configuring and utilising these protocols [9].Strong verification procedures for critical requests and multi-factor authentication (MFA) for all systems can provide an additional layer of protection and stop unauthorised access or actions, even if an attacker uses social engineering to obtain someone's login credentials [5, 14]. Additionally, organisations must develop incident response plans tailored to address AI-enhanced attacks [5]. In order to minimise potential harm, these plans should specify how to promptly identify, contain, and resolve AI-driven social engineering incidents [5].

In summary, research into watermarking AI-generated content for detection is still ongoing and appears to be a promising future defence against these attacks [3]. Although there are still issues with making it dependable and widely applicable, integrating hidden signals into AI-generated text, audio, and video could offer a technical means of identifying synthetic content [3].

VI. Future Directions in AI-Driven Social Engineering

The future of AI-driven social engineering will likely involve even more advanced and personalised attacks as LLMs get better [3]. These attacks might become almost impossible to tell apart from real communications, highly tailored to individual weaknesses and able to mimic human emotions more accurately [3, 4].

The combination of LLMs with other new technologies, like Augmented Reality (AR) and Virtual Reality (VR), could also lead to new and more immersive ways to attack [3]. Imagine realistic and interactive social engineering scenarios within AR/VR environments powered by LLMs, making deception even more convincing [3]. Also, the field of adversarial AI will likely play a bigger role in social engineering [3].

Attackers will probably use adversarial AI techniques to create attacks specifically designed to avoid AI-powered detection systems, leading to a continuous battle between attackers and defenders [3].

Ultimately, cybersecurity in the age of AI will become a field that changes very quickly, requiring constant adaptation and new ideas from both sides [3]. Organisations and individuals need to be proactive and adaptable in their security, constantly updating their defences to keep up with the evolving threats [3, 5].

VII. Ethical Implications of LLM Use in Social Engineering

Using LLMs for harmful purposes like social engineering raises serious ethical questions [3]. Using advanced technology for malicious activities brings up moral issues about who is responsible, what their intentions are, and the potential for widespread harm [3, 13]. Creating and using tools specifically for cybercrime, like dark LLMs, makes these ethical concerns even bigger [3]. The increasing number of AI-generated deceptions can damage public trust and security in digital communications [12, 13]. If AI can convincingly pretend to be anyone and create realistic fake content, it becomes much harder for users to trust that online interactions are real, which could hurt legitimate communication and business [12].

To reduce the misuse of LLMs, we need to establish rules and ethical guidelines for how they are developed and used, especially in cybersecurity [3]. Governments and industry organisations need to work together to create policies and standards that ensure LLMs are used responsibly and the risk of malicious use is minimised [3].

Finally, it's important to remember that AI in cybersecurity has a dual nature [3]. The same AI technologies that help attackers can also be used to create better security solutions, leading to a constant cycle of innovation in the fight between attackers and defenders [3].

VIII. Methods and Preventive Methods

This section outlines the methods used in AI-enhanced social engineering attacks and the preventive methods.

A. **Methods Used in AI-Enhanced Social Engineering Attacks**

1. **LLM-Driven OSINT Gathering***:* Attackers use LLMs to automatically collect and analyse publicly available information to target individuals.

2. **Automated Phishing Email Generation*:*** LLMs are used to quickly create many personalised phishing emails with good language and grammar.

3. **Realistic Fake Persona Creation*:*** LLMs help create believable fake online identities for impersonation on different platforms.

4. **Voice Cloning and Deepfake Video Generation*:*** LLMs enable the creation of realistic voice copies and fake videos for advanced audio and visual impersonation.

5. **Malicious LLM Utilisation*:*** Cybercriminals use specialised LLMs like FraudGPT and WormGPT for harmful activities, including creating malware and stealing information.

6. **LLM Safety Guardrail Bypass*:*** Attackers use special prompts and techniques to get around safety features in LLMs and generate harmful content.

IX. Preventive Methods and Best Practices

1. **Enhanced User Awareness Training:** Implement training programmes to teach users about AI-driven social engineering tactics, including recognising sophisticated phishing, deepfakes, and AI-generated voices.

2. **AI-Powered Anomaly Detection and Threat Intelligence:** Use AI-driven platforms to monitor communication patterns, identify unusual behaviour, and use threat intelligence for defence.

3. **Strengthening Email Authentication:** Implement and correctly configure SPF, DKIM, and DMARC protocols to prevent email spoofing.

4. **Multi-Factor Authentication (MFA):** Enforce MFA across all systems to add an extra layer of security beyond passwords.

5. **Robust Verification Processes:** Establish strict verification procedures for all unusual or sensitive requests, especially those involving financial transactions or data access.

6. **Incident Response Planning:** Develop and maintain plans to respond to AI-enhanced social engineering attacks.

7. **Regular System Updates and Patching:** Keep all systems and software up-to-date with security patches to fix weaknesses.

8. **Strong Access Controls and Encryption:** Implement strong access controls and encryption to protect sensitive data.

9. **Information Sharing:** Share information with others in the industry to stay informed about new threats and best practices.

10. **AI-Based Security Controls:** Use AI-powered security tools for better threat detection and response.

11. **Context-Based Defences:** Implement security measures that understand the context of communications to better identify threats.

12. **Multi-Layered Security:** Use a comprehensive security strategy with multiple layers of defence.

13. **Advanced Obfuscation Detection:** Use algorithms that can recognise and decode advanced text-hiding techniques.

14. **Regular Rotation of Security Credentials:** Frequently update and change email security keys and passwords to prevent exploitation of compromised keys.

X. Conclusion

In conclusion, the rise of large language models has brought a new level of sophistication and scale to social engineering attacks. LLMs make these attacks much better through improved personalisation, automation, better language quality, advanced impersonation, and the spread of malicious AI tools. Detecting these AI-generated attacks is hard because the content looks so real, and attackers use tricks to avoid detection. Fighting this growing threat requires a comprehensive and adaptable security approach. No single solution is enough; instead, organisations and individuals need to use a combination of technical controls, user education, and strong processes to protect themselves from AI-enhanced social engineering. Constant adaptation and new ideas are important to keep up with the quickly changing threat landscape.

To protect against this evolving threat, individuals and organisations should focus on staying informed about the latest AI-driven social engineering tactics. Using strong security practices, like multi-factor authentication and good email security protocols, is essential. Teaching employees and individuals how to recognise and report suspicious activity, especially involving AI-generated content, is still very important. Investing in and using AI-powered security solutions that can analyse behaviour and spot unusual activity is also crucial. Ultimately, creating a culture of security awareness and being cautious about digital communications will be key to dealing with the challenges of AI-enhanced social engineering in the future.

References

1. *K. Haselton, "ZachXBT: The Crypto Detective Uncovering Blockchain Scams", CCN, Oct. 2023*

2. *Jafri, "ZachXBT tracks $330M Bitcoin stolen in social engineering scam from elderly American," CryptoSlate, Apr. 30, 2025.*

3. *"Defending against Social Engineering Attacks in the Age of LLMs" by Lin Ai, Tharindu Kumarage, Amrita Bhattacharjee, Zizhou Liu, and Zheng Hui.*

4. *"A Study on the Psychology of Social Engineering-Based Cyberattacks and Existing Countermeasures" by Murtaza Ahmed Siddiqi, Wooguil Pak.*

5. *P. Schaab, K. Beckers, and S. Pape, ''Social engineering defence mechanisms and counteracting training strategies,'' Inf. Comput. Secur., vol. 25,no. 2, pp. 206–222, Jun. 2017.*

6. *Salahdine and N. Kaabouch, ''Social engineering attacks: A survey,''Future Internet, vol. 11, no. 4, 2019.*

7. *Yasin, R. Fatima, L. Liu, A. Yasin, and J. Wang, ''Contemplating social engineering studies and attack scenarios: A review study,'' Secure Privacy, vol. 2, no. 4, pp. 1–14, Jul. 2019.*

8. *S. M. Albladi and G. R. S. Weir, ''User characteristics that influence judgment of social engineering attacks in social networks,'' Hum.-Centric Comput. Inf. Sci., vol. 8, no. 1, p. 5, Dec. 2018.*

9. *Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar, ''SoK: A comprehensive reexamination of phishing research from the security perspective,''IEEE Commun. Surveys Tuts., vol. 22, no. 1, pp. 671–708, 2020.*

10. *Z. Wang, H. Zhu, and L. Sun, ''Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods,'' IEEE Access,vol. 9, pp. 11895–11910, 2021.*

11. *S. M. Albladi and G. R. S. Weir, ''Predicting individuals' vulnerability to social engineering in social networks,'' Cybersecurity, vol. 3, no. 1, 2020.*

12. *P. van Schaik, J. Jansen, J. Onibokun, J. Camp, and P. Kusev, ''Security and privacy in online social networking: Risk perceptions and precautionary behaviour,'' Comput. Hum. Behav., vol. 78, pp. 283–297, 2018.*

13. *R. A. M. Lahcen, B. Caulkins, R. Mohapatra, and M. Kumar, ‘‘Review and insight on the behavioral aspects of cybersecurity,’’ Cybersecurity, vol. 3,no.1, p.10, Dec. 2020*

14. *P. Schaab, K. Beckers, and S. Pape, ‘‘Social engineering defence mechanisms and counteracting training strategies,’’ Inf. Comput. Secur., vol. 25 no. 2, pp. 206–222, Jun. 2017.*

15. *L. Ai, T. Kumarage, A. Bhattacharjee, Z. Liu, and Z. Hui, "Defending against social engineering attacks in the age of LLMs," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, Dec. 2023, pp. 9173–9187.*

16. *M. A. Siddiqi and W. Pak, "A study on the psychology of social engineering-based cyberattacks and existing countermeasures," Appl. Sci., vol. 12, no. 12, art. no. 6042, Jun. 2022.*

17. *[5] P. Schaab, K. Beckers, and S. Pape, "Social engineering defence mechanisms and counteracting training strategies," Inf. Comput. Secur., vol. 25, no. 2, pp. 206–222, Jun. 2017.*

18. *[6] F. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," Future Internet, vol. 11, no. 4, art. no. 89, Apr. 2019.*

19. *[7] A. Yasin, R. Fatima, L. Liu, A. Yasin, and J. Wang, "Contemplating social engineering studies and attack scenarios: A review study," Secur. Privacy, vol. 2, no. 4, p. e86, Jul./Aug. 2019.*

20. *[8] S. M. Albladi and G. R. S. Weir, "User characteristics that influence judgment of social engineering attacks in social networks," Hum.-Centric Comput. Inf. Sci., vol. 8, no. 1, art. no. 5, Dec. 2018.*