Towards Enhanced Human Mitigation of Vishing Attacks: Leveraging Large Language Models for Real-Time User Guidance

Gaetano Cimino[1], Vincenzo Deufemia[1]

[1]University of Salerno, via Giovanni Paolo II, Fisciano (SA), 84084, Italy

Abstract

Vishing attacks, a prevalent manifestation of social engineering, exploit human trust and manipulation over phone calls to illicitly obtain sensitive information. As these attacks evolve in sophistication, traditional defense mechanisms struggle to maintain efficacy, necessitating the exploration of alternative solutions. In this context, Large Language Models (LLMs) emerge as a cornerstone for fortifying defenses against vishing attacks. Through harnessing the profound linguistic knowledge embedded within LLMs, there exists the potential to comprehensively analyze conversations, identify subtle indicators characteristic of vishing, and dynamically generate adaptive countermeasures in real-time. This position paper underscores the promising role of LLMs in enhancing cybersecurity defenses against vishing, thereby laying the groundwork for further exploration and advancement in this critical domain.

1. Introduction

With the exponential advancement of communication technologies, the landscape of fraudulent activities has evolved, presenting new challenges for cybersecurity. One such challenge is the rise of vishing, a form of social engineering where attackers manipulate individuals into divulging sensitive information over the phone under false pretenses (Salahdine and Kaabouch 2019; Tulkarm 2021). According to statistics provided by the Federal Trade Commission, during the sole year of 2020, there were over 128,000 occurrences of fraudulent schemes perpetrated through telephone communications, culminating in a significant financial detriment amounting to $108 million for the victims involved[1]. Vishing, derived from "*voice*" and "*phishing*", exploits human psychology and trust, making it a valuable tool for cybercriminals to exploit vulnerabilities in individuals and organizations alike. This trend has been further magnified by the widespread adoption of remote work and virtual interactions on a global scale, particularly accelerated by events such as the COVID-19 pandemic, which has created a fertile environment for the proliferation of vishing attacks. As individuals increasingly rely on digital communication channels for work, socializing, and commerce, they become more susceptible to manipulation and deception over the phone. The absence of visual cues in telephonic conversations exacerbates this vulnerability, as individuals are forced to rely solely on auditory cues and verbal interactions to assess the authenticity of the caller. Furthermore, vishing attacks often target specific

First International Workshop on Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES (DAMOCLES). Workshop co-located with AVI 2024, June 4th, 2024, Arenzano, Genoa, Italy.

g.cimino@unisa.it (G. Cimino); deufemia@unisa.it (V. Deufemia)

0000-0001-8061-7104 (G. Cimino); 0000-0002-6711-3590 (V. Deufemia)

[1]https://blog.knowbe4.com/vishing-attacks-yield-phone-fraud-take-of-over-100-million

demographics or industries, exploiting social or cultural norms to increase their effectiveness. For example, attackers may leverage knowledge obtained from social media or data breaches to personalize their attacks, enhancing their credibility and persuasiveness. Alternatively, they may impersonate trusted authority figures, such as bank representatives or government officials, to instill a sense of urgency or fear in their targets. An illustration of such a scenario was documented by The Wall Street Journal, wherein a vishing attack led to the CEO of a British energy company transferring $243,000 to the bank account of an assailant under the erroneous belief that he was engaged in a legitimate conversation with his superior[2].

Traditional defense mechanisms, such as spam filters and blacklisting (Miramirkhani, Starov, and Nikiforakis 2016), are proving inadequate in light of evolving attack tactics and the sheer volume of fraudulent calls. Therefore, in response to the escalating threat posed by vishing attacks, cybersecurity researchers and practitioners are exploring innovative detection approaches (Dissanayake et al. 2023; Moussavou Boussougou and Park 2023; Lee and Park 2023). While technical solutions can effectively support users in detecting vishing attacks, they may not impart users with the knowledge necessary to enhance their accuracy in recognizing vishing attempts (Huang et al. 2022). Indeed, users must be equipped to comprehend the consequences of their actions and learn to engage in more secure practices (Desman 2003). Concerning this matter, there has been a rapid proliferation of security and awareness training initiatives (Siponen 2000; Kävrestad and Nohlberg 2021), which aim to educate users on what actions to take, why they are necessary, and how to execute them. This shift towards the development of human-technical solutions involves the generation of adaptive aids in real-time to prevent users from inadvertently falling victim to social engineering attacks and reduce their susceptibility (Huang et al. 2022). In this context, Large Language Models (LLMs) possess the potential to serve as an effective tool for mitigating vishing. By leveraging LLMs to analyze conversations and detect linguistic patterns indicative of attack attempts, the aim is to equip individuals and organizations with timely guidance against social engineering attacks (Uddin and Sarker 2024; Heiding et al. 2024). Moreover, the iterative nature of LLM-based approaches enables continuous learning and adaptation to new attack vectors, enhancing their effectiveness in dynamic threat environments.

This position paper seeks to explore the intersection of vishing attacks, communication technologies, and LLM-based mitigation strategies, with a focus on developing an approach to address the evolving challenges posed by vishing attacks, ultimately enhancing the resilience of individuals and organizations against fraudulent activities perpetrated over telecommunication channels.

2. Related Work

In the early stages of vishing detection, the strategy of phone blacklisting emerged as a prevalent method. This entailed the compilation of databases containing identified scam numbers, typically sourced from user submissions. Tran *et al.* (Tran, Hoai, and Choo 2020) delved into this technique further, exploring a blacklisting and whitelisting-based detection approach through their iCaMs system. This system utilizes machine learning (ML) for validating phone numbers within a

[2]https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

client-server architecture. Similarly, Zhang and Gurtov (Zhang and Gurtov 2009) proposed a detection system that relies on recipient whitelists to establish connections with callers. Jeong and Lim (Jeong and Lim 2019) addressed vishing through an intelligence-based detection model, which integrates blacklisting-based and scenario-based rule models with convolutional neural networks (CNNs) to enhance the accuracy of detecting abnormal financial transactions. However, the efficacy of traditional methods relying on blacklisted numbers has waned due to advancements in Voice-over-Internet Protocol (VoIP) technology.

Alternative approaches have been explored for vishing detection. Brabin and Bojjagani (Brabin and Bojjagani 2023) introduced a mechanism employing a Central Banking Server as an Authentication Server alongside a nationwide unique phone number. This authentication mechanism aims to validate phone numbers' authenticity and mitigate vishing attacks. Through simulation and analysis using Scyther, a protocol verification tool, the authors demonstrated that their mechanism offers enhanced protection against vishing attacks. Dissanayake *et al.* (Dissanayake et al. 2023) proposed a system leveraging third-party threat intelligence services to assess the reputation of suspicious artifacts in call conversations. It also utilizes natural language processing (NLP) and ML techniques to examine the content of voice calls, identifying suspicious elements such as phishing keywords, sensitive information, and contextual cues. Boussougou and Park (Moussavou Boussougou and Park 2023) introduced an artificial neural network architecture for detecting Korean vishing attacks. This model integrates a 1-dimensional CNN, a Bidirectional Long Short-Term Memory, and Hierarchical Attention Networks to effectively extract and learn features from word embedding vectors. Furthermore, the model incorporates attention mechanisms to emphasize crucial features, thereby enhancing detection performance.

Given the rapidity of vishing occurrences, real-time detection has become a paramount area of research. Song *et al.* (Song, Kim, and Gkelias 2014) pioneered the iVisher approach, employing Session Initiation Protocol-based (SIP) VoIP to authenticate caller IDs and combat spoofing attempts in real-time. Despite its reliance on user responsiveness and organizational cooperation, iVisher enhances telephone communication security by ensuring the consistency between displayed names and actual caller IDs. Similarly, Kang *et al.* (Kang et al. 2022) introduced DeepDetection, utilizing autoencoders for two-fold authentication against vishing while preserving privacy through local voice data preprocessing. Yoon *et al.* (Yoon and Choi 2022) proposed a federated learning-based approach, prioritizing user data privacy while improving detection accuracy. Despite its emphasis on privacy preservation, their study primarily focuses on the accuracy of detection algorithms, leaving room for comprehensive evaluation using additional performance metrics. Zhao *et al.* (Zhao et al. 2018) proposed an Android application that utilizes NLP techniques and ML algorithms for dynamic call content analysis, while Kale *et al.* (Kale et al. 2021) employed Naive Bayes and CNN algorithms to classify fraudulent calls based on conversation transcripts' intent. Finally, Lee and Park (Lee and Park 2023) proposed a real-time vishing detection system tailored specifically for the Korean language. Through the utilization of fundamental ML models and the transformation of voice files into text using NLP techniques, the system is capable of promptly identifying vishing attempts as they occur. The primary emphasis lies in achieving swift detection rather than the development of intricate models.

While these technical solutions have significantly advanced vishing detection, it is essential to acknowledge the significance of human-technical solutions that empower users with knowledge

and awareness (Breve, Cimino, Deufemia, et al. 2022). By equipping individuals with essential information, human-technical solutions complement technical approaches, thereby creating a comprehensive defense against vishing attacks.

3. Advancing Vishing Attack Mitigation with Large Language Models

Human susceptibility to manipulation and the propensity to trust in interpersonal interactions render individuals the weakest link within the security framework (Pokrovskaia and Snisarenko 2017). Malicious actors exploit this vulnerability to psychologically coerce individuals into divulging confidential information or circumventing security protocols. Consequently, vishing attacks persist despite the availability of technical solutions, highlighting the imperative of user education and support in mitigating such threats. Cybercriminals resort to vishing attacks when conventional methods of exploiting technical vulnerabilities are unfeasible (Aroyo et al. 2018). Furthermore, the expeditious nature of vishing incidents has propelled real-time detection to the forefront of research priorities within our society.

Given the intricate and continually evolving nature of vishing attacks, LLMs play a crucial role in supporting users during real-time interactions with potential fraudsters. Leveraging sophisticated NLP techniques, LLMs meticulously analyze ongoing conversations, discerning linguistic patterns indicative of vishing attempts. Through this analytical process, LLMs provide users with personalized and contextually appropriate answers designed to thwart social engineering attacks and diminish the likelihood of succumbing to fraudulent schemes. In contrast to conventional defense mechanisms reliant on static rules or pattern matching, LLMs possess the inherent capacity to adapt and evolve in response to the evolving landscape of attack techniques and linguistic nuances. Continuously learning from new data and interactions, LLMs refine their comprehension of vishing patterns, thereby augmenting the efficacy of their suggested responses over time. Moreover, the real-time nature of LLM-based assistance empowers users to make informed decisions and take measures to safeguard their personal information and assets. Figure 1 illustrates an example workflow delineating the generation of privacy-preserved answers. Specifically, when a fraudster solicits a user to perform an action during a phone call, the request is submitted to an LLM for analysis. Utilizing either a speech-to-text module to transcribe spoken language into text or a multimodal model capable of handling voice data directly, the pipeline initiates iterative interaction between the user and the LLM to identify potential vishing attack risks associated with the fraudster's request. Upon detection of a data leakage risk, the user seeks recommendations on crafting a response conducive to preserving privacy while gathering information about the attacker. Subsequently, the generated response is relayed to the fraudster, with the user awaiting their subsequent move.

An approach for implementing the iterative interaction between the user and the LLM is referred to as *Iterative Refinement* (Feng et al. 2023). This approach aims to iteratively generate new prompts, which serve as formal instructions for engaging with the LLM, based on the model outputs. Starting with the initial output, the model systematically refines its predictions over successive iterations. Alternatively, a predetermined set of prompts can be crafted following the principles of the *Manual Template Engineering* approach (Liu et al. 2023), wherein domain

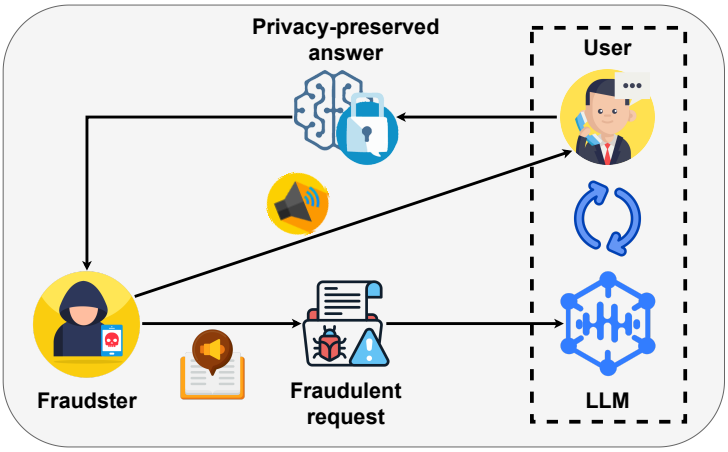

Figure 1: An illustration of the workflow utilized for generating privacy-preserved answers leveraging outputs from an LLM.

experts analyze the task at hand and manually devise prompts for interacting with the LLM. For example, in the context of mitigating a vishing attack, it may prove beneficial to employ three prompts:

(i) soliciting the model's analysis of the risk posed by a request,
(ii) requesting guidance on handling the request, and
(iii) seeking advice on obtaining information about the fraudster.

The assessment of risk inherent in a fraudulent request holds paramount significance. It is imperative to meticulously inform users about potential risks and, in instances of extreme severity, to intervene judiciously to preemptively mitigate any potential harm. However, a critical consideration lies in minimizing user involvement throughout the iterative process. Users should primarily focus on avoiding repercussions from attempted vishing attacks, without becoming unduly preoccupied with the intricacies of LLM interaction. Hence, careful attention must be directed towards disseminating knowledge to users effectively. In this regard, the utilization of intelligent multimodal interfaces may prove invaluable in facilitating the interpretation of LLM outputs and expediting the process of formulating answers (Ahmad et al. 2024; Wang et al. 2024). For instance, LLMs may generate multiple alternative recommendations on how to respond to a fraudulent request, and users may find it cumbersome to sift through lengthy responses presented in a raw format. Consequently, it is essential to ensure that the LLM's response is concise, considering that the user is engaged in a phone call. To enhance this capability, LLMs can be fine-tuned on datasets specific to privacy and cybersecurity contexts (Breve, Cimino, and Deufemia 2022). This fine-tuning process improves their ability to produce contextually appropriate and precise responses. Lastly, given the user's limited attention and the real-time nature of the interaction, the LLM's recommendations should be succinct and to the point, allowing the user to quickly grasp the necessary information without disrupting the ongoing conversation. Therefore, the paramount importance lies in effectively communicating the recommendations of the LLM.

Evaluating the proposed approach for utilizing LLMs in mitigating vishing attacks necessitates a comprehensive assessment framework that encompasses various dimensions:

Accuracy. The accuracy of the LLM in identifying and responding to vishing attempts should be rigorously evaluated through controlled experiments and real-world simulations. This evaluation should include measures such as precision, recall, and F1-score to quantify the model's ability to correctly identify fraudulent requests and provide appropriate responses.

Efficiency. The scalability and computational efficiency of the approach need to be assessed to ensure its feasibility for deployment in real-time environments with large-scale usage. Evaluating the responsiveness of the system and the latency of generating and delivering responses in real-time is essential to ensure a seamless user experience and timely mitigation of vishing attacks. Techniques for optimizing the computational efficiency of the LLM, such as model compression, distributed computing, and caching strategies, should be explored and evaluated to minimize resource consumption and improve system performance.

Usability. The usability of the LLM-based system plays a crucial role in its practical effectiveness. User acceptance and satisfaction with the system interface, as well as the clarity and effectiveness of the generated responses, should be evaluated through user studies and feedback sessions. Moreover, the impact of the system on users' decision-making processes and their ability to effectively navigate vishing scenarios should be assessed to gauge the practical utility of the approach.

Security and privacy. Beyond technical performance and usability, the security and privacy implications of deploying an LLM-based solution for vishing mitigation must also be carefully evaluated. Potential vulnerabilities and risks associated with the system, such as adversarial attacks or unintended information disclosure, should be thoroughly examined and addressed to ensure the integrity and confidentiality of user interactions.

The integration of LLMs into vishing defense strategies represents a promising avenue for bolstering user resilience and reducing susceptibility to fraudulent schemes perpetrated over telecommunication channels. As the capabilities of LLMs continue to advance, their role in enhancing cybersecurity measures is expected to become even more indispensable, paving the way for a safer and more secure digital environment.

Acknowledgments

This work has been supported by the Italian Ministry of University and Research (MUR) and by the European Union - NextGenerationEU, under grant PRIN 2022 PNRR "DAMOCLES: Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES" (Grant P2022FXP5B).

References

Ahmad, Awais, Sohail Jabbar, Sheraz Akram, Paul Anand, Umar Raza, and Nuha Alshuqayran. 2024. "Enhancing ChatGPT's querying capability with voice-based interaction and a CNN-based impaired vision detection model." *Computers, Materials and Continua.*

Aroyo, Alexander Mois, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. 2018. "Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble?" *IEEE Robotics and Automation Letters* 3 (4): 3701–3708.

Brabin, DR Denslin, and Sriramulu Bojjagani. 2023. "A Secure Mechanism for Prevention of Vishing Attack in Banking System." In *2023 International Conference on Networking and Communications (ICNWC),* 1–5. IEEE.

Breve, Bernardo, Gaetano Cimino, and Vincenzo Deufemia. 2022. "Identifying security and privacy violation rules in trigger-action IoT platforms with NLP models." *IEEE Internet of Things Journal* 10 (6): 5607–5622.

Breve, Bernardo, Gaetano Cimino, Vincenzo Deufemia, et al. 2022. "Towards Explainable Security for ECA Rules." In *EMPATHY@ AVI,* 26–30.

Desman, Mark B. 2003. "The ten commandments of information security awareness training." *Inf. Secur. J. A Glob. Perspect.* 11 (6): 39–44.

Dissanayake, Dushan, Prarthana Gamage, Niroopama Kumarasinghe, and Gamage Upeksha Ganegoda. 2023. "Leveraging Artifact Reputation Analysis and Contextual Sentiment Analysis for Advanced Detection of Vishing and Smishing Attacks." In *2023 8th International Conference on Information Technology Research (ICITR),* 1–5. IEEE.

Feng, Jiazhan, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2023. "Knowledge refinement via interaction between search engines and large language models." *arXiv preprint arXiv:2305.07402.*

Heiding, Fredrik, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter S Park. 2024. "Devising and detecting phishing emails using large language models." *IEEE Access.*

Huang, Linan, Shumeng Jia, Emily Balcetis, and Quanyan Zhu. 2022. "Advert: an adaptive and data-driven attention enhancement mechanism for phishing prevention." *IEEE Transactions on Information Forensics and Security* 17:2585–2597.

Jeong, Eui-seok, and Jong-in Lim. 2019. "Study on intelligence (AI) detection model about telecommunication finance fraud accident." *Journal of the Korea Institute of Information Security & Cryptology* 29 (1): 149–164.

Kale, Neha, Shivangi Kochrekar, Rishita Mote, and Surekha Dholay. 2021. "Classification of fraud calls by intent analysis of call transcripts." In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT),* 1–6. IEEE.

Kang, Yeajun, Wonwoong Kim, Sejin Lim, Hyunji Kim, and Hwajeong Seo. 2022. "DeepDetection: Privacy-Enhanced Deep Voice Detection and User Authentication for Preventing Voice Phishing." *Applied Sciences* 12 (21): 11109.

Kävrestad, Joakim, and Marcus Nohlberg. 2021. "Evaluation strategies for cybersecurity training methods: a literature review." In *Human Aspects of Information Security and Assurance: 15th IFIP WG 11.12 International Symposium, HAISA 2021, Virtual Event, July 7–9, 2021, Proceedings 15,* 102–112. Springer.

Lee, Minyoung, and Eunil Park. 2023. "Real-time Korean voice phishing detection based on machine learning approaches." *Journal of Ambient Intelligence and Humanized Computing* 14 (7): 8173–8184.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." *ACM Computing Surveys* 55 (9): 1–35.

Miramirkhani, Najmeh, Oleksii Starov, and Nick Nikiforakis. 2016. "Dial one for scam: A large-scale analysis of technical support scams." *arXiv preprint arXiv:1607.06891.*

Moussavou Boussougou, Milandu Keith, and Dong-Joo Park. 2023. "Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection." *Mathematics* 11 (14): 3217.

Pokrovskaia, Nadezhda N, and Svetlana O Snisarenko. 2017. "Social engineering and digital technologies for the security of the social capital'development." In *2017 International Conference" Quality Management, Transport and Information Security, Information Technologies"(IT&QM&IS),* 16–18. IEEE.

Salahdine, Fatima, and Naima Kaabouch. 2019. "Social engineering attacks: A survey." *Future internet* 11 (4): 89.

Siponen, Mikko T. 2000. "A conceptual foundation for organizational information security awareness." *Information management & computer security* 8 (1): 31–41.

Song, Jaeseung, Hyoungshick Kim, and Athanasios Gkelias. 2014. "iVisher: Real-Time Detection of Caller ID Spoofing." *ETRI Journal* 36 (5): 865–875.

Tran, Manh-Hung, Trung Ha Le Hoai, and Hyunseung Choo. 2020. "A third-party intelligent system for preventing call phishing and message scams." In *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications: 7th International Conference, FDSE 2020, Quy Nhon, Vietnam, November 25–27, 2020, Proceedings 7,* 486–492. Springer.

Tulkarm, Palestine. 2021. "A Survey of Social Engineering Attacks: Detection and Prevention Tools." *Journal of Theoretical and Applied Information Technology* 99 (18).

Uddin, Mohammad Amaz, and Iqbal H Sarker. 2024. "An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach." *arXiv preprint arXiv:2402.13871.*

Wang, Jiayin, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. "Understanding User Experience in Large Language Model Interactions." *arXiv preprint arXiv:2401.08329.*

Yoon, Jun Yong, and Bong Jun Choi. 2022. "Privacy-Friendly Phishing Attack Detection Using Personalized Federated Learning." In *International Conference on Intelligent Human Computer Interaction*, 460–465. Springer.

Zhang, Ruishan, and Andrei Gurtov. 2009. "Collaborative reputation-based voice spam filtering." In *2009 20th International Workshop on Database and Expert Systems Application*, 33–37. IEEE.

Zhao, Qianqian, Kai Chen, Tongxin Li, Yi Yang, and XiaoFeng Wang. 2018. "Detecting telecommunication fraud by understanding the contents of a call." *Cybersecurity* 1:1–12.