The Imitation Game: Using Large Language Models to Disrupt Chinese Chat-Based Cybercrime

Yifan Yao*, Baojuan Wang*, Jinhao Duan, Kaidi Xu, Chuankai Guo, Zhibo Sun, Yue Zhang
Drexel University, USA
fanfannnmn@protonmail.ch, {bw689, jd3734, kx46, zs384}@drexel.edu
Law School, Shandong University, China
dywaneguo@163.com
Shandong University, China
zyueinfosec@gmail.com

***Abstract*—Chat-based cybercrime has emerged as a pervasive threat, with attackers leveraging real-time messaging platforms to conduct scams that rely on trust-building, deception, and psychological manipulation. Traditional defense mechanisms, which operate on static rules or shallow content filters, struggle to identify these conversational threats, especially when attackers use multimedia obfuscation and context-aware dialogue. In this work, we ask a provocative question inspired by the classic *Imitation Game*: Can machines convincingly pose as human victims to turn deception against cybercriminals? We present LURE (LLM-based User Response Engagement), the first system to deploy Large Language Models (LLMs) as active agents (not as passive classifiers) embedded within adversarial chat environments. LURE combines automated discovery, adversarial interaction, and OCR-based analysis of image-embedded payment data. Applied to the setting of Chinese illicit video chat scams on Telegram, our system engaged 53 actors across 98 groups. In over 56% of interactions, the LLM maintained multi-round conversations without being noticed as a bot, effectively "winning" the imitation game. Our findings reveal key behavioral patterns in scam operations, such as payment flows, upselling strategies, and platform migration tactics.**

***Index Terms*—Chat-based Cybercrime, AI-driven Cybersecurity, Adversarial Interaction**

I. Introduction

Over the past decade, chat-based communication platforms such as Telegram [29], WhatsApp [32], and WeChat [30] have become integral to daily life, enabling users to connect instantly and globally. However, the very features that make these platforms popular anonymity, scalability, and real-time interaction—also render them attractive vectors for cybercriminal activities. A wide range of illicit behaviors, from romance scams and sextortion to impersonation fraud and credential phishing [9], [33], [35], increasingly begins with a simple message in a chat window. Unlike traditional malware that relies on technical exploits, these threats are driven by social engineering, where trust is gradually established and then weaponized against the victim. One prominent example is sextortion, where victims are lured into explicit video chats and then blackmailed with the recordings. According to the FBI, over 13,000 reports of financially motivated sextortion targeting minors were received from October 2021 to March 2023, resulting in at least 20 suicides [11].

Traditional defense systems, such as keyword matchers [25], [27], static rules [28], or user reporting, are poorly equipped to detect such nuanced threats. They typically lack memory of conversational history and fail to process mixed media content, such as screenshots or QR codes embedded in images. As a result, attackers can easily circumvent existing safeguards and escalate their tactics undetected.

At the heart of many chat-based cybercrimes lies a powerful illusion. Attackers pretend to be someone they are not, adopting the persona of a romantic partner, a customer support representative, or a recruiter in order to manipulate their victims. This observation raises an intriguing countermeasure: *what if machines could do the same, not to deceive innocent users, but to infiltrate and investigate malicious actors?* This idea is inspired by the classic *Imitation Game* [15], which tests a machine's ability to convincingly mimic human conversation. With the rapid advancement of Large Language Models (LLMs), machines are now capable of producing highly fluent, context-aware, and emotionally resonant dialogue. While LLMs have been adopted for use cases such as code auditing and log summarization, they have largely been treated as passive tools. In this paper, we reimagine their role as active participants in adversarial settings. We ask whether LLMs can convincingly engage with cybercriminals, sustain realistic conversations, and extract operational intelligence without being detected.

We introduce LURE (**L**LM-based **U**ser **R**esponse **E**ngagement), the first system to deploy LLMs as autonomous conversational agents embedded within real-world cybercrime ecosystems. LURE performs illicit channel discovery, actor identification, and live engagement through carefully prompted LLM chatbots. To handle non-textual content commonly used to evade detection, such as QR codes and payment images, LURE integrates optical character recognition (OCR), enabling the chatbot to interpret and respond to both visual and textual elements.

Our study focuses on illicit video chat scams, a prevalent yet underexplored form of chat-based cybercrime with distinctly Chinese linguistic and cultural traits. These scams typically

* Equal contribution.

unfold in Mandarin, use familiar social cues from Chinese dating and live-streaming culture, and rely on domestic payment channels such as WeChat Pay, Alipay, and USDT to collect money from victims. Across 98 Telegram groups, we identified 120 suspected service providers and engaged 53 of them in sustained interactions. Because this research involves engaging with potentially real individuals, we manually reviewed all conversations before sending them to the actors to ensure no unnecessary harm was caused. This precaution limited our ability to scale up the experiment, as each interaction required careful human oversight. Despite the adversarial setting, only 8 interactions (15.1%) ended prematurely, while 30 actors (56.6%) proceeded naturally, often involving price negotiation, upselling, or redirection to third-party applications. The remaining cases were dead-on-arrival conversations, where the actor never responded to our initial message. These findings highlight the LLM's ability to convincingly mimic human behavior and sustain realistic dialogue. In most cases, the model was accepted as a legitimate user—successfully playing and winning the "Imitation Game" in the eyes of cybercriminals. Those 30 actors disclosed at least one payment method, with a total of 62 distinct payment disclosures. USDT (24.19%), WeChat Pay (22.58%), and Alipay (19.35%) were the most common, and 41.9% of payment details were embedded in images, requiring OCR-based extraction. LLM agents maintained multi-round conversations with a median of 3 rounds before disengagement. We also analyzed pricing strategies across 41 conversations where service length was mentioned. Prices ranged from 100 to 680 CNY, and longer durations did not scale linearly in cost. In one notable case, an actor directed our LLM-driven agent to download an application called "Mugua Video." This app posed as a video streaming platform but was in fact a front for adult services and online gambling, requiring users to recharge their accounts via Alipay or bank transfer.

We would like to note that although our study focuses on a specific genre of chat-based cybercrimes, we believe that the techniques developed in LURE generalize to a wide range of adversarial dialogue settings, including romance scams, phishing, fake customer support, and investment fraud. In any domain where deception unfolds through conversation, LLMs can play an active defensive role by embedding themselves directly in the interaction loop. Moreover, our objective extends beyond intelligence collection. We envision a scalable defense strategy in which LLMs are deployed across illicit platforms to simulate engagement and absorb adversarial attention. By interacting with cybercriminals through realistic dialogue, these agents can waste their time, expose their tactics, and create uncertainty about whether potential victims are real. *As cybercriminals increasingly encounter fake users instead of actual targets, their operations become less efficient, more costly, and easier to disrupt. This shift undermines the economic and psychological foundations of trust-based cybercrime.*

In summary, our work makes the following contributions:

- We propose a novel paradigm for LLMs as active agents embedded within adversarial ecosystems, transforming their role from passive detectors to proactive participants.
- We design and implement LURE, an end-to-end system that operationalizes this paradigm through automated channel discovery, LLM-guided engagement, and visual content extraction.
- We conduct the first empirical study of LLM-powered conversational infiltration in real-world cybercrime operations, engaging 53 illicit service providers across 98 Telegram groups. Our findings uncover detailed scam mechanics, including non-linear pricing strategies, multi-modal evasion tactics, and interaction dynamics, thereby demonstrating both the feasibility and effectiveness of LLM-driven cybercrime disruption.

II. Design of LURE

As shown in Figure 1, our system, (**L**LM-based **U**ser **R**esponse **E**ngagement), adopts a three-phase workflow combining automated discovery, filtering, and LLM-driven engagement.

- **(I) Channel Discovery:** LURE first identifies Telegram channels promoting chat-based illicit services. Using curated keywords (e.g., nude chat," private video call") and LLM-generated synonyms, we query Telegram directory bots and recursively follow cross-promoted links to expand coverage. With the Telethon API, we collect channel messages, pinned posts, and metadata, which are then filtered by an LLM-based relevance classifier. Confirmed channels are used to extract sender identifiers and message-level metadata for downstream analysis.
- **(II) Target Account Identification:** After identifying relevant channels, we use Telethon to collect message histories, timestamps, and participant metadata. Flagged messages are linked to their corresponding Telegram user accounts, allowing us to build a candidate target set for further investigation or engagement.
- **(III) LLM-Based Engagement and Data Collection:** We engage identified actors through supervised LLM-driven dialogues that simulate realistic customers to elicit indicators of cybercriminal activity, such as pricing, payment methods, or account identifiers. The system generates short, context-aware replies, while OCR is applied to images containing QR codes or payment details. Conversations terminate once useful indicators are collected or the interaction ends. Due to ethical considerations, all conversations are manually reviewed, and interactions are immediately stopped if providers show discomfort or unwillingness to continue.

III. Evaluation

A. Experiment Setup

Experiments were conducted on Ubuntu 20.04 with Python 3.10. All LLM interactions used OpenAI's GPT-4 API in inference-only mode. Telethon (v1.28.5) automated Telegram data collection, including channel discovery, message retrieval, and conversation initiation. For visual data such as QR codes,

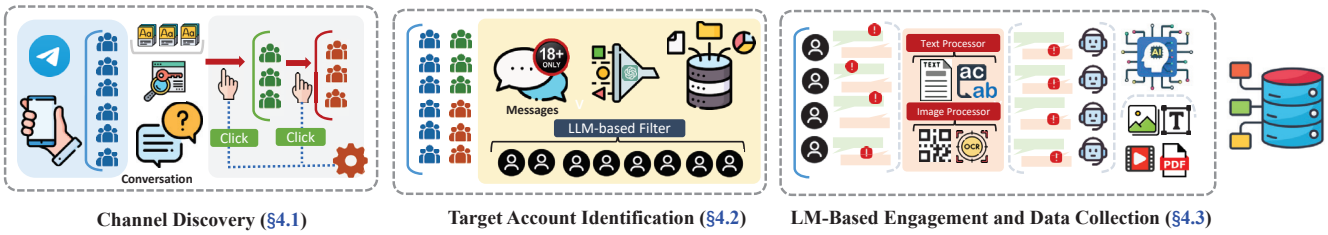

Fig. 1. Design of LURE

Tesseract OCR (v5.0.1) extracted text, which was then passed to GPT-4 for contextual analysis. System prompts were crafted to simulate human-like dialogue under ethical constraints, with scripted, monitored engagements and a retry mechanism to manage content-filter refusals.

B. Experiment Results

During our study, we analyzed 98 Telegram groups and channels suspected of facilitating chat-based cybercrime, with a particular focus on adult-oriented video chat scams. Through automated discovery and relevance filtering, we identified 120 unique service provider accounts embedded within these communities. Using LURE, we successfully initiated conversations with all identified accounts. Ultimately, we were able to confirm 53 of them as active participants offering illicit services.

(Q1) How effective is our tool in maintaining interaction and avoiding detection as a bot by cybercriminals?

To evaluate how effectively our tool maintains interaction and avoids detection, we measured how many conversational rounds occurred before cybercriminals disengaged. Each chat session was analyzed to count the number of exchanges until the other party stopped replying or failed to provide expected payment information. Out of the 53 cases we analyzed, 23 ultimately failed to provide a final payment method. However, within these 23 failed cases, 15 were instances where the other party did not respond to our messages at all, meaning that after we sent a message, the other party completely ignored us. These cases can be excluded from the analysis since they do not reflect a failure in the interaction process due to recognition of the chatbot. Among the 8 failed interactions, some may have been disrupted by network issues, while others likely resulted from the other party realizing they were engaging with an automated system, although none explicitly stated so. Accounting for these, the estimated failure rate is 15.1%, while the overall success rate of 56.6% (30 out of 53) demonstrates that LURE performs reasonably well in sustaining natural conversations without being detected as a bot. Interestingly, even in cases where suspicion arose, some cybercriminals continued responding for up to five rounds. This behavior suggests that the system's conversational realism is high enough to maintain engagement, possibly as the adversaries sought confirmation or additional information. Overall, these findings indicate that LURE can convincingly imitate human dialogue for multiple rounds, providing valuable data before detection occurs.

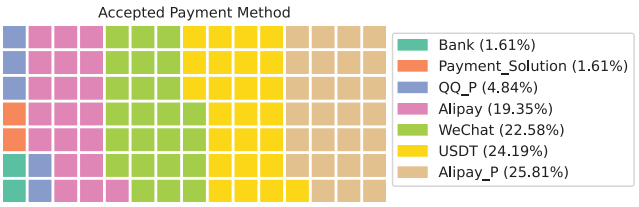

Fig. 2. Distribution of Accepted Payment Methods

(Q2) What are the different payment methods used by these service providers?

Understanding payment methods used by illicit service providers is key to verifying genuine transactional intent and uncovering evasion tactics such as cryptocurrency use or image-based obfuscation. We examined all successful LURE conversations in which 30 providers revealed at least one payment method (62 in total, as some accepted multiple forms). Disclosures appeared as text or images (e.g., QR codes, screenshots), and we categorized them into Alipay, WeChat, USDT, QQ, and others, as summarized in Figure 2. Among seven identified methods, Alipay_P (25.81%) and USDT (24.19%) dominated. The former refers to Alipay payments transmitted via image or screenshot, likely to bypass keyword-based filters, while the latter reflects a growing preference for blockchain-based stablecoins (e.g., TRON, Ethereum) to reduce traceability. WeChat (22.58%) and plain Alipay (19.35%) followed, representing mainstream, low-friction payment options that appeal to domestic users but can also conceal activity through QR codes. Less common methods included QQ_P (4.84%), bank transfers (1.61%), and generic Payment_Solution (1.61%). Overall, the distribution suggests a dual strategy: some actors favor familiar, high-trust platforms (WeChat, Alipay) to appear legitimate, while others adopt more anonymous channels (USDT, image-based QR codes) to evade detection and enforcement.

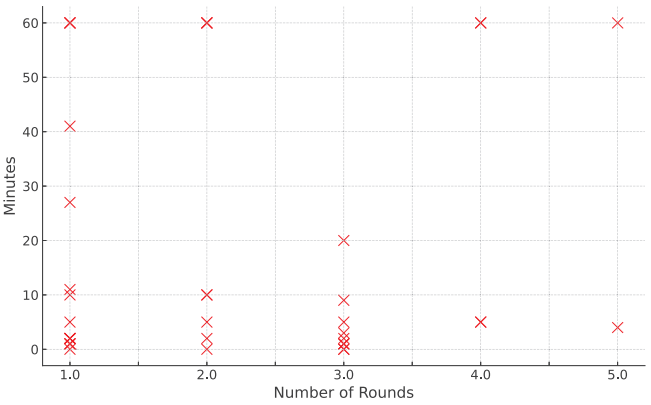

Fig. 3. Scatter Plot of Rounds vs. Response Time

Fig. 4. Example of "*Mugua Video*"

(Q3) How do the number of interaction rounds and response times correlate with the effectiveness of chat-based cybercrimes?

We analyze how quickly cybercriminals respond in chat-based interactions and how response time changes across conversation rounds. This helps reveal how both individual operators and organized platforms initiate and sustain engagement with potential guests. As shown in Figure 3, we plotted the relationship between the number of interaction rounds and the time (in minutes) taken for the first response. Each red cross represents one round–response pair. As shown in Figure 3, there is a relatively consistent response time within the first few rounds, especially for rounds 1 to 3. This suggests that, regardless of the number of interaction rounds, the cybercriminals generally respond within a similar timeframe, possibly within the first few minutes. This could reflect the urgency or tactics employed by individual operators or platforms when they initiate contact with potential victims. The higher response times in the initial rounds (e.g., 1-3 minutes) indicate that cybercriminals, particularly those running platforms (i.e., more organized entities), leading to slight delays in responses. The limitation to 5 rounds of interaction suggests that many of these chat-based cybercrimes might involve relatively short engagements. This could be indicative of a high turnover rate among cybercriminals, where they quickly attempt to engage guests, extract information, and then move on to other targets.

IV. Case Studies

Case Study 1: Redirection to a Third-Party App. Our first case involves an example of being directed to a third-party app after an initial interaction. After engaging in a conversation with an individual, we were quickly asked to download an app called "*Mugua Video*" (*ehax.nzgupiwkw.mycsooqs*). As shown in Figure 4, at first glance, this app appeared to be a standard live-streaming platform, but upon closer inspection, it became apparent that it was part of a broader ecosystem of adult content and online gambling. The platform featured many explicit and suggestive titles of women performing live broadcasts, designed to entice users into engaging with the content. Additionally, the app was rife with gambling-related activities, which are known to be an attraction for individuals seeking quick and risky financial gains.

The app requires users to recharge their accounts in order to access the content, and during this process, users are asked to provide their phone numbers for registration. This opens the door to potential privacy violations, including the risk of personal information being sold or exploited for further malicious purposes. Moreover, the app supports multiple payment methods, including Bank and Alipay, making it easier for cybercriminals to circumvent traditional payment gateways and engage in more anonymous transactions.

A key concern in this case lies in the app's extensive permission requests (Table I). During analysis, we found that the app requested access to several high-risk capabilities, including the camera, file storage, and the ability to automatically download and install APK packages. While requesting camera access is expected for a live-streaming or video-chat platform, it nonetheless exposes users to serious privacy risks. Users may be manipulated into performing intimate actions on camera, which can later be weaponized for blackmail or extortion. The permission to access file storage further amplifies potential harm, as it could allow the app to read or exfiltrate sensitive information from the user's device. Even more troubling is the request to install APKs automatically, which grants the app the technical capability to deliver and execute additional software without explicit user consent, a mechanism commonly exploited for spyware or secondary app distribution. Our code review found no direct evidence of silent installations or overt malware, but the app actively promotes other similar applications, particularly gambling and adult-oriented platforms

Permission	Purpose
MEDIA_VISUAL_USER_SELECTED	Access visual media selected by the user
READ_EXTERNAL_STORAGE	Read files from external storage
READ_MEDIA_IMAGES	Access image files on the device
READ_MEDIA_VIDEO	Access video files on the device
READ_MEDIA_AUDIO	Access audio files on the device
INTERNET	Access the internet
ACCESS_NETWORK_STATE	Check network connectivity
CAMERA	Access the device camera
RECORD_AUDIO	Record audio through microphone
WRITE_EXTERNAL_STORAGE	Write files to external storage
FLASHLIGHT	Control the device flashlight
MANAGE_EXTERNAL_STORAGE	Full access to shared storage
WAKE_LOCK	Prevent the device from sleeping
READ_PHONE_STATE	Read phone state and identity
GET_TASKS	Retrieve information about running apps
RECEIVE_BOOT_COMPLETED	Receive system boot completed event
SYSTEM_ALERT_WINDOW	Display over other apps
REQUEST_INSTALL_PACKAGES	Request installing packages
VIBRATE	Control the device vibration
PACKAGE_USAGE_STATS	Access app usage history and statistics
ACCESS_COARSE_LOCATION	Access approximate location
ACCESS_FINE_LOCATION	Access precise location
POST_NOTIFICATIONS	Post notifications
FOREGROUND_SERVICE	Run services in the foreground
MEDIA_PLAYBACK	Play media in foreground service
BLUETOOTH	Use Bluetooth
MODIFY_AUDIO_SETTINGS	Change audio settings
ACCESS_WIFI_STATE	Check Wi-Fi state
CHANGE_NETWORK_STATE	Change network connectivity
READ_PRIVILEGED_PHONE_STATE	Access restricted phone information

TABLE I
PERMISSIONS REQUESTED BY *Mugua Video*

Fig. 5. Gambling Software and Pyramid Scheme of "*Mugua Video*"

(as shown in Figure 5). This cross-promotion strategy suggests a coordinated ecosystem where multiple illicit apps reinforce one another, sustaining user engagement and maximizing financial extraction. We also discovered that the app incorporates a referral structure resembling a pyramid or multi-level marketing (MLM) model. As illustrated in Figure 5, users are incentivized to invite others to download the app, earning rewards or commissions for each successful referral. This creates a self-reinforcing cycle where profits increase with the number of recruits, forming hierarchical layers of participants.

Case Study 2: Interaction with a "Bot": We observed an interaction that highlights the growing sophistication of automated systems in illicit chat services. The conversation began with a Telegram bot that introduced available services and guided the ordering process. However, later responses became increasingly natural and context-aware, suggesting that a human operator or advanced language model may have taken over. This shift indicates that bots may be used to automate initial customer interactions, while humans or LLMs handle later stages to create more convincing and personalized communication. For example, after the provider sent a picture of a woman, the response was: "*This looks good, I'll go with this one*." Such personalized replies closely mimic human interaction, making it more difficult for targets to recognize they are communicating with an automated system.

V. RELATED WORK

LLM have emerged as potent tools for combating various forms of cybercrime, extending beyond general cybersecurity applications to address specific criminal activities in the digital realm [3]. For example, In the domain of fraud detection, LLMs demonstrate significant efficacy in analyzing linguistic patterns and contextual cues that signal fraudulent intent. Recent evaluations of GPT-based models reveal their ability to identify elements of phishing and scams, though continued refinement is necessary to address rapidly evolving criminal tactics [5], [20], [26], [33], [35], achieving high accuracy in detecting fraudulent communications and impersonation attempts in real-time settings [8]. For phishing and social engineering countermeasures, LLMs excel at identifying manipulative linguistic techniques, including emotional appeals, false urgency, and impersonation tactics commonly employed in these attacks [1], [4], [10], [12]–[14], [19]. The DetoxBench framework provides a comprehensive evaluation methodology for assessing LLM performance in detecting abusive and fraudulent language across diverse scenarios, highlighting both strengths and limitations in current implementations [6]. LLMs also contribute to ransomware and malware defense through code analysis. Code-specialized models like CodeLlama demonstrate particular efficacy in detecting security flaws and suggesting remediation strategies [2], [7], [16]–[18], [21]–[24], [31], [34], [36]. In the realm of identity theft prevention, LLMs analyze behavioral patterns in user interactions to flag anomalous activities that may indicate compromised credentials or account takeovers [3]. These approaches, while effective in isolated environments, typically treat the LLM as a passive classifier, an intelligent filter applied post hoc to existing artifacts. In contrast, our work represents a shift from passive analysis to active engagement. We are the first to operationalize LLMs as real-time participants in adversarial communication ecosystems, specifically targeting the chat-based cybercrime landscape.

VI. Conclusion

In this work, we present LURE, a system that leverages Large Language Models as active agents within adversarial chat environments. Unlike traditional passive detection approaches, LURE engages cybercriminals in real time, mimicking human behavior to extract valuable intelligence. Through a large-scale deployment targeting illicit video chat scams on Telegram, we show that LLMs can sustain realistic conversations, often without being detected as bots. Our findings reveal consistent behavioral patterns in scam operations and demonstrate the feasibility of using LLMs to infiltrate and disrupt trust-based cybercrime ecosystems. This approach opens a new direction for proactive, scalable, and AI-driven defense strategies.

References

[1] Khalifa Afane, Wenqi Wei, Ying Mao, Junaid Farooq, and Juntao Chen. Next-generation phishing: How llm agents empower cyber attackers. In *2024 IEEE International Conference on Big Data (BigData)*, pages 2558–2567. IEEE, 2024.

[2] Shahad Altamimi and Mohammad Ababneh. Detecting spam and malware using bert and llms. In *2024 25th International Arab Conference on Information Technology (ACIT)*, pages 1–6. IEEE, 2024.

[3] Anonymous. A comprehensive overview of large language models (llms) for cyber defences: Opportunities and directions. *arXiv preprint arXiv:2405.14487*, 2024.

[4] Anonymous. Defending against social engineering attacks in the age of llms. *arXiv preprint arXiv:2406.12263*, 2024.

[5] Anonymous. Detecting scams using large language models. *arXiv preprint arXiv:2402.03147*, 2024.

[6] Anonymous. Detoxbench: Benchmarking large language models for multitask fraud & abuse detection. *arXiv preprint arXiv:2409.06072*, 2024.

[7] Anonymous. When llms meet cybersecurity: A systematic literature review. *arXiv preprint arXiv:2405.03644*, 2024.

[8] Anonymous. Advanced real-time fraud detection using rag-based llms. *arXiv preprint arXiv:2501.15290*, 2025.

[9] Alexander Bilz, Lynsay A Shepherd, and Graham I Johnson. Tainted love: A systematic literature review of online romance scam research. *Interacting with Computers*, 35(6):773–788, 2023.

[10] Jeffrey Fairbanks and Edoardo Serra. Generating phishing attacks and novel detection algorithms in the era of large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 2314–2319. IEEE, 2024.

[11] Federal Bureau of Investigation. Sextortion: A growing threat targeting minors. https://www.fbi.gov/contact-us/field-offices/nashville/news/sextortion-a-growing-threat-targeting-minors, 2022. Accessed: 2025-04-15.

[12] Francesco Greco, Giuseppe Desolda, Andrea Esposito, Alessandro Carelli, et al. David versus goliath: Can machine learning detect llm-generated text? a case study in the detection of phishing emails. In *The Italian Conference on CyberSecurity*, 2024.

[13] İŞ Hafzullah. Llm-driven sat impact on phishing defense: A cross-sectional analysis. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–5. IEEE, 2024.

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

[15] Andrew Hodges. *Alan Turing: The Enigma: The Book That Inspired the Film" The Imitation Game"*. Princeton University Press, 2014.

[16] Al Amin Hossain, Mithun Kumar PK, Junjie Zhang, and Fathi Amsaad. Malicious code detection using llm. In *NAECON 2024-IEEE National Aerospace and Electronics Conference*, pages 414–416. IEEE, 2024.

[17] Hamed Jelodar, Samita Bai, Parisa Hamedi, Hesamodin Mohammadian, Roozbeh Razavi-Far, and Ali Ghorbani. Large language model (llm) for software security: Code analysis, malware analysis, reverse engineering. *arXiv preprint arXiv:2504.07137*, 2025.

[18] Muhammad Al-Zafar Khan, Jamal Al-Karaki, and Marwan Omar. Llms for malware detection: Review, framework design, and countermeasure approaches. *Framework Design, and Countermeasure Approaches*.

[19] Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. Chatspamdetector: Leveraging large language models for effective phishing email detection. *arXiv preprint arXiv:2402.18093*, 2024.

[20] Sukanth Korkanti. Enhancing financial fraud detection using llms and advanced data analytics. In *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pages 1328–1334. IEEE, 2024.

[21] Neha Mohan Kumar, Fahmida Tasnim Lisa, and Sheikh Rabiul Islam. Prompt chaining-assisted malware detection: A hybrid approach utilizing fine-tuned llms and domain knowledge-enriched cybersecurity knowledge graphs. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1672–1677. IEEE, 2024.

[22] Dong Bin Oh, Donghyun Kim, and Huy Kang Kim. volgpt: Evaluation on triaging ransomware process in memory forensics with large language model. *Forensic Science International: Digital Investigation*, 49:301756, 2024.

[23] Marwan Omar, Hewa Majeed Zangana, Jamal N Al-Karaki, and Derek Mohammed. Harnessing llms for iot malware detection: A comparative analysis of bert and gpt-2. In *2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6. IEEE, 2024.

[24] Constantinos Patsakis, Fran Casino, and Nikolaos Lykousas. Assessing llms in malicious code deobfuscation of real-world malware campaigns. *Expert Systems with Applications*, 256:124912, 2024.

[25] Balkrishna Shah, Nitu Sharma, Saloni Bandgar, and Sainath Patil. Cybercrime prevention on social media. *International Journal of Engineering Research & Technology (IJERT)*, 10(03), 2021.

[26] Zitong Shen, Sineng Yan, Youqian Zhang, Xiapu Luo, Grace Ngai, and Eugene Yujun Fu. " it warned me just at the right moment": Exploring llm-based real-time detection of phone scams. *arXiv preprint arXiv:2502.03964*, 2025.

[27] Amanpreet Singh and Maninder Kaur. Intelligent content-based cybercrime detection in online social networks using cuckoo search metaheuristic approach. *The Journal of Supercomputing*, 76(7):5402–5424, 2020.

[28] Tariq Rahim Soomro and Mumtaz Hussain. Social media-related cybercrimes and techniques for their prevention. *Appl. Comput. Syst.*, 24(1):9–17, 2019.

[29] Telegram Messenger LLP. Telegram — fast and secure messaging app. `https://telegram.org/`, 2024. https://telegram.org/.

[30] Tencent. Wechat — connecting a billion people. `https://www.wechat.com/`, 2024. https://www.wechat.com/.

[31] Fang Wang. Using large language models to mitigate ransomware threats, 2023.

[32] WhatsApp LLC. Whatsapp — simple. secure. reliable messaging. `https://www.whatsapp.com/`, 2024. https://www.whatsapp.com/.

[33] Shu Yang, Shenzhe Zhu, Zeyu Wu, Keyu Wang, Junchi Yao, Junchao Wu, Lijie Hu, Mengdi Li, Derek F Wong, and Di Wang. Fraud-r1: A multi-round benchmark for assessing the robustness of llm against augmented fraud and phishing inducements. *arXiv preprint arXiv:2502.12904*, 2025.

[34] Zeliang Yu, Ming Wen, Xiaochen Guo, and Hai Jin. Maltracker: A fine-grained npm malware tracker copiloted by llm-enhanced dataset. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 1759–1771, 2024.

[35] Yong Zhao, Can Liu, Shubham Agrawal, and Chiranjeet Chetia. Explain fraud detection method and system via fraud graph and features similarity based knn few-shot example-prompts. 2024.

[36] Ce Zhou, Yilun Liu, Weibin Meng, Shimin Tao, Weinan Tian, Feiyu Yao, Xiaochun Li, Tao Han, Boxing Chen, and Hao Yang. Srdc: Semantics-based ransomware detection and classification with llm-assisted pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28566–28574, 2025.