Volume-10 Issue 01, January -2026 ISSN: 2456-9348
Impact Factor: 8.232

International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

WEAPONIZING LARGE LANGUAGE MODELS: AUTOMATED PHISHING, SOCIAL ENGINEERING, AND MALWARE GENERATION

Syed Mutiuddin
Department of Computer Science, College of Science,
Northern Border University, Saudi Arabia.
syed.syed@nbu.edu.sa

Omar Ibrahim
Department of Computer Science, College of Science,
Northern Border University, Saudi Arabia.
omar.ibrahim@nbu.edu.sa

Karim Rayan
Department of Computer Science, College of Science,
Northern Border University, Saudi Arabia.
Karim.rayan@nbu.edu.sa

ABSTRACT
Large language models (LLMs) have essentially changed the game in natural-language generation. So far, open models and a few powerful commercial assistants have been able to write emails, code, and answer questions with a human-level fluency. Sadly, the factors that bring about the great success of LLMs - contextual awareness, coherence, and ease of use - are the very factors that are being exploited by adversaries.
Phishing as the most common method of social engineering is still going strong: around 96 % of attacks are delivered by emails, and lately, several studies have inferred that 82 % of phishing emails are written with the help of LLMs . Results of evaluations conducted in mid-2025 showed that messages of spear-phishing generated by LLMs had click-through rates of more than 30 %, thus, they were more successful than those created by humans . Simultaneously, the underground communities are jailbreaking open models and launching “dark LLM” services like WormGPT, FraudGPT, and DarkBERT, which get rid of safety guardrails and provide ready-made prompts for business-email compromise, code obfuscation, and malware generation, respectively, as their main functionalities. Besides that, the latest studies reveal that autonomous prompt-injection worms can go on RAG systems, and malware examples like PromptLock and LameHug can have LLMs as a part of their payload that they directly embed into the payload.
This article goes through how LLMs are exploited for phishing, social engineering, and malware, explains the recent case studies, and talks about the ethics of building the resilient AI systems. We, by integrating academic research, industry reports, and dark-web observations, give a well-structured overview of the current capabilities, point out the gaps in the defense, and suggest the responsible AI development as a way forward.

I. INTRODUCTION
Since the initial Generative Pre-traiend Transformer, generative AI has been evolving at a breakneck speed. Current LLMs are able to summarise lengthy documents, write code as well as conduct multi-turn interactions with humans in a nearly indistinguishable way. Companies have welcomed these models for enhancing the efficiency of their employees while the bad actors have leveraged the same capabilities to escalate social engineering and malware campaigns to a level of industrialisation.
Phishing – deceptive communications that trick recipients into giving away their credentials, transferring funds or installing malware – is still the predominant way to gain access to cyber intrusions. A systematisation of knowledge (SoK) study in 2025 shows that almost 96 % of phishing attempts are done through email and 82 % of those emails contain LLM-generated content. In other words, attackers do not have to personally write emotionally impactful messages anymore; they can use public or jailbroken models to generate customised phishing emails which impersonate a company's tone, grammar and style. This research also revealed that spear-phishing emails created by LLMs had the click-through rates over 30 %, which is significantly higher than those

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

human-written campaigns. In other words, AI-assisted phishing is not just a hypothetical danger; it is already leading to an increase in attacker success rates.

In addition to that, LLMs also help adversaries in growing and personalising social engineering. In general, phishing usually makes use of standard bait which can easily be identified by spam filters. Nevertheless, generative models are capable of getting public data (for example, LinkedIn profiles or past emails) and producing customised messages which will strike a chord with a single recipient. The Generative AI Social Engineering Framework considers the issue of realistic content generation, advanced targeting and personalisation, as well as automated attack infrastructure, the three main pillars of this menace. Merging the functionalities in this way attackers receive the capability of producing mail, cloning sites and automating bulk phishing campaigns.

Another trend on the dark web is the emergence of malicious LLM products. The unseen developers in the underground have launched LLMs by which are trained on malware repositories and social engineering scripts, marketed as Phishing-as-a-Service. Among the examples there are WormGPT and FraudGPT – the blackhat variations of GPT-J and GPT-3 transitioned models that have the feature-removing safety guardrails and are able to generate malware code, phishing emails, and fake web pages. The instrument is distributed through a subscription-based sales model (most of the time, around €60 to €200 per month), and it promises capabilities such as generation of phishing in different languages, session-memory for targeted follow-ups, context retention, and code obfuscation. Also, some researchers have developed LLM-embedded malware families such as PromptLock and LameHug/PROMPTSTEAL, which at runtime utilize LLM prompts for command generation and data exfiltration. Thus, a cybercrime ecosystem has been formed wherein perpetrators can escalate their operations with little to no technical skills while the defenders are challenged with identifying and attributing the attacks at an unprecedented level.

This document wants to lay down a clear picture of how the Large Language Models (LLMs) are being turned into weapons. We start with the generative AI social-engineering framework and list the various deception categories in brief, i.e., phishing, disinformation, malware generation and malicious LLMs specialised in certain fields. Next, these dangers are real and we have the examples to prove it: from WormGPT and FraudGPT, blackhat LLM services to LLM-embedded malware (PromptLock, LameHug) and autonomous AI worms. At last, we consider ethical aspects and defensive proposals. In the entire work, along with the risks, we keep stressing the necessity of the responsible AI development, human control and regulatory barriers.

II. SOURCES AND CASE STUDIES

Generative AI social-engineering framework

Researchers at Oxford University suggest that generative AI socially engineered via three main areas: (1) realistic content generation – LLMs are able to produce convincing emails, make a clone of websites, create images or voice deepfakes and even simulate interactive chat conversations; (2) advanced targeting and personalisation – AI models can investigate digital footprints and prepare highly personalised messages deriving from the victim's or organisation's history; and (3) automated attack infrastructure – generative models are capable of automating the phishing email creation, distribution and optimisation processes, learning from user interactions to increase success rates. Figure 1 shows this framework.

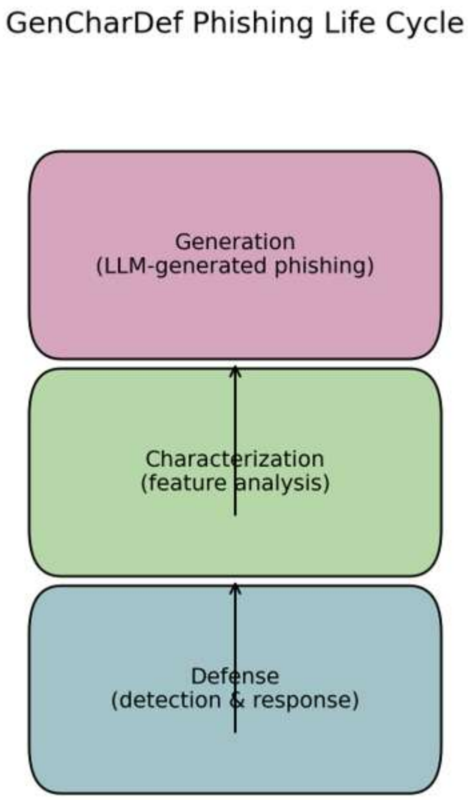
Generative AI social-engineering framework

These columns illustrate the ways in which opponents use generative AI as a weapon. Contemporary LLMs are able to generate fluent, well-formatted messages, thus, they do not have the grammatical errors which gave away the phishing attempts of the early days. Besides that, they enable exact targeting by taking the victim's context for argument and consequently, personalizing the wording. At last, automation set-up allows for situations where AI agents can dispatch several thousands of phishing emails, change their language depending on the answers and coordinate with email servers or chatbots so as to have the efficient, adaptable campaigns.

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

Misuse categories

Surveys of both academic and industry research suggest a Generation–Characterisation–Defense (GenCharDef) framework as a way to comprehend the use of LLMs in the phishing lifecycle. The circle goes on with generation, where attackers employ LLMs or dark models to produce phishing emails, code or deepfake content; with characterisation, where defenders examine these artefacts to locate features and identify malicious patterns; and with defense, where detection and mitigation measures are implemented (Figure 2). The paper is about the generation stage only and the different methods in which adversaries turn LLMs into their weapons. Table 1 gives an overview of primary misuse categories.



GenCharDef phishing life cycle

Table 1. Categories of LLM misuse and examples

Misuse category	Capabilities & examples
Phishing & social engineering	LLMs create very believable spear-phishing emails that are coherent, the language is personalized for the individual targets and the corporate tone is used; a few dark LLMs can produce content in several languages and also allow continuations . In addition, models become a source of writing the text for voice phishing audios or the cloning of websites to get credentials.
Disinformation & deepfakes	Generative models are capable of creating fake news articles, social-media posts, and images. A scam like CEO fraud can be facilitated by voice cloning and video deepfakes. For instance, Rapid7 shared the story of a Ferrari executive who prevented a deepfake call by asking a question that only the CEO could answer.
Malware generation & exploitation	LLMs are capable of producing complicated code, polymorphic malware, and scripts that exploit vulnerabilities. Research highlights that although models can create harmful code, the code is generally flawed and needs human intervention for debugging and finalizing . In any case, attackers continue to use these means for gathering information and finding vulnerabilities.
Specialised malicious LLMs	Secret online communities have models that are jailbroken and are trained on malware and fraud data. Some of these models are named WormGPT, FraudGPT, DarkBard, and

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

Misuse category	Capabilities & examples
	WolfGPT. These paid services deregulate the safety guardrails and produce phishing content, malware code, and exploit kits; the price of subscriptions varies from €60 to $200 per month .
LLM-enabled malware & autonomous worms	Recent proof-of-concepts reveal that malware is incorporating LLM calls. PromptLock employs runtime prompts to create a summary of file contents, recognize sensitive data and then generate the commands for exfiltration. LameHug/PROMPTSTEAL has implanted hundreds of API keys and is utilizing an LLM to create system shell commands for data exfiltration. The Morris-II AI worm is a demonstration of how prompt-injection payloads can efficiently spread to the next layer of retrieval-augmented generation systems without human intervention.

Malicious LLM services and case studies

Proof-of-concept malware examples highlight how attackers can integrate large language models (LLMs) in their malicious software. For instance, at runtime, PromptLock fetches an external LLM to get a summary of the file contents, locate sensitive information and create data-exfiltration commands. APT28, with which LameHug/PROMPTSTEAL is associated, is said to have implanted hundreds of API keys in secret and now uses a HuggingFace model to generate shell commands for system reconnaissance and file exfiltration. These instances of malware illustrate the possibility of such harmful software offloading the logic to a language model thus, they are able to operate under the radar of static signatures and make it hard to be put in the blacklist. In addition to malware, the Morris-II AI worm, as an example, demonstrates that prompt-injection payloads can only-invoked-procedures diffuse the said payloads across retrieval-augmented generation systems and thereby cleverly send out constructed messages as well as get hold of the confidential data.

On the other hand, AI-powered generative models can be exploited to produce deepfake content for social engineering. Rapid7 explains the situation in which a Ferrari executive received a voice-imitation call from the CEO, and the fraud was only prevented when the target asked the deepfake a personal question that it couldn't answer. Similar to that are the cases where the fake videos created by AI depict the executives requesting urgent payments or issuing false emergency alerts just to cause panic. The mentioned instances are just some of the ways in which the advancement of multimodal generation technology has widened the social engineering armory.

The table beneath that presents the most prominent malicious LLM services and proof-of-concept malware that have been referred to in the case studies.

Tool/Case	Summary
WormGPT & FraudGPT	Blackhat chatbots marketed by subscription that create phishing emails, malware code, and fraud scripts; based on GPT-J or GPT-3 variants.
DarkBard, WolfGPT & Wrappers	Jailbroken or wrapper models that unbundle guardrails from popular assistants (e.g., Grok, Mixtral), thereby allowing the generation of exploits and creation of remote access tools .
PromptLock	Ransomware prototype, conceptually, that employs LLM prompts for file analysis and, consequently, the generation of exfiltration commands during runtime.
LameHug/PROMPTSTEAL	Malware associated with APT28 that has the capability of implanting a large number of API keys in its code and later on uses an LLM to generate system commands which then transfers the stolen data to a location.
Morris-II AI worm	A self-replicating prompt-injection worm that spreads to retrieval-augmented AI systems one by one and steals confidential data.

III. ETHICAL AI AND DEFENCES

A self-replicating program that uses prompt injection in a gradual manner across retrieval-augmented AI systems, it steals confidential data.

The dual-use nature of LLMs calls for a solid ethical framework. Developers, organisations and policymakers have the shared responsibility of preventing the misuse while keeping the door open for the benefits of generative AI. The main defensive measures are summarised in the principles below:

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

Reliable guardrails: Regularly red-team models to uncover methods of jailbreaking and ensure that safety policies are up to the task. Use multi-layer content filters, keep track of API usage for unusual prompts, and provide prompt injection by allowing strict boundaries between untrusted input and model execution to lessen the risk of prompt injections.
Detection and reaction: Understand that static signatures may be inefficient when malware moves logic to LLMs. Look for embedded API keys and unusual prompt patterns , review LLM API calls for anomalous sequences , and consider generative models as high-risk systems with limited access and trained users.
Open ness and control: Record model training data, biases and limitations; give terms of use that are easy to understand and allow for reports of misuse. Regulators should stop the selling of malicious models and require that AI generated content be labelled. Awareness of the issue and digital literacy through education can prepare users to recognise deepfake scams and verify requests that are unusual.

IV. CONCLUSION

Large language models have democratised advanced text generation, but the same capabilities also empower adversaries. Dark-web services like WormGPT and FraudGPT make phishing and malware generation inexpensive and accessible, while proof-of-concept malware such as PromptLock and LameHug embed LLMs directly into payloads. Even self-propagating worms demonstrate that generative ecosystems can be attack surfaces. Yet defenders are not powerless: applying the GenCharDef framework helps identify detection points, and organisations can mitigate risk through red-team testing, strict AI usage policies and user education. Developers must implement robust guardrails and remain transparent about limitations, and policymakers should regulate malicious AI services and require disclosure of AI-generated content. Ultimately, ethical AI development and collaboration among researchers, industry and governments are essential to harness the benefits of generative models while curbing their abuse.

REFERENCES

1) R. Sura, A. K. Chaudhary, K. Chaganti and S. M. Kona, "AI-Driven Ethical Compliance in Hyper Automation: Building Transparent RPA Systems for Regulatory Adherence," 2025 International Conference on Computing Technologies (ICOCT), Bengaluru, India, 2025, pp. 1-8, doi: 10.1109/ICOCT64433.2025.11118464.
2) Ande, B.R. (2025). AI-Driven Continuous Authentication: Integrating Deep Learning with Multimodal Biometrics for Enhanced Identity Verification. In: Mishra, D., Yang, X.S., Unal, A., Jat, D.S. (eds) Data Science and Big Data Analytics. IDBA 2025. Learning and Analytics in Intelligent Systems, vol 56. Springer, Cham. https://doi.org/10.1007/978-3-032-05373-2_42
3) Chaudhary AK, Pandey AK, Dubey AK. Computer Aided Taguchi-Fuzzy Multi-Optimization of laser cutting process. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*. 2014;26(2):doi:10.3233/IFS-130770
4) A. Sharma, B. B. Chaudhari and S. Kabade, "Artificial Intelligence-Powered Network Intrusion Detection System (IDS) with Hybrid Deep Learning Approach in Cloud Environments," *2025 IEEE North-East India International Energy Conversion Conference and Exhibition (NE-IECCE)*, Silchar, India, 2025, pp. 1-6, doi: 10.1109/NE-IECCE64154.2025.11182913.
5) Fnu, H., Soni, V., Asundi, S., Thummarakoti, S., Kaushik, K., & Soni, M. (2025, May). Design and Deployment of AI-Driven Fraud Detection Models in Cloud-Based Banking Ecosystems. In *2025 International Conference on Networks and Cryptology (NETCRYPT)* (pp. 223-228). IEEE.
6) Kabade, S., Sharma, A., & Kagalkar, A. (2024). Securing Pension Systems with AI-Driven Risk Analytics and Cloud-Native Machine Learning Architectures. *International Journal of Emerging Research in Engineering and Technology*, *5*(2), 52-64. https://doi.org/10.63282/3050-922X.IJERET-V5I2P106
7) Oxford University Paper (2023). “Generative AI Social Engineering Framework.” Presents the social-engineering pillars of realistic content, advanced targeting and automated infrastructure .
8) Fnu, H., Kaushik, K., Gaur, P., Saratchandran, D. V., Thapliyal, A. G., Sidhu, K. S., & Singh, V. (2025, April). Multi-Agent Systems for Collaborative Financial Decision-Making Over Distributed Network Architectures. In *2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT)* (pp. 939-944). IEEE.

IJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

9) Gopal Varma, S. C., & Chaudhari, B. (2025). Federated Learning in Financial Data Privacy: A Secure AI Framework for Banking Applications. *International Journal of Emerging Trends in Computer Science and Information Technology*, 101-110. https://doi.org/10.56472/ICCSAIML25-112
10) .NET Core Microservices for Zero-Downtime AuthHub Migrations. (2025). *European Journal of Engineering and Technology Research*, *10*(5), 1-4. https://doi.org/10.24018/ejeng.2025.10.5.3288
11) AI-Driven Behavioral Biometrics for 401(k) Account Security. (2025). *International Research Journal of Advanced Engineering and Technology*, *2*(06), 23-26. https://doi.org/10.55640/irjaet-v02i06-04