

ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ/INFORMATICS AND INFORMATION PROCESSES

DOI: https://doi.org/10.60797/itech.2026.11.9 EDN: XRLKLP

RISKS OF USING GENERATIVE LANGUAGE MODELS IN PHISHING ATTACKS: ANALYSIS AND THREAT MODELING

Research article

Negoda A.N.[1, *]

[1] GPM Digital Innovations LLC, Moscow, Russian Federation

* Corresponding author (alexandernegoda95[at]gmail.com)

Suggested: 18.05.2026; Accepted: 17.06.2026; Published: 14.07.2026

Abstract

The article examines the risks associated with the use of large language models and generative artificial intelligence (hereinafter referred to as AI) in phishing attacks. The purpose of the study is to analyze how LLMs transform phishing, what new vulnerabilities emerge as a result, and how a threat model for such attacks can be constructed. It was found that the use of LLMs changes phishing in several ways: it accelerates OSINT preparation, increases the plausibility of texts, simplifies message personalization, expands the range of delivery channels, and enhances the reproducibility of attacker actions. The authors propose an original threat model linking threat actors, assets, delivery channels, vulnerabilities, and consequences. It is concluded that generative AI should be regarded as an independent factor increasing cyber risk, while countering AI-enabled phishing should rely on the combined implementation of organizational, technical, and analytical measures.

Keywords: generative artificial intelligence, large language models, phishing, social engineering, AI threat model, AI phishing.

AI КАК ИСТОЧНИК УГРОЗ (LLM, ГЕНЕРАТИВНЫЙ ИИ). РИСКИ ИСПОЛЬЗОВАНИЯ ГЕНЕРАТИВНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ В ФИШИНГОВЫХ АТАКАХ: АНАЛИЗ И МОДЕЛЬ УГРОЗ

Научная статья

Негода А.Н.[1, *]

[1] ООО «ГПМ Цифровые Инновации», Москва, Российская Федерация

* Корреспондирующий автор (alexandernegoda95[at]gmail.com)

Предложена: 18.05.2026; Принята: 17.06.2026; Опубликована: 14.07.2026

Аннотация

В статье рассматриваются риски использования больших языковых моделей и генеративного искусственного интеллекта (далее — ИИ) в фишинговых атаках. Цель исследования состоит в анализе того, как LLM трансформируют фишинг, какие новые уязвимости при этом формируются и каким образом может быть построена модель угроз для таких атак. Установлено, что применение LLM изменяет фишинг по нескольким направлениям: ускоряет OSINT-подготовку, повышает правдоподобие текста, упрощает персонализацию сообщений, расширяет набор каналов воздействия и повышает воспроизводимость атакующих действий. Предложена авторская модель угроз, которая связывает субъектов атаки, активы, каналы доставки, уязвимости и последствия. Сделан вывод о том, что генеративный ИИ следует рассматривать как самостоятельный фактор повышения киберриска, при этом противодействие AI-фишингу должно опираться на совместную реализацию организационных, технических и аналитических мер.

Ключевые слова: генеративный искусственный интеллект, большие языковые модели, фишинг, социальная инженерия, модель угроз ИИ, AI-фишинг.

Introduction

The proliferation of large language models has transformed the nature of emerging digital threats. While early discussions of risks associated with generative AI focused primarily on adversarial attacks on models and issues of output reliability, attention has increasingly shifted toward the risk of generating malicious content using AI. From this perspective, LLMs function as a tool for preparing attack vectors, which is particularly relevant in the context of phishing, as this form of attack is inherently built around text, trust, the imitation of legitimate communication, and the exploitation of user behavioral responses. For this reason, generative AI should be considered as an independent source of new cyber threats [1]. At the same time, phishing remains one of the most widespread forms of cyberattacks, relying on fraudulent communication to obtain credentials, financial information, or access to protected resources. The contemporary academic literature emphasizes that victim vulnerability is determined by a combination of factors: outcomes are influenced by cognitive, emotional, social, and situational conditions, including time pressure, trust in the perceived authority of the sender, patterns of digital behavior, and the overall level of cybersecurity awareness. In the context of the active use of generative AI, this dependency is further amplified, as attackers gain the ability to rapidly tailor message content to a specific recipient and situation [2].

The aim of this study is to analyze the risks associated with the use of generative language models in phishing attacks and to develop a threat model that links the technological capabilities of LLMs with typical attack scenarios, victim vulnerabilities, and organizational consequences.

The research materials consisted of academic publications addressing the cybersecurity risks of generative AI, the human factor in phishing, AI-enhanced social engineering, tools for generating phishing content, methods for its detection, and security issues of systems incorporating large language models. The study is based on methods of comparative analysis, problem-oriented thematic classification, structured description of AI-enabled phishing scenarios, and threat modeling.

Main results

The emergence of generative language models is transforming phishing primarily by increasing its accessibility. Previously, the preparation of a plausible attack depended on the attacker's competence (at a minimum, language proficiency), the ability to imitate corporate communication styles, and the manual crafting of messages; however, with the widespread availability of LLMs, much of this work is delegated to the model. Existing research on offensive-side scenarios demonstrates that publicly available tools can be used to rapidly generate multiple variations of similar phishing messages, implement phishing schemes involving voice modulation, and even configure chatbots capable of conducting conversations on behalf of the attacker. Notably, generative models are involved already at the attack preparation stage, as they assist in collecting and structuring information about the target, constructing a credible pretext, generating message content, selecting an appropriate tone, and maintaining communication after the initial contact [3].

At the content level, the threat is associated with several key changes (Fig. 1).

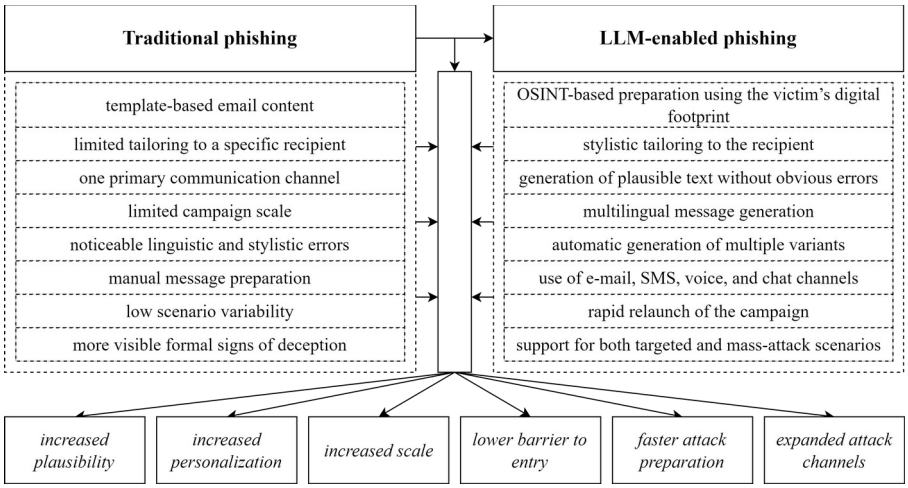


Figure 1 - The impact of generative models on the transformation of phishing and associated threats
DOI: https://doi.org/10.60797/itech.2026.11.9.1

The first change concerns OSINT preparation, in which various data sources – such as social media, corporate websites, public appearances, and other digital traces – are used to generate messages that account for the recipient's position, field of activity, current projects, and communication style. The second aspect relates to text quality, as AI-generated messages no longer reveal themselves through poor style, spelling errors, or templated phrasing. At the same time, scalability becomes possible, as attackers can generate dozens of message variations for different target groups, vary delivery channels, and quickly relaunch campaigns after previous templates are blocked.

In addition, contemporary threats are becoming increasingly multimodal; attackers may employ not only email, but also voice phishing (vishing), smishing, spoofed video calls, and chatbots that sustain the victim's engagement as the attack unfolds. In the context of AI-enhanced social engineering, the academic literature identifies three key properties that provide attackers with the highest likelihood of successful phishing: plausibility, personalization, and automation. Taken together, these factors enable an unprecedented level of influence over the user during an attack [4].

In light of this, the following classification of AI-enabled phishing can be proposed (Fig. 2).

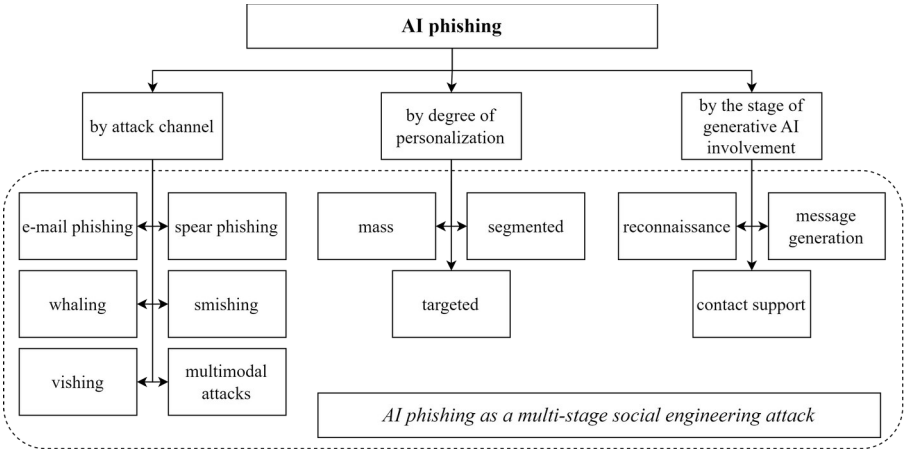


Figure 2 - Classification of attacks associated with AI-enabled phishing
DOI: https://doi.org/10.60797/itech.2026.11.9.2

In practice, the greatest risk arises from the combination of all three levels, as it turns LLMs into a tool for conducting multi-stage social engineering. At the same time, several key directions remain predominant in the academic literature on the study of AI in phishing (Table 1):

Table 1 - Comparative analysis of existing academic approaches to the study of AI in phishing

DOI: https://doi.org/10.60797/itech.2026.11.9.3

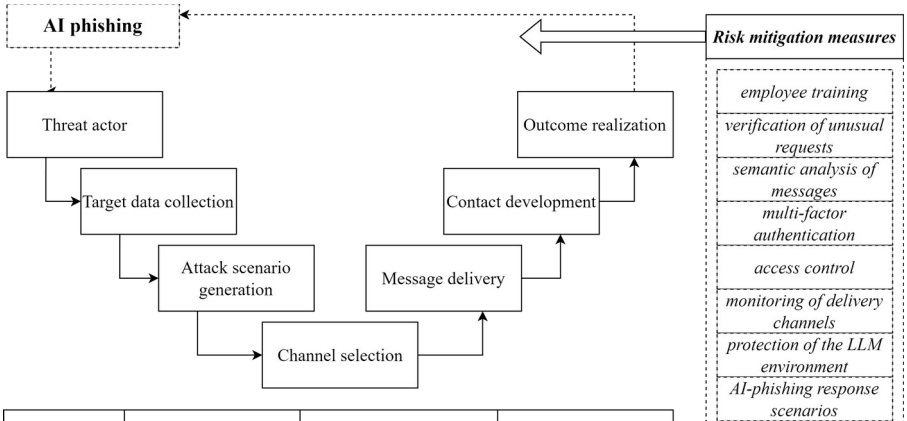
Direction	Focus of the study	Purpose	Relevance for threat modeling
Human factor	Recipient behavior and vulnerabilities	Explains the causes of attack success	Enables the identification of victim vulnerabilities
AI-enhanced social engineering	AI as a tool of persuasive influence	Demonstrates increased personalization and plausibility	Links LLMs with influence techniques
Phishing content generation	Creation of phishing messages using LLMs	Reveals offensive capabilities of AI	Describes the preparation and delivery stages
Detection of AI-enabled phishing	Methods for identifying AI-generated attacks	Forms the basis for new defensive solutions	Establishes foundations for countermeasure
Protection of LLM-based systems	Threats to LLM infrastructure and protection measures	Expands risk understanding beyond the message itself	Incorporates protection of the LLM environment into the threat model
Multi-agent detection systems	Architectures for intelligent threat detection	Enhances prospects for threat detection	Refines future directions of defensive strategies

Note: compiled by the author based on [5], [6], [7], [8]

Thus, LLMs can be used to bypass traditional filtering systems and to generate messages that account for anti-spam features already at the construction stage. In one of the most illustrative studies, the Phish-Master algorithm was proposed, combining Chain-of-Thought, MetaPrompt, and domain-specific prompts. The authors demonstrated that emails generated in this manner were successfully delivered in 99% of cases in a real network environment; accordingly, the detection model developed on their basis achieved very high performance on test data [5].

In addition, it is widely recognized that traditional defense approaches have long relied on URL-based features, email header analysis, domain anomalies, and other formal indicators. However, these methods are no longer sufficient for detecting LLM-generated phishing, as such messages can be grammatically correct, stylistically consistent, and contextually relevant while still containing malicious intent. As a result, the importance of semantic analysis and the detection of underlying persuasion patterns is increasing. In particular, the academic literature proposes a two-stage model in which the presence of persuasion principles is first assessed, followed by binary classification of the message based on this evaluation. In one study, the authors used a dataset of 2,995 AI-generated emails, achieving a detection accuracy of 94% with an AUC of 98%. Notably, this approach analyzes the message as a mechanism of psychological influence on the recipient [6].

Based on this, a threat model for the use of LLMs in phishing attacks can be proposed (Fig. 3), the construction of which is grounded in the vulnerabilities of LLM-based systems themselves when they are integrated into an organization’s business processes [7], [8].



Criterion	Traditional mass phishing	LLM-enhanced spear phishing	Multimodal AI phishing
Preparation	template-based	targeted	multi-stage
Victim data	general	OSINT profile	extended digital footprint
Text quality	medium/low	high	high and consistent
Personalization	low	high	very high
Channel	one	one or two	multiple channels
Detection	easier	more difficult	most difficult
Potential damage	moderate	high	very high

Figure 3 - Threat model for the use of LLMs in phishing attacks
DOI: https://doi.org/10.60797/itech.2026.11.9.4

In this context, the threat model of AI-enabled phishing can be represented as a sequence of interconnected elements. At the reconnaissance stage, the generative model assists in organizing OSINT data and constructing a victim profile. At the preparation stage, a pressure scenario, delivery channel, and communication style are selected. At the delivery stage, an email, message, voice fragment, or video communication script is generated. During the attack support stage, follow-up contact may occur, including clarification of details, reduction of suspicion, and guiding the victim toward the intended action. This is followed by the impact stage, involving data disclosure, financial transfers, deployment of malicious payloads, or unauthorized access to internal systems. Compared to traditional phishing, the key distinction of this scenario lies in the ability to rapidly adapt text, tone, argumentation, and persuasion tactics in response to the victim's behavior. Accordingly, based on this model, the development of defensive strategies and mechanisms becomes feasible.

Conclusion

Thus, the analysis demonstrates that the use of generative language models transforms both the nature of traditional phishing and the approaches to its execution, affecting the entire attack lifecycle. At a minimum, LLMs accelerate reconnaissance, increase the likelihood of successful message delivery, facilitate personalization, and support the scaling of attacks across multiple channels. As a result, phishing becomes less costly to prepare, more precise in targeting, and more difficult to detect. The core of the threat lies in the human factor, as well as in the limited capabilities of existing defenses and detection methods against AI-generated phishing. In this context, the most effective approach appears to be the integration of the identified threats within a unified model, followed by the development of appropriate defensive strategies and mechanisms.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Namiot D.E. On the cyber risks of generative artificial intelligence / D.E. Namiot, E.A. Ilyushin // International Journal of Open Information Technologies. — 2024. — Vol. 12, № 10. — P. 109–118.

2. Jabir R. Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors / R. Jabir, J. Le, C. Nguyen // AI. — 2025. — Vol. 6. — Art. 174.

3. Matecas A.-R. Social Engineering with AI / A.-R. Matecas, P. Kieseberg, S. Tjoa // Future Internet. — 2025. — Vol. 17. — Art. 515.

4. Gonzaga K. AI-Powered Social Engineering: Emerging Attack Vectors, Vulnerabilities, and Multi-Layered Defense Strategies / K. Gonzaga, S. Serra, M. Gomes [et al.] // Computers. — 2026. — Vol. 15. — Art. 128.

5. Han W. Phish-Master: Leveraging Large Language Models for Advanced Phishing Email Generation and Detection / W. Han, J. Zhu, C. Zhang [et al.] // Applied Sciences. — 2025. — Vol. 15. — Art. 12203.

6. Tooher P. Two-Stage Deep Learning Framework for AI-Driven Phishing Email Detection Based on Persuasion Principles / P. Tooher, H.S. Lallie // Computers. — 2025. — Vol. 14. — Art. 52.

7. Evglevskaya N.V. Ensuring the security of complex systems with the integration of large language models: analysis of threats and protection methods / N.V. Evglevskaya, A.A. Kazantsev // Economics and Quality of Communication Systems. — 2024. — № 4(34). — P. 129–144.

8. Astashkina A.A. Study of phishing detection methods using multi-agent intelligent systems / A.A. Astashkina, D.V. Bukin // Paradigm. — 2024. — № 5-5. — P. 211.