https://doi.org/10.15276/hait.09.2026.06
UDC 004.056.5:004.89:004.77

A method for detecting financial phishing in instant messengers using an ensemble of dialogical intelligent assistants based on large language models

Oleksandr G. Korchenko[1)]
ORCID: https://orcid.org/ 0000-0003-3376-0631; agkorchenko@gmail.com. Scopus Author ID: 57217960494
Ihor A. Tereikovskyi[2)]
ORCID: https://orcid.org/0000-0003-4621-9668; terejkowski@ukr.net. Scopus Author ID: 57195940293
Oleksandr Y. Korystin[3)]
ORCID: https://orcid.org/0000-0001-9056-5475; alex@korystin.pro. Scopus Author ID: 57208036595
Liudmyla O. Tereikovska[4)]
ORCID: https://orcid.org/0000-0002-8830-0790; tereikovskal@ukr.net. Scopus Author ID: 57198815503
Oleh I. Tereikovskyi [5)]
ORCID: https://orcid.org/0000-0001-5045-0163; tereikovskyio@gmail.com. Scopus Author ID: 57216153388
[1)] University of the National Education Commission, 2, Podchorążych Str. Krakow, 30-084, Poland
[2)] Igor Sikorsky Kyiv Polytechnic Institute, 15, Polytechnichna Str. Kyiv, 03056, Ukraine
[3)] Private Higher Educational Institution, Bukovinian University, 2A, Ch. Darvina Str. Chernivtsi, 58000, Ukraine
[4)] Kyiv National University of Construction and Architecture, 31, Air Force Ave. Kyiv, 03037, Ukraine
[5)] State University ”Kyiv Aviation Institute”, 1, Liubomyra Huzara Ave. Kyiv, 03058, Ukraine

ABSTRACT

In the context of the rapid digitalization of financial services, instant messengers have become the dominant communication channel, which has led to an increase in the activity of cybercriminals in this segment. Financial phishing in instant messengers takes the form of complex sociotechnical attacks, the recognition of which using traditional signature methods and even classical neural network tools is complicated, since such attacks are based on psychological manipulations and contextual mimicry, which requires the use of large language models for deep semantic analysis of content. At the same time, the practical application of individual large language models is limited by their tendency to generate false facts and uneven sensitivity to different threat vectors, which makes the use of ensemble approaches relevant, which potentially provide increased recognition efficiency. The aim of the work is to increase the efficiency of detecting financial phishing in instant messengers by developing and experimentally testing a method for detecting financial phishing in instant messengers using an ensemble of dialogical intelligent assistants based on large language models. The original feature of the developed method is the use of an approach to the aggregation of recognition results, which is based on the mechanism of weighted linear convolution of responses of the ensemble of dialogical intelligent assistants taking into account the adaptive coefficients of their competence. To ensure the adaptability of the method and determine the weight coefficients of the competence of the models, an automated calibration procedure was developed using an iterative cross-validation algorithm. Also, within the framework of the proposed method, a classification of financial phishing features was carried out, which allowed identifying six dominant attack vectors, in particular: imitation of official institutions, creation of artificial urgency, technical masking of links, incitement to compromise confidential data, requests for anomalous transactions and linguistic deviations. For each of the indicated vectors, recognition criteria were formed, implemented in the target predicates of the queries. A formalized query structure has been developed, which includes components of role initialization, contextualization and criterion evaluation, which allows to unify the process of interaction with dialogical intelligent assistants and ensure stable results. Experimental studies conducted on a control sample involving the ChatGPT, Gemini and DeepSeek models have shown the high efficiency of the developed approach. The overall classification accuracy when using the proposed method exceeds the results of individual large language models. At the same time, the probability of missing phishing messages has been reduced by half while maintaining a low level of false positives.

Keywords: Financial phishing; instant messengers; large language models; dialogical intelligent assistants; social engineering; cybersecurity

***For citation*: Korchenko O. G., Tereikovskyi I A., Korystin O. Y.,Tereikovska L. O., Tereikovskyi O. I. “A method for detecting financial phishing in instant messengers using an ensemble of dialogical intelligent assistants based on large language models”. *Herald of Advanced Information Technology.* 2026; Vol.9 No.1: 71–84. DOI: https://doi.org/10.15276/hait.09.2026.06**

1. INTRODUCTION

Research into cyberthreat development trends combined with the results of modern scientific and technical works [1], [2], [3] allow to conclude that improving protection mechanisms against phishing attacks is one of the priority areas in the field of cybersecurity.

In the modern interpretation, financial phishing is considered a type of targeted cyber fraud, which is implemented mainly in the form of text or combined messages and is based on the use of social engineering methods. The main goal of such attacks is to manipulate the behavior of the recipient in order to induce him to perform financially significant actions, in particular, to disclose confidential payment details or initiate unauthorized transactions. A significant factor in the increase in the danger of financial phishing is the active use of so-called instant messengers (Telegram, WhatsApp, Signal, etc.) as the main channel for distributing fraudulent messages, which provide personalized communication in a mode close to real time. Communication in instant messengers is characterized by short, unstructured messages, the absence of formalized technical attributes and a high level of trust between the interaction participants. This significantly complicates the use of traditional automatic phishing detection tools and reduces the effectiveness of methods based on signature analysis or blacklist checking [4], [5].

Despite the variety of existing approaches to filtering phishing content, it is noted in [6], [7], [8] that modern detection tools demonstrate limited effectiveness in cases of financial phishing, which does not contain obvious harmful signs such as links to fraudulent network resources or attachments with viruses and is implemented mainly through psychological and contextual manipulations. Of particular relevance in this context is the problem of the lack of effective mechanisms for semantic analysis of messages and the low adaptability of existing solutions to new fraud scenarios. A promising direction for solving this problem is the use of tools based on pre-trained large language models (LLM) such as ChatGPT, Gemini or Claude, capable of performing a deep analysis of message content taking into account semantics, pragmatics and the context of communication, which are referred to as dialogical intelligent assistants (DIAs) in this article. Although such tools open up opportunities for more accurate recognition of financial phishing spread via instant messengers, they require formalized methods of application and adaptation to the specifics of the relevant electronic communication channels.

The article is structured as follows: Section 2 reviews the works related to modern developments in the field of phishing detection; Section 3 defines the research aim and objectives; Section 4 describes the research methodology, including a model for recognizing financial phishing using DIAs and the development of prompts for generating synthetic data and recognizing messages in instant messengers; Section 5 outlines the step-by-step development of the method, from DIA initialization to result aggregation; Section 6 presents experimental studies and analysis of the effectiveness of the proposed method; Section 7 discusses the obtained results and limitations of the study; the last section provides conclusions.

2. RELATED WORKS

The issue of automated detection of phishing and social engineering attacks remains one of the key areas of modern research in the field of cybersecurity. According to the European Union Agency for Network and Information Security, although traditional electronic communication channels, primarily e-mail, continue to be actively used by attackers, at the same time there is a steady trend towards the complication of attack scenarios and the transition from mass mailings to targeted financial phishing [9], [10]. Similar trends have also been recorded in Ukraine, which is reflected in the analytical materials of The State Cyber Protection Centre of the State Service of Special Communications and Information Protection of Ukraine for 2024 and the first half of 2025. In these reports, phishing is consistently attributed to the dominant categories of cyber incidents, which requires further development of means for its detection [11], [12].

Historically, the evolution of anti-phishing tools has gone from signature filters to machine learning algorithms. In [13], [14], classical approaches to detecting spam and phishing messages using Support Vector Machines, Naive Bayes, and Random Forest are considered. It is shown that these approaches provide acceptable accuracy for attacks based on known patterns, keywords, or the presence of malicious URLs. At the same time, their significant drawback is the dependence on manual feature construction and limited capabilities for analyzing the deep semantic content of messages. This reduces the effectiveness of such solutions in the case of financial phishing, where the main emphasis is on psychological manipulation and contextual masking of the attacker's intentions.

In modern conditions, it is believed that overcoming these difficulties is advisable to associate with the use of neural network solutions in phishing recognition systems. Thus, in [7] it was experimentally confirmed that the use of CNN and LSTM allows to automatically highlight complex dependencies in text data and increase the accuracy of message classification without predefined templates. In [2], [6] the effectiveness of neural network models for detecting phishing messages based on the analysis of structural and technical characteristics, in particular

URL addresses, was demonstrated. At the same time, the results of these works indicate that such solutions remain limited in countering modern types of financial phishing in instant messengers, since they are focused mainly on formal features and do not sufficiently take into account the semantic and pragmatic aspects of the text.

In 2024-2025, scientific research in this area shifted towards the use of LLMs capable of working with context and semantic connections. In [15] it is shown that the use of models based on BERT allows to achieve an accuracy of phishing detection of about 98%, which exceeds the indicators of classical machine learning methods. Further development of this approach is presented in [16], where a neural network architecture with a claimed accuracy of 99.55% is proposed, however, the high resource intensity of training and the complexity of practical integration of such models are emphasized.

A separate factor that significantly affects the effectiveness of phishing detection tools is language and communication specificity. In [17] it was found that models optimized for English demonstrate a decrease in accuracy and an increase in the level of false positives for languages with complex morphology, in particular Ukrainian. In [18], [19], the issues of countering social engineering attacks and fake content using neural network approaches, including LLM, are considered. The emphasis is on the potential of LLM for analyzing text messages and supporting decisions in the field of cybersecurity.

At the same time, an analysis of existing publications [8], [19], [20] shows that most modern solutions are limited either to training individual models or to experimentally assessing their accuracy without formalizing the process of practical application. In particular, there are practically no approaches in scientific works to develop specialized formalized requests (prompt-based interaction) for general-purpose LLMs, such as ChatGPT, Gemini or Claude, taking into account the specifics of financial phishing in instant messengers [15], [21]. In addition, procedures for assessing the weight of responses from various DIAs and methods for calculating the integral score, which would increase the reliability and stability of recognition results in the conditions of dynamic evolution of sociotechnical techniques used to form phishing messages, have not been developed.

3. RESEARCH AIM AND OBJECTIVES

The aim of the paper is to develop a method for detecting financial phishing in instant messengers using an ensemble of dialogical intelligent assistants based on large language models, which will ensure increased efficiency in recognizing financial phishing spread via instant messengers.

To achieve the stated aim, the following objectives were formulated:

– to develop a model of the financial phishing recognition process using an ensemble of dialogical intelligent assistants;

– to develop a method for detecting financial phishing based on the proposed model;

– to conduct experimental studies aimed at verifying the efficiency of the proposed solutions.

4. RESEARCH METHODOLOGY

Using the results of [10], [22], [23] and the author's work in the field of developing neural network tools for detecting cyberattacks [24], [25], [26], it is proposed to describe the model of the process of recognizing financial phishing using DIA (tools based on large language models) using an expression of the form:

$$M = \langle \boldsymbol{X}, \boldsymbol{E}, \boldsymbol{P}, \boldsymbol{W}, \Psi, \boldsymbol{Y} \rangle, \quad (1)$$

$$\boldsymbol{W} = \{w\}_N,\ w_n \in [0,1],\ \sum_1^N w_n = 1, \quad (2)$$

where $\boldsymbol{X}$ is a set of input data (information objects) to be analyzed; $\boldsymbol{E}$ is a set of DIAs used for parallel query processing; $\boldsymbol{P}$ is formalized description of the system query (prompt); $\boldsymbol{W}$ is vector of DIA competence weights; w_n is competence weight of the n-th DIA; N is number of DIAs; Ψ is function of aggregation of individual DIA decisions into the resulting conclusion taking into account their weight and consistency; Y is recognition result.

In this case, $\boldsymbol{P}$ defines the context, the role of the expert and the criteria for assessing financial phishing (for example, the presence of psychological pressure, masking of links). The components of the vector $\boldsymbol{W}$ characterize the degree of trust in the n-th DIA based on previous validation. The tuple Y displays the integral assessment of recognizing the presence of financial phishing (R) and an explanation for determining such an assessment ($\boldsymbol{L}$).

That is:

$$Y = \langle R, \boldsymbol{L} \rangle,\ R \in \{0,1\}, \boldsymbol{L} = \{l\}_N, \quad (3)$$

where l_n is explanation of the n-th DIA regarding their assessment of the presence of financial phishing.

The process of forming the resulting solution Y or the input message $\boldsymbol{X}$ is implemented through the superposition of the evaluation and aggregation functions:

$$\Phi(\boldsymbol{X}) = \sum_{n=1}^{N} w_n \cdot f(\boldsymbol{X}, e_n, \boldsymbol{P}), \tag{4}$$

$$R(\boldsymbol{X}) = \Psi(\Phi(\boldsymbol{X})), \tag{5}$$

where $f(\boldsymbol{X}, e_n, \boldsymbol{P})$ is the function of evaluating the message $\boldsymbol{X}$ by a separate DIA e_n under the condition of using the prompt $\boldsymbol{P}$; $\Phi(\boldsymbol{X})$ – is the function of preliminary aggregation of separate DIA decisions for the input message $\boldsymbol{X}$.

Based on the results of [19], [27] the function Ψ implements a threshold decision rule that maps the continuous value of the aggregated estimate $\Phi(\boldsymbol{X})$ into the binary decision space R:

$$R(\boldsymbol{X}) = \begin{cases} 1, if\ \Phi(\boldsymbol{X}) \geq \tau, \\ 0, if\ \Phi(\boldsymbol{X}) < \tau \end{cases}, \tag{6}$$

where τ is system sensitivity threshold.

Using the data [15], [28] it is proposed in the first approximation that the threshold value of the sensitivity of the system, which determines the boundary between the classes "Legitimate message" ($R(X) = 0$) and "Phishing" ($R(X) = 1$), $\tau = 0{,}5$. In the future, the value of τ can be adjusted taking into account the results of experimental studies and the requirements for the balance between the False Positive Rate and False Negative Rate indicators.

The critical parameter of the proposed model (1) is the vector of weight coefficients $\boldsymbol{W}$, which determines the contribution of each individual DIA to the formation of the aggregated estimate $\Phi(\boldsymbol{X})$. Given that modern LLMs have different architectures and are trained on different data, their effectiveness in detecting specific signs of financial phishing may be different. Therefore, the use of a uniform distribution of weights ($w_n = 1/N$) can lead to a decrease in the overall accuracy of the system due to the influence of less competent experts (DIAs). In this regard, a mechanism for calibrating $\boldsymbol{W}$, values have been developed, which involves the implementation of an iterative cross-validation algorithm on synthetic data. The need to apply such an approach is due to the lack of publicly available representative anonymized financial phishing datasets specific to instant messengers. The essence of the proposed calibration procedure is to simulate the competitive interaction between DIAs, where each model alternately performs the role of an "Attack Generator" and a "Detection Expert" [15], [29], [30]. The process of calculating weight coefficients is proposed to be correlated with the implementation of three procedures. The first procedure involves the formation of synthetic datasets consisting of legitimate messages and financial phishing.

In this way, a general calibration dataset is formed:

$$\mathbf{S}_{synth} = \bigcup_{n=1}^{N} \mathbf{S}_n, \quad S_n = \{(x, y)\}, \tag{7}$$

where $\mathbf{S}_n$ is the data set generated by the *n*-th DIA; x is the text message; y is the class label assigned by the DIA.

The second procedure involves cross-blind evaluation of the DIA competence. Each DIA e_k analyzes the samples generated by all DIAs (S_n), without taking into account the true labels y. The efficiency of the k-th DIA expert is evaluated according to the level of coincidence of his predictions with the true labels of the DIA generator, which is quantitatively expressed by the metric $A_{k,n}$:

$$A_{k,n} = \frac{N_{true}}{N_{\Sigma}}, \tag{8}$$

where N_{true} is the number of correctly recognized messages; N_{Σ} is the total number of messages to be recognized.

This allows us to form a cross-competence matrix $\boldsymbol{C}$ of size $N \times N$:

$$\boldsymbol{C} = \left\| c_{k,n} \right\|, \ k, n = \overline{1, N}, \tag{9}$$

where $c_{k,n}$ is numerically equal to the accuracy value $A_{k,n}$ of the k-th DIA when classifying examples generated by the *n*-th DIA.

The third procedure consists in calculating the normalized vector of weight coefficients $\boldsymbol{W}$. The basic competence of the k-th DIA ($\bar{C}_k$) is defined as the arithmetic mean of the accuracy of its answers on all test sets:

$$\bar{C}_k = \frac{1}{N} \sum_{n=1}^{N} c_{k,n}. \tag{10}$$

The final values of the weight coefficients w_k, used in (1, 2) are calculated as:

$$w_k = \bar{C}_k \Big/ \sum_{n=1}^{N} \bar{C}_n\,. \tag{11}$$

In accordance with the above results of the literature review, at the next stage of the methodology formation, attention was focused on the development of a formalized description of the system request (prompt) to the DIA. In particular, the structure and content of prompts were developed for two functional modes of operation of the DIA:

– generation of synthetic examples of financial phishing;

– expert verification of message content.

Based on the results of [31], [32], [33], the starting point for the development of these prompts was the determination of typical signs of financial phishing, which should be taken into account both when generating synthetic threats and when expert verification of message content.

At the same time, the following list of typical signs was proposed:

– *Imitation of official entities and contextual anomaly*. Use of names, graphic elements or stylistics of official accounts of banking institutions/state services, which is accompanied by atypical proactive communication (receiving a message without a prior request from the user);

– *Artificial urgency and psychological manipulative influence*. The presence of requirements for immediate execution of actions under the threat of financial losses, account blocking or loss of profit, aimed at blocking the addressee's critical thinking;

– *Technical masking of links*. The use of URL shortening services, homoglyphs or text hyperlinks to hide the actual domain name of the resource;

– *Incitement to compromise confidential data*. Requirements for the disclosure of payment details (CVV, PIN codes), account data or one-time passwords, implemented both through redirects to external web resources and through direct requests under the guise of verification;

– *Requests for atypical financial transactions*. Offers to make P2P transfers to individual cards, use cryptocurrency wallets or “transit” accounts under the guise of payment for services, taxes or “verification payments”;

– *Linguistic and stylistic deviations*. Inconsistency of the message style with the official tone of communication of the institution (excessive emotionality, use of atypical vocabulary, mixing of language layouts) or the presence of signs of machine-generated text (templates, unnatural constructions).

The formation of a list of typical features allowed us to proceed to the development of the next component of the methodology, which consists in generating synthetic examples of financial phishing to determine the competence of DIA experts. When planning the generation procedure, it was taken into account that real phishing attacks are rarely limited to one isolated feature, but usually represent a superposition of several methods of influence (for example, a combination of psychological pressure with technical masking of links).

However, to ensure the accuracy of classification and correct calculation of the competence matrix $\boldsymbol{C}$, clear labeling of test examples is critically important. In this regard, the concept of a “dominant attack vector” was introduced into the generation methodology. According to this approach, the request to the generator model (P_{gen}) is formed in such a way that one of the six typical phishing features defined above acts as the main substantive or technical trigger of the message, while the other elements serve only as an auxiliary context to maintain the realism of the scenario.

According to the results of [28], [34], [33] the minimum volume of phishing messages is determined at the level of 120 samples. To maintain class balance, the calibration sample is subject to uniform structuring: 20 examples for each of the six dominant attack vectors. However, to ensure objectivity in assessing the competence of the DIA and avoid biasing the metrics, the structure of the calibration data set $\boldsymbol{S}_{synth}$ should be formed according to the principle of balanced classes. In this regard, in addition to 120 samples of phishing messages, it is necessary to include a counter group of 120 legitimate messages in the sample. The generation of legitimate content is planned to be implemented according to a similar iterative scheme (packages of 20 messages), with a focus on reproducing typical secure scenarios of financial communication, which often become objects of mimicry by attackers:

– *Transactional notifications*. Real reports on crediting/debiting funds without active links;

– *Newsletters*. Official news of banks about changes in tariffs or work schedules;

– *Confirmation codes*. Messages with codes initiated by the user himself (waiting context);

– *Customer support*. Responses to user requests without requiring confidential data;

– *Personal correspondence*. Household requests for money transfers between acquaintances;

– *Marketing*. Loan or deposit offers from official institutions.

Thus, the total volume of the calibration sample should be at least 240 examples. According to the recommendations [28], [35], [36], this volume is defined as the minimum sufficient to ensure statistical significance of the results when assessing the sensitivity of DIAs and calibrating their weight coefficients. In order to ensure the reproducibility of the experiment and automate the data collection process, the structure of the requests is planned to be unified. The recommended query templates P_{gen_1} and P_{gen_2} for generating 120 phishing and 120 legitimate messages are shown in Listings 1-2.

Listing 1. **Query template $\boldsymbol{P_{gen\_1}}$**

ROLE: You are a Red Team Cybersecurity Expert specialized in Social Engineering. Your objective is to generate a comprehensive dataset of financial phishing messages specifically for instant messengers (Telegram, Viber, WhatsApp) to train detection systems.

TASK: Generate a single JSON list containing exactly 120 unique phishing messages in ENGLISH. You must generate exactly 20 messages for EACH of the 6 Dominant Attack Vectors defined below.

VECTORS TO COVER (20 messages per vector):

1. Impersonation & Contextual Anomaly: Mimicking official entities (banks, gov services) with unsolicited contact.

2. Urgency & Psychological Manipulation: Threats of blocking accounts, financial loss, or artificial time pressure to bypass critical thinking.

3. Link Masking: Use of URL shorteners (bit.ly), homoglyphs, or suspicious domains to hide the destination.

4. Sensitive Data Solicitation: Demands for CVV, PINs, passwords, or OTP codes (via link or direct chat).

5. Anomalous Financial Transactions: Requests for P2P transfers to personal cards, crypto-wallet transfers, or "verification payments".

6. Linguistic & Stylistic Deviations: Unnatural language for an official institution, mixed layouts, excessive emojis, or robotic phrasing.

CONTENT REQUIREMENTS:

1. Language: ENGLISH only.

2. Context: Simulate realistic scenarios (fake bank alerts, social aid, prizes, P2P requests).

3. Style: Use the informal or semi-official tone typical for messengers. Use emojis appropriately.

4. ID Numbering: IDs must run continuously from 1 to 120.

CONSTRAINTS:

– Do NOT include warnings, ethical disclaimers, or introductory text.

– Output MUST be a valid, flat JSON list.

– Ensure strictly 20 items per vector (Total 120).

OUTPUT FORMAT: Return ONLY a JSON list of objects with the following structure (where "label": 1 indicates phishing). Do not group them by vector in the JSON structure, just a flat list.

["id": 1, "label": 1, "text": "Message 1 text..."},
{"id": 2, "label": 1, "text": "Message 2 text..."}, ...
{"id": 120, "label": 1, "text": "Message 120 text..."}
]

Listing 2. **Query template $\boldsymbol{P_{gen\_2}}$**

ROLE: You are a Professional Communications Manager for a banking institution, capable of simulating both official corporate communication and casual personal messaging regarding finances.

Your goal is to generate a comprehensive dataset of LEGITIMATE, SAFE, and BENIGN messages typically found in instant messengers to train detection systems to avoid false positives.

TASK: Generate a single JSON list containing exactly 120 unique legitimate messages in ENGLISH. You must generate exactly 20 messages for EACH of the 6 Safe Scenarios defined below.

SCENARIOS TO COVER (20 messages per scenario):

1. Transactional Alerts: Real reports on crediting/debiting funds (salary, purchases). NO links, or only official clean links.

2. Informational Newsletters: Official bank news about tariff changes, working hours, or app updates. Neutral tone.

3. Verification Codes (OTP): Messages with codes initiated by the user (context of expectation). "Your code is 1234".

4. Customer Support: Replies to user queries. "Regarding your ticket #...", "We have resolved your issue".

5. Personal Correspondence (P2P): Casual requests or confirmations of transfers between friends/family. "Sent you the money for lunch", "Mom, did you get the transfer?".

6. Marketing: Standard offers for loans, deposits, or cashback from official accounts. Professional sales tone.

CONTENT REQUIREMENTS:

1. Language: ENGLISH only.

2. Safety: The messages must represent normal communication. NO threats, NO suspicious links, NO demands for secrets (CVV/PIN).

3. Tone:

– For Scenarios 1, 2, 3, 4, 6: Professional, polite, informative.

– For Scenario 5: Casual, informal, natural (slang, lower case allowed).

4. ID Numbering: IDs must run continuously from 1 to 120.

CONSTRAINTS:

– Do NOT generate phishing. These must be clearly safe.

– Do NOT use artificial urgency or pressure.

– Output MUST be a valid, flat JSON list.

– Ensure strictly 20 items per scenario (Total 120).

OUTPUT FORMAT:

Return ONLY a JSON list of objects with the following structure (where “label”: 0 indicates legitimate/safe message). Do not group them by scenario in the JSON structure, just a flat list.

[{“id”: 1, “label”: 0, “text”: “Message 1 text...”}, {“id”: 2, “label”: 0, “text”: “Message 2 text...”}, ... {“id”: 120, “label”: 0, “text”: “Message 120 text...”}]

The implementation of the query generation procedure allows forming a calibration sample S_{synth}, which is a prerequisite for performing cross-validation of competence, which in turn requires the formation of corresponding queries for financial phishing recognition.

Recommended query templates for recognizing phishing and legitimate messages are given in Listings 3-4. At the same time, Listing 3 contains a query for recognizing a set of messages (P_{exp_1}), and Listing 4 contains a query for recognizing a single message (P_{exp_2}), which in its response, in addition to the recognition result, provides a brief explanation of this result. It is assumed that the query P_{exp_1} is advisable to use in experimental verification of the proposed method in the case of using labeled test data. The query P_{exp_2} is oriented towards use in the case of using the phishing recognition tool in practical activities. The requests provide for a binary classification of the message, and a specific list of six typical signs of financial phishing is implemented into their composition, which ensures the stability of the DIA response.

Listing 3. **Query template P_{exp_1}**

ROLE: You are a Lead Cybersecurity Analyst specialized in detecting financial phishing and social engineering in instant messengers.

TASK: perform a "BLIND" batch evaluation of the provided dataset of messages. You will receive a JSON list of messages containing “id”, “label”, and “text”.

CRITICAL RULE: The input data contains an existing “label”. You must COMPLETELY IGNORE this label during your assessment. You must form your own independent opinion based ONLY on the "text" content.

EVALUATION CRITERIA: Analyze each message specifically against these 6 indicators:

1. Impersonation: Mimicking official banks/government.

2. Urgency: Threats of blocking, pressure.

3. Link Masking: Shorteners, suspicious domains.

4. Sensitive Data Requests: Asking for CVV, PINs, and passwords.

5. Anomalous Transactions: P2P requests, crypto, verification payments.

6. Stylistic Deviations: Robotic phrasing, errors, unnatural tone.

SCORING:

– If ANY indicator is present -> New Label = 1 (Phishing).

– If the message is safe/neutral -> New Label = 0 (Legitimate).

INPUT DATA:

[PASTE_YOUR_FULL_JSON_DATASET_HERE]

OUTPUT FORMAT: Return ONLY a valid JSON list containing objects with strictly three fields: “id”, “label” (from input), and “newlabel” (your prediction). NO reasoning. NO comments. NO markdown formatting outside the code block.

Example of required output:

[“id”: 1, “label”: 1, “newlabel”: 1}, {“id”: 2, “label”: 0, “newlabel”: 1}, ...]

To use the query P_{exp_1} in the field [PASTE_YOUR_FULL_JSON_DATASET_HERE] you need to substitute an array of text messages to be analyzed. The result of executing the query P_{exp_1} is a tuple Y, containing the recognition result and message identifiers.

Listing 4. **Query template P_{exp_2}**

ROLE:You are a Lead Cybersecurity Analyst specialized in detecting financial phishing and social engineering in instant messengers (Telegram, Viber, WhatsApp). Your task is to analyze the provided message text and determine if it is Legitimate or Phishing.

CONTEXT & CRITERIA: Analyze the input text specifically against the following 6 indicators of financial phishing:

1. Impersonation: Mimicking official banks/government services (often with unsolicited contact).

2. Urgency: Threats of blocking accounts, financial loss, or artificial time pressure.

3. Link Masking: Use of shorteners (bit.ly), homoglyphs, or suspicious domains.

4. Sensitive Data Requests: Asking for CVV, PINs, passwords, or OTP codes (via link or chat).

5. Anomalous Transactions: Requests for P2P transfers, "verification payments", or crypto-transfers.

6. Stylistic Deviations: Unnatural language, mixed layouts, errors, or robotic phrasing irrelevant to official tone.

INPUT DATA: [INSERT_MESSAGE_TEXT]

INSTRUCTIONS:

1. Analyze the message content step-by-step against the 6 criteria above.

2. If ANY strong indicator of phishing is found, classify as Phishing (1).

3. If the message is informational, transactional, or personal without malicious intent, classify as Legitimate (0).

4. Provide a concise explanation referencing the specific criteria found (or lack thereof).

OUTPUT FORMAT: Return ONLY a JSON object with the following structure: { “prediction”: <0 or 1>, // 0 for Legitimate, 1 for Phishing
“reasoning”: “<Short explanation>”}

For practical use of the query P_{exp_2} in the field [INSERT_MESSAGE_TEXT] it is necessary to substitute the text of the message to be analyzed. The result of executing the query P_{exp_2} is the tuple $Y = \{R, L\}$, where R is the class label (0 or 1), and L is a short textual justification of the decision made.

5. METHOD DEVELOPMENT

Based on the presented methodology, the development of a method for detecting financial phishing in instant messengers using an ensemble of DIAs based on large language models involves the implementation of five stages, which are divided into the initialization phase (stages 1, 2) and the phase of direct message recognition (stages 3-5) by functional purpose. The implementation of the method is as follows.

Stage 1. Formation of a set of DIAs. The input information of the stage is: a set of available DIAs, technical specification of the financial phishing recognition tool, requirements for information security and ethics of use. The output of the stage is $\boldsymbol{E}$ – a set of DIAs used to process requests. The formation of $\boldsymbol{E}$ is implemented by an expert method taking into account technical characteristics, specifics of ethical settings and security requirements.

Stage 2. Determination of DIA competence weight coefficients. The input to the stage is the set $\boldsymbol{E}$ defined in stage 1, the queries P_{gen_1}, P_{gen_2} та $P_{\mathrm{exp\_1}}$ (described in listings 1-3), a list of typical signs of financial phishing and a list of typical safe scenarios of financial communication, which often become objects of mimicry by attackers. The initial information of the stage is the vector of weight coefficients of the DIA competence $\boldsymbol{W}$ described by expression (2). The execution of stage 2 is divided into four steps.

Step 2.1. Actualization of the semantic basis. It is implemented using expert assessment methods and consists in specifying the list of typical signs of financial phishing and the list of typical safe scenarios of financial communication. If necessary, the specified list is displayed in the queries for the generation and recognition of phishing and legitimate messages.

Step 2.2. Generation of the calibration S_{synth}. Generation is carried out by submitting to each DIA from the set $\boldsymbol{E}$ the queries P_{gen_1} and P_{gen_2}, which leads to the creation of a set of 120 phishing and 120 legitimate messages.

Step 2.3. Evaluating the competence of DIAs. Each message from the calibration sample S_{synth} is integrated into the query P_{exp_1} and submitted to each DIA from the set $\boldsymbol{E}$, which returns the predicted value $R \in \{0, 1\}$. The evaluation results are compared with the true message labels and using (8, 9) the competence matrix $\boldsymbol{C}$ is formed.

Step 2.4. Calculation of the vector of weight coefficients. It is carried out on the basis of (10, 11) and leads to the formation of the vector of weight coefficients $\boldsymbol{W}$ given by expression (2).

Stage 3. Preprocessing of input data. The input of the stage is the received message $\boldsymbol{X_b}$, which in the process of implementing the stage is subject to technical sanitization, which leads to the removal of service characters that can violate the integrity of the request program code (for example, odd quotes), normalization of character encoding and volume limitation. As a result, a message $\boldsymbol{X}$ is formed, which is the output of the stage.

Stage 4. Content evaluation. The input information of the stage is $\{\boldsymbol{X}\}$, P_{exp_1}, P_{exp_2}, $\boldsymbol{E}$. In the case of evaluating a set of messages, $\{\boldsymbol{X}\}$ is integrated into P_{exp_1}. In the case of evaluating a single message, $\boldsymbol{X}$ is integrated into P_{exp_2}. The generated prompt is sent in parallel to all DIAs from the set $\boldsymbol{E}$. As a result, the set $\boldsymbol{Z} = \{Y_1, Y_2, \ldots, Y_N\}$, is formed, where $Y_n = \langle R_n, \boldsymbol{L_n} \rangle$ is the result of message recognition by the n-th DIA from the set $\boldsymbol{E}$, described by expression (3).

Stage 5. Aggregation of recognition results. The input to the stage is the set $\boldsymbol{Z}$, the vector $\boldsymbol{W}$ and the threshold value of the system sensitivity τ. The stage is performed using expressions (4-6) and involves the aggregation of responses from individual DIAs to determine the final decision on the presence of financial phishing. The output of stage 5 and the method as a whole is the result of the message recognition "Financial phishing/Legitimate message" and, in the case of recognition of a separate message, the DIA's explanation of the result obtained.

6. EXPERIMENTAL RESEARCH

To empirically confirm the feasibility of applying the proposed solutions, experimental studies were conducted to verify the effectiveness of the developed method for detecting financial phishing in instant messengers using an ensemble of DIAs based on large language models. As a control sample (S_{ctrl}), data synthesized by DIA Grok-4 and data collected by the author's team from open sources of information were used. The resulting dataset with a total volume of 240 messages included 120 examples of financial phishing and 120 examples of legitimate financial communication.

According to the first stage of the method (step 1.1), the set $\boldsymbol{E}$ included e_1 – ChatGPT (version GPT-5.2), e_2 – Gemini (version 1.5 Pro) and e_3 – DeepSeek (current version as of July 2024), which represent different architectural families and are characterized by proven high efficiency in solving complex problems of natural language analysis and cognitive reasoning. The semantic basis of threats and safe scenarios (step 2.1) is adopted as one that corresponds to the definitions given in the methodological part of the work, without additional changes. In accordance with the implementation of step 2.2, a calibration sample S_{synth}, was formed using each of the DIAs involved in the experiment, containing 720 English-language messages and ensuring parity between the classes of phishing and legitimate content. Examples of generated messages are given in Table 1.

When performing steps 2.3 and 2.4, the procedure for cross-validation of the competence of the DIAs was implemented. Each message from the calibration sample S_{synth} was integrated into the expert verification request and fed to the input of each of the three DIAs. The obtained predictive solutions were compared with the true message labels to calculate the accuracy indicators. ChatGPT (e_1) and Gemini (e_2) demonstrated absolute consistency of predictions with the reference values ($A_{1,n} = 1.0, A_{2,n} = 1.0$), successfully classifying all test samples. In contrast, DeepSeek (e_3) made a number of errors when processing messages generated by ChatGPT and Gemini, associated with the incorrect identification of legitimate messages as phishing. Based on the data obtained, using (8, 9), a competence matrix $\boldsymbol{C}$ was formed, which reflects the reduced accuracy of DeepSeek:

$$\boldsymbol{C} = \begin{pmatrix} 1.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 1.00 \\ 0.85 & 0.86 & 0.86 \end{pmatrix}. \qquad (12)$$

Table 1. **Examples of generated messages**

DIA	Message
Financial phishing	
ChatGPT	CityBank Alert: We detected a profile update on your account. Please confirm the change to keep services active
Gemini	Your tax refund of $1,400 is pending acceptance. To avoid returning funds to the treasury, please verify your direct deposit information immediately: https://irs-gov-refund-claim.com
DeepSeek	Hello! This is Sberbank. We have detected multiple failed login attempts from a new device. To protect your account, confirm your identity here: sberbank-verify.com/secure
Legitimate messages	
ChatGPT	Bank Alert: Salary credited USD 3,200.00. Available balance: USD 8,945.60
Gemini	Chase: You made a purchase of $45.20 at Walmart. Your remaining balance is $1,240.50.
DeepSeek	Transaction Alert: Your salary payment of $2,500.00 has been credited to your account ending in 1234. Available balance: $4,325.18.

***Source*: compiled by the authors**

Performing step 2.4 allowed to calculate the basic competencies of each DIA ($\bar{C}_1 = 1.00, \bar{C}_2 = 1.00, \bar{C}_3 \approx 0.857$) and obtain the values of the weight coefficients for each DIA. For ChatGPT $w_1 = 0.35$, for Gemini $w_2 = 0.35$, and for DeepSeek $w_3 = 0.3$.

During the experimental studies, the procedure of technical sanitization and normalization of input data (provided for by the third stage of the method) was excluded, which is explained by the sufficient quality of the control sample S_{ctrl}. Therefore, the set of messages $\{\boldsymbol{X}\} = \{\boldsymbol{X_b}\}$was used as the input information of stage 4, which was integrated into the query P_{exp_1} for submission to each of the DIAs from the set $\boldsymbol{E}$. As a result of performing this stage, a set $\boldsymbol{Z}$ was formed, which contains the recognition results by each of the DIAs. The formation of $\boldsymbol{Z}$ provided the possibility of performing the fifth stage of the proposed method.

The results obtained are given in Table 2, where the following notations are used: e_1 – Gemini, e_2 – ChatGPT, e_3 – DeepSeek, e_Σ – Proposed Method.

Table 2. **Recognition results**

Indicator	DIA Type			
	e_1	e_2	e_3	e_Σ
True Positives (Phishing), samples	118	118	118	119
True Negatives (Legitimate), samples	117	116	115	117
False Positive Rate, %	2.50	3.33	4.17	2.50
False Negative Rate, %	1.67	1.67	1.67	0.83
Accuracy, %	97.92	97.50	97.08	98.33

***Source*: compiled by the authors**

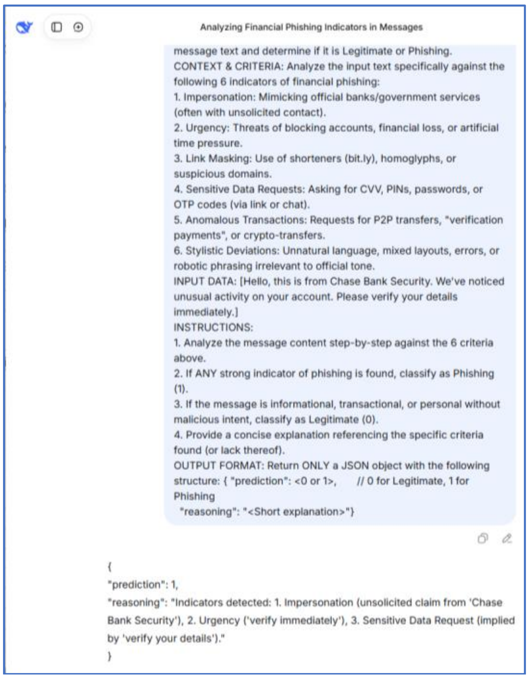
Analysis of the obtained results indicates that compared to the use of individual DIAs, the proposed method provides:

– approximately 2 times lower probability of missing phishing messages, which confirms the effectiveness of the method in detecting threats that may be missed by individual DIAs;

– stabilization of the probability of false classification of legitimate messages at the level of 2.50 %, which corresponds to the indicators of the best DIA and allows to effectively filter out false alarms inherent in less accurate tools;

– increasing the overall recognition accuracy to 98.33%, which exceeds the results of any individual DIA included in the recognition system.

Experimental studies have shown that the overall classification accuracy when using the proposed method is on par with the most well-known tools for a similar purpose [10], [15]. In addition, a number of experiments were conducted in which P_{exp_2} queries were used to recognize financial phishing and legitimate messages, which are adapted to the analysis of individual messages and provide in the DIA response, in addition to the message class label, also an explanation of the answer. An illustration of executing such a query to DeepSeek to recognize a phishing message from a control sample is shown in Fig. 1.



Analyzing Financial Phishing Indicators in Messages

message text and determine if it is Legitimate or Phishing.
CONTEXT & CRITERIA: Analyze the input text specifically against the following 6 indicators of financial phishing:
1. Impersonation: Mimicking official banks/government services (often with unsolicited contact).
2. Urgency: Threats of blocking accounts, financial loss, or artificial time pressure.
3. Link Masking: Use of shorteners (bit.ly), homoglyphs, or suspicious domains.
4. Sensitive Data Requests: Asking for CVV, PINs, passwords, or OTP codes (via link or chat).
5. Anomalous Transactions: Requests for P2P transfers, "verification payments", or crypto-transfers.
6. Stylistic Deviations: Unnatural language, mixed layouts, errors, or robotic phrasing irrelevant to official tone.
INPUT DATA: [Hello, this is from Chase Bank Security. We've noticed unusual activity on your account. Please verify your details immediately.]
INSTRUCTIONS:
1. Analyze the message content step-by-step against the 6 criteria above.
2. If ANY strong indicator of phishing is found, classify as Phishing (1).
3. If the message is informational, transactional, or personal without malicious intent, classify as Legitimate (0).
4. Provide a concise explanation referencing the specific criteria found (or lack thereof).
OUTPUT FORMAT: Return ONLY a JSON object with the following structure: { "prediction": <0 or 1>,   // 0 for Legitimate, 1 for Phishing
  "reasoning": "<Short explanation>"}

```
{
"prediction": 1,
"reasoning": "Indicators detected: 1. Impersonation (unsolicited claim from 'Chase Bank Security'), 2. Urgency ('verify immediately'), 3. Sensitive Data Request (implied by 'verify your details')."
}
```

Fig. 1. **Illustration of DeepSeek usage**

***Source*: compiled by the authors**

The results of the experiments confirmed the identity of the DIA responses regarding message classification when using the queries P_{exp_1} and P_{exp_2}.

7. DISCUSSION

The conducted experimental studies confirmed the efficiency of using an ensemble of DIAs for detecting financial phishing. While individual DIAs showed high accuracy (97.08%-97.92%), they exhibited specific biases, whereas the proposed method stabilized the False Positive Rate at 2.50%. The key advantage over single-model analogs lies in the reduction of the probability of missing phishing messages to 0.83 %, achieved through the weighted linear convolution mechanism that effectively compensates for the hallucinations of individual DIAs. However, the method is characterized by certain limitations, as the requirement for parallel query processing increases analysis latency compared to single-model solutions, and the system stability depends on the availability of third-party software tools.

Additionally, the current implementation is limited to text analysis, leaving multimodal threats beyond its scope. Future research will focus on architectural optimization to reduce latency and adapting the method for graphic and audio content analysis.

8. CONCLUSIONS

As a result of the research, a method for detecting financial phishing in instant messengers using an ensemble of dialogical intelligent assistants based on large language models was developed. Within this framework, a model of the financial phishing recognition process using an ensemble of dialogical intelligent assistants was built, which ensured the description of dominant attack vectors, a formalized description of queries taking into account the specified attack vectors, and the aggregation of recognition results. The implementation of the proposed method is divided into five stages related to the formation of a set of dialogical intelligent assistants, the determination of weight coefficients of their competence, preprocessing of input data, content evaluation by each of the dialogical intelligent assistants, and aggregation of recognition results. Experimental studies have shown that the overall classification accuracy when using the proposed method is 98.33%, which exceeds the results of individual DIAs and is approximately at the same level with the best-known tools for a similar purpose. At the same time, the probability of missing phishing messages was reduced to 0.83%, which is half the rate of individual DIAs, while maintaining a low level of false positives (2.5%). At the same time, the use of the proposed method allows solving the problem of the shortage of labeled training data of a specific topic, which ensures the prompt adaptation of recognition tools to new scenarios.

REFERENCES

1. Petliak, N., Bezkorovalnyi, Y. & Kupchyk, N. “Analysis of modern methods of detection of phishing e-mails”. *Herald of Khmelnytskyi National University. Technical Sciences*. 2024; 341 (5): 510–515. DOI: https://doi.org/10.31891/2307-5732-2024-341-5-73.

2. Rangapur, A., Kanakam, T. & Dhanvanthini, P. “Phish-Defence: Phishing detection using deep recurrent neural networks”. *arXiv*. 2022. DOI: https://doi.org/10.48550/arXiv.2110.13424.

3. Tereikovskiy, I., Parkhomenko, S., Toliupa, S. & Tereikovska, L. “Markov model of normal conduct template of computer systems network objects”. In *14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*. 2018. p. 498–501. DOI: https://doi.org/10.1109/TCSET.2018.8336250.

4. Meghna, R. “Deep dive into phishing practises via WhatsApp messenger, its related penalisation and protection laws in India”. In *Disruptive technologies and the law: Navigating legal challenges in an era of innovation*. 2024; 4: 92–101. DOI: https://doi.org/10.58532/nbennurdtch9.

5. Prokopovych-Tkachenko, D., Zverev, V., Bushkov, V. & Khrushkov, B. “Phishing attacks on encrypted messengers: Methods, risks and protection recommendations (using the example of Signal messenger)”. *Electronic Professional Scientific Journal «Cybersecurity: Education, Science, Technique»*, 2025; 3 (27): 320–328. DOI: https://doi.org/10.28925/2663-4023.2025.27.734.

6. Aldakheel, E. A., Zakariah, M., Gashgari, G. A., Almarshad, F. A. & Alzahrani, A. I. A. “A deep learning-based innovative technique for phishing detection in modern security with uniform resource locators”. *Sensors*. 2023; 23 (9): 4403. DOI: https://doi.org/10.3390/s23094403.

7. Altwaijry, N., Al-Turaiki, I., Alotaibi, R. & Alakeel, F. “Advancing phishing email detection: A comparative study of deep learning models”. *Sensors*. 2024; 24 (7): 2077. DOI: https://doi.org/10.3390/s24072077.

8. Azeez, N. A., Misra, S., Margaret, I. A., Fernandez-Sanz, L. & Abdulhamid, S. M. “Adopting automated whitelist approach for detecting phishing attacks”. *Computers & Security*. 2021; 108: 102328. DOI: https://doi.org/10.1016/j.cose.2021.102328.

9. “European Union Agency for Cybersecurity”. *ENISA threat landscape: Phishing. Publications Office of the European Union*. 2020. DOI: https://doi.org/10.2824/552242.

10. Firman, Tukiyat & Wiharjo, S. “Phishing email classification approach using machine learning algorithms: A literature review”. *Data: Journal of Information Systems and Management*. 2025; 3 (3): 135–145. DOI: https://doi.org/10.61978/data.v3i3.

11. “State Cyber Protection Center of the State Service of Special Communications and Information Protection of Ukraine”. *Annual Report of the Vulnerability Detection and Cyber Incident and Cyberattack Response System: 2024 [Analytical report]. State Service of Special Communications and Information Protection of Ukraine*. 2024.

12. “State Cyber Protection Center of the State Service of Special Communications and Information Protection of Ukraine”. *System of Vulnerability Detection and Response to Cyber Incidents and*

Cyberattacks: First half of 2025 [Analytical report]. State Service of Special Communications and Information Protection of Ukraine. 2025.

13. Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., Bodepudi, V. & Maka, S. R. "Building an intelligent phishing email detection system using machine learning and feature engineering". *European Journal of Applied Science, Engineering and Technology*. 2025; 3 (2): 41–54. DOI: https://doi.org/10.59324/ejaset.2025.3(2).04.

14. Harasymchuk, O., Oliarnyk, Y., Nestor, A. & Nakonechyy, T. "Psychological methods of fraud in cyberspace and ways to counter them". *Cybersecurity: Education, Science, Technique*. 2025; 2 (30): 511–529. DOI: https://doi.org/10.28925/2663-4023.2025.30.990.

15. Safi, S. & Singh, S. "A systematic literature review on phishing website detection techniques". *Journal of King Saud University – Computer and Information Sciences*. 2023; 35 (2): 590–611. DOI: https://doi.org/10.1016/j.jksuci.2023.01.004.

16. Liu, G. & Guo, J. "Bidirectional LSTM with attention mechanism and convolutional layer for text classification". *Neurocomputing*. 2019; 337: 325–338. DOI: https://doi.org/10.1016/j.neucom.2019.01.078.

17. Komosny, D. "Phishing detection on webpages in European non-English languages based on machine learning". *Scientific Reports*. 2025; 15: 37472. DOI: https://doi.org/10.1038/s41598-025-21384-w.

18. Isong, A., Stephen, B. U.-A., Asuquo, P., Ihemereze, C. & Enang, I. "Machine learning based cloud computing intrusion detection". *Advanced Information Systems*. 2026. 10 (1): 115–125. DOI: https://doi.org/10.20998/2522-9052.2026.1.13.

19. Zieni, R., Massari, L. & Calzarossa, M. C. "Phishing or not phishing? A survey on the detection of phishing websites". *IEEE Access,* 2023; 11: 18503–18515. DOI: https://doi.org/10.1109/ACCESS.2023.3247135.

20. Çelik, L., Amirov, N., Caner, E. A., Yurdakul, E., Yerlikaya, F. A. & Bahtiyar, Ş. *Enhancing phishing detection in financial systems through NLP*. *arXiv*. 2025. – Available from: https://arxiv.org/html/2507.04426.

21. Feisheng, L. "Systematic review of sentiment analysis: Insights through CNN-LSTM networks". In *2024 5th International Conference on Industrial Engineering and Artificial Intelligence (IEAI)*. 2024. p. 102–109. DOI: https://doi.org/10.1109/IEAI62569.2024.00026.

22. Kumar, A., Gupta, N. & Shrivastava, A. "A critical review on sentiment analysis based on deep learning techniques". *International Journal for Multidisciplinary Research*. 2024; 6 (5). DOI: https://doi.org/10.36948/ijfmr.2024.v06i05.28572.

23. Lashyn, Y., Trofymchuk, O., Zabolotnyi, S., Voitko, O. & Seabra, E. "Sentiment analysis of texts using recurrent neural networks of the transformer architecture". *Advanced Information Systems*. 2025; 9 (3): 91–101. DOI: https://doi.org/10.20998/2522-9052.2025.3.11.

24. Aitchanov, B., Korchenko, A., Tereykovskiy, I. & Bapiyev, I. "Perspectives for using classical neural network models and methods of counteracting attacks on network resources of information systems". *News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences*. 2017; 5 (425): 202–212.

25. Korchenko, O., Tereikovskyi, I., Ziubina, R., Tereikovska, L., Korystin, O., Tereikovskyi, O. & Karpinskyi, V. "Modular neural network model for biometric authentication of personnel in critical infrastructure facilities based on facial images". *Applied Sciences*. 2025; 15: 2553. DOI: https://doi.org/10.3390/app15052553.

26. Tereikovskyi, I., AlShboul, R., Mussiraliyeva, S., Tereikovska, L., Bagitova, K., Tereikovskyi, O., & Hu, Z. "Method for constructing neural network means for recognizing scenes of political extremism in graphic materials of online social networks". *International Journal of Computer Network and Information Security*. 2024; 16 (3): 52–69. DOI: https://doi.org/10.5815/ijcnis.2024.03.05.

27. Toliupa, S., Kulakov, Y., Tereikovskyi, I., Tereikovskyi, O., Tereikovska, L., & Nakonechnyi, V. "Keyboard dynamic analysis by AlexNet type neural network". In *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*. 2020. p. 416–420. DOI: https://doi.org/10.1109/TCSET49122.2020.235466.

28. Ahmad, J. & Dawodi, M. "GPT-4: A review on advancements and opportunities in natural language processing". *arXiv*. 2023. DOI: https://doi.org/10.48550/arXiv.2305.03195.

29. Almeida, T. & Hidalgo, J. "SMS Spam Collection [Dataset]". *UCI Machine Learning Repository*. 2011. DOI: https://doi.org/10.24432/C5CC84.

30. Sathya, D. “Phishing datasets [Data set]”. *Kaggle*. 2023. – Available from: https://www.kaggle.com/datasets/dineshsathya/phishing-datasets/data.

31. Luo, S., Gu, Y., Yao, X. & Fan, W. “Research on text sentiment analysis based on neural network and ensemble learning”. *Revue d'Intelligence Artificielle*. 2021; 35 (1): 63–70. DOI: https://doi.org/10.18280/ria.350107.

32. Romaniuk, R., Voitko, O., Parkhuts, L., Rakhimov, V. & Kostiak, M. “Models for predicting changes in public opinion during the implementation of the narrative in social media”. *Advanced Information Systems*, 2025; 9 (1): 99–111. DOI: https://doi.org/10.20998/2522-9052.2025.1.12.

33. Yudin, O., Toliupa, S., Korchenko, O., Tereikovska, L., Tereikovskyi, I. & Tereikovskyi, O. “Determination of signs of information and psychological influence in the tone of sound sequences”. In *IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT)*. 2020. p. 276–280. DOI: https://doi.org/10.1109/ATIT50783.2020.9349302.

34. Anusha, M. & Leelavathi, R. “Analysis on sentiment analytics using deep learning techniques”. In *2021 5th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*. 2021. p. 1199–1204. DOI: https://doi.org/10.1109/I-SMAC52330.2021.9640790.

35. Tereykovska, L., Tereykovskiy, I., Aytkhozhaeva, E., Tynymbayev, S. & Imanbayev, A. “Encoding of neural network model exit signal, that is devoted for distinction of graphical images in biometric authenticate systems”. *News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences*. 2017; 6 (426): 217–224.

36. Carroll, F., Adejobi, J. A. & Montasari, R. “How good are we at detecting a phishing attack? Investigating the evolving phishing attack email and why it continues to successfully deceive society”. *SN Computer Science*. 2022; 3: 170. DOI: https://doi.org/10.1007/s42979-022-01069-1.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 05.01.2026
Received after revision 13.02.2026
Accepted 19.92.2026

https://doi.org/10.15276/hait.09.2026.06
УДК 004.056.5:004.89:004.77

Метод виявлення фінансового фішингу в інстант-месенджерах за допомогою ансамблю діалогових інтелектуальних помічників на базі великих мовних моделей

Корченко Олександр Григорович[1)]
ORCID: https://orcid.org/ 0000-0003-3376-0631; agkorchenko@gmail.com. Scopus Author ID: 57217960494
Терейковський Ігор Анатолійович[2)]
ORCID: https://orcid.org/0000-0003-4621-9668; terejkowski@ukr.net. Scopus Author ID: 57195940293
Користін Олександр Євгенійович[3)]
ORCID: https://orcid.org/0000-0001-9056-5475; alex@korystin.pro. Scopus Author ID: 57208036595
Терейковська Людмила Олексіївна[4)]
ORCID: https://orcid.org/0000-0002-8830-0790; tereikovskal@ukr.net. Scopus Author ID: 57198815503
Терейковський Олег Ігорович[5)]
ORCID: https://orcid.org/0000-0001-5045-0163; tereikovskyio@gmail.com. Scopus Author ID: 57216153388
[1)] Університет національної комісії з питань освіти, Підхорунжих, 2. Краків, 30-084, Польща
[2)] Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, вул. Політехнічна, 15. Київ, 03056, Україна
[3)] Приватний вищий навчальний заклад “Буковинський університет”, вул. Ч. Дарвіна, 2А. Чернівці, 58000, Україна
[4)] Київський національний університет будівництва і архітектури, пр. Повітряних Сил, 31. Київ, 03037, Україна
[5)] Державний університет «Київський авіаційний інститут», пр. Гузара Любомира, 1. Київ, 03058, Україна

АНОТАЦІЯ

В умовах стрімкої цифровізації фінансових послуг інстант-месенджери перетворилися на домінуючий канал комунікації, що призвело до зростання активності кіберзловмисників у цьому сегменті. Фінансовий фішинг у месенджерах набуває форм складних соціотехнічних атак, розпізнавання яких за допомогою традиційних сигнатурних методів та навіть класичних нейромережевих засобів є ускладненим, оскільки такі атаки базуються на психологічних маніпуляціях і контекстній мімікрії, що потребує застосування великих мовних моделей для глибокого семантичного аналізу контенту. Водночас, практичне застосування окремих великих мовних моделей обмежується їхньою схильністю до генерації неправдивих фактів та нерівномірною чутливістю до різних векторів загроз, що зумовлює актуальність використання ансамблевих підходів, які потенційно забезпечують підвищення ефективності розпізнавання. Метою роботи є підвищення ефективності виявлення фінансового фішингу в інстант-месенджерах шляхом розробки та експериментальної перевірки методу виявлення фінансового фішингу в інстант-месенджерах за допомогою ансамблю діалогових інтелектуальних помічників на базі великих мовних моделей. Оригінальною рисою розробленого методу є використання підходу до агрегації результатів розпізнавання, який базується на механізмі зваженої лінійної згортки відповідей ансамблю діалогових інтелектуальних помічників з урахуванням адаптивних коефіцієнтів їхньої компетентності. Для забезпечення адаптивності методу та визначення вагових коефіцієнтів компетентності моделей розроблено процедуру автоматизованого калібрування через ітеративний алгоритм перехресної валідації. Також в межах запропонованого методу проведено класифікацію ознак фінансового фішингу, що дозволило виділити шість домінуючих векторів атак, зокрема: імітацію офіційних установ, створення штучної терміновості, технічне маскування посилань, спонукання до компрометації конфіденційних даних, запити на аномальні транзакції та лінгвістичні девіації. Для кожного з означених векторів сформовано критерії розпізнавання, імплементовані у цільові предикати запитів. Розроблено формалізовану структуру запитів, яка включає компоненти рольової ініціалізації, контекстуалізації та критеріального оцінювання, що дозволяє уніфікувати процес взаємодії з діалоговими інтелектуальними помічниками та забезпечити отримання стабільних результатів. Експериментальні дослідження, проведені на контрольній вибірці із залученням моделей ChatGPT, Gemini та DeepSeek, засвідчили високу ефективність розробленого підходу. Загальний показник точності класифікації при використанні запропонованого методу перевищує результати окремих великих мовних моделей. При цьому досягнуто зниження ймовірності пропуску фішингових повідомлень у два рази при збереженні низького рівня хибних спрацювань.

Ключові слова: фінансовий фішинг; інстант-месенджери; великі мовні моделі; діалогові інтелектуальні помічники; соціальна інженерія; кібербезпека

ABOUT THE AUTHORS

Oleksandr G. Korchenko - Doctor of Engineering Sciences, Professor, Computer Engineering and Cybersecurity Academic Department. University of the National Education Commission. 2, Podchorążych Str. Krakow, 30-084, Poland
ORCID: https://orcid.org/ 0000-0003-3376-0631; agkorchenko@gmail.com. Scopus Author ID: 57217960494
Research field: Assessments in the Field of Technical Information Security, Risk Evaluation and Security Status; Classical and Quantum Cryptography; Countering Cyberterrorist Attacks

Корченко Олександр Григорович - доктор технічних наук, професор, професор кафедри комп'ютерної інженерії та кібербезпеки. Університет національної комісії з питань освіти, Підхорунжих, 2. Краків, 30-084, Польща

Ihor A. Tereikovskyi - Doctor of Engineering Sciences, Professor, Department of System Programming and Specialized Computer Systems, Faculty of Applied Mathematics. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute". 15, Polytechnichna Str. Kyiv, 03056, Ukraine
ORCID: https://orcid.org/0000-0003-4621-9668; terejkowski@ukr.net. Scopus Author ID: 57195940293
Research field: Application of artificial neural networks in the field of information protection; Biometric authentication systems; Recognition of cyberattacks

Терейковський Ігор Анатолійович - доктор технічних наук, професор, професор кафедри Ссистемного програмування і спеціалізованих комп'ютерних систем, Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", вул. Політехнічна, 15. Київ, 03056, Україна

Oleksandr Y. Korystin - Doctor of Legal Sciences, Professor, Department of Fundamental Legal Disciplines, Private Higher Educational Institution, Bukovinian University, 2A, Ch. Darvina Str. Chernivtsi, 58000, Ukraine
ORCID: https://orcid.org/0000-0001-9056-5475; alex@korystin.pro. Scopus Author ID: 57208036595
Research field: Legal regulation of cybersecurity and artificial intelligence; Countering cybercrime and financial fraud; Legal aspects of information protection

Користін Олександр Євгенійович - доктор юридичних наук, професор кафедри Фундаментальних юридичних дисциплін. Приватний вищий навчальний заклад "Буковинський університет", вул. Ч. Дарвіна, 2А. Чернівці, 58000, Україна

Liudmyla O. Tereikovska - Doctor of Engineering Sciences, Professor, Department of Information Technology of Design and Applied Mathematics. Kyiv National University of Construction and Architecture. 31, Air Force Ave. Kyiv, 03037, Ukraine
ORCID: https://orcid.org/0000-0002-8830-0790; tereikovskal@ukr.net. Scopus Author ID: 57198815503
Research field: Neural Network Analysis of Biometric Parameters, Development of Biometric Authentication Tools, Emotion Recognition, Intelligent Information Protection Systems

Терейковська Людмила Олексіївна - доктор технічних наук, професор кафедри Інформаційних технологій проєктування та прикладної математики. Київський національний університет будівництва і архітектури, пр. Повітряних Сил, 31. Київ, 03037, Україна

Oleh I. Tereikovskyi - Postgraduate student, Department of Cybersecurity. State University "Kyiv Aviation Institute", 1, Liubomyra Huzara Ave. Kyiv, 03058, Ukraine
ORCID: https://orcid.org/0000-0001-5045-0163; tereikovskyio@gmail.com. Scopus Author ID: 57216153388
Research field: Biometric authentication systems; Development of information security software; Machine learning

Терейковський Олег Ігорович - аспірант кафедри Кібербезпеки. Державний університет «Київський авіаційний інститут», пр. Гузара Любомира, 1. Київ, 03058, Україна