DOI: https://doi.org/10.15276/hait.09.2026.42
UDC 004.056.5:004.89:004.77

A method for detecting financial phishing in email based on a judicial interaction model of intelligent agents

Liudmyla O. Tereikovska[1)]
ORCID: https://orcid.org/0000-0002-8830-0790; tereikovskal@gmail.com. Scopus Author ID: 57198815503
Iurii V. Krak[2)]
ORCID: https://orcid.org/ 0000-0002-8043-0785; yuri.krak@gmail.com. Scopus Author ID: 6602577533
Anna O. Korchenko[3)]
ORCID: https://orcid.org/ 0000-0003-0016-1966; anna.korchenko@uken.krakow.pl. Scopus Author ID: 56029291400
Olena I. Balina[1)]
ORCID: https://orcid.org/ 0000-0003-4621-9668; elena.i.balina@gmail.com. Scopus Author ID: 57223105709
Oleh I.Tereikovskyi[2)]
ORCID: https://orcid.org/0000-0001-5045-0163; tereikovskyio@gmail.com. Scopus Author ID: 57216153388
[1)] Kyiv National University of Construction and Architecture, 31, Air Force Ave. Kyiv, 03037, Ukraine
[2)] Taras Shevchenko National University of Kyiv, 60, Volodymyrska Str. Kyiv, 01033, Ukraine
[3)] University of the National Education Commission, 2, Podchorążych Str. Krakow, 30-084, Poland

ABSTRACT

Relevance: stems from the need to improve the detection of financial phishing in email, in particular when fraudulent messages use sophisticated social engineering techniques and contain no obvious technical indicators of compromise. Traditional filtering tools and solutions based on individual large language models may demonstrate insufficient effectiveness in detecting such phishing because they typically evaluate the message text in isolation and do not compare different interpretations of its content. The aim of the article is to develop a method for detecting financial phishing in email based on a model of judicial interaction among intelligent agents, which is intended to improve the effectiveness of its recognition. **Objectives:** to develop a model of a multi-agent environment of judicial interaction among intelligent agents with assessments represented as fuzzy numbers; to develop a method for detecting financial phishing in email that implements the proposed model as sequential stages of analyzing the incoming message and aggregating the fuzzy assessments of the agents; and to evaluate the effectiveness of the method experimentally. **Methods:** the study applies a multi-agent model of judicial interaction among intelligent agents based on large language models, together with fuzzy logic. Individual verdicts are expressed as linguistic assessments mapped to triangular fuzzy numbers and aggregated using competence weighting coefficients. **Scientific novelty** lies in adapting the adversarial procedure of judicial proceedings to the organization of interaction among intelligent agents, which involves the formation of opposing reasoned positions of the prosecution and the defense, their evaluation by a panel of judges, and the aggregation of the judges' verdicts based on fuzzy logic, thereby making it possible to improve the accuracy of financial phishing recognition. **Practical significance:** the proposed method can be used in email protection tools where accuracy takes priority over resource consumption, providing a textual justification of the message classification decision. **Results:** experimental studies on a corpus formed by the authors, using Claude, ChatGPT, Gemini, and DeepSeek, showed a lower classification error rate for the proposed method than for individual large language models and for the weighted integration method without judicial interaction, under the same experimental protocol. **Conclusions:** the proposed method makes it possible to improve the accuracy of financial phishing recognition in email on the basis of judicial interaction among intelligent agents implemented using large language models, without additional training of these models.

Keywords: Financial phishing; email; large language models; intelligent agents; social engineering; cybersecurity; multi-agent system; fuzzy logic; judicial interaction

***For citation*: Tereikovska L. O., Krak Iu. V., Korchenko A. O., Balina O. I., Tereikovskyi O. I. "A method for detecting financial phishing in email based on a judicial interaction model of intelligent agents". *Herald of Advanced Information Technology.* 2026; Vol.9 No.4: 683–699. DOI: https://doi.org/10.15276/hait.09.2026.42**

1. INTRODUCTION

Analysis of current trends in the field of cyber threats, supported by the results of scientific and technical research [1], [2], [3], indicates that the development and improvement of methods for countering phishing attacks remains one of the critically important tasks of maintaining information security.

In the modern scientific paradigm, financial phishing is classified as a destructive form of targeted cyber fraud, which is executed through text or multimedia content and relies on social engineering tools [4], [5]. The key destructive impact of such attacks is aimed at deforming the behavioral patterns of the recipients in order to induce them to perform legally or financially harmful actions, in particular, unauthorized transfer of funds or compromise of authentication and payment data.

© Tereikovska L., Krak Iu., Korchenko A., Balina O., Tereikovskyi O., 2026

This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/deed.uk)

The effective solution to this scientific and practical problem is directly correlated with the international priorities of critical infrastructure protection and national cybersecurity doctrines worldwide [5], [6]. This vector is also a cornerstone in ensuring the stability of the digital space of Ukraine, since countering cognitive threats and social engineering is defined as one of the strategic priorities of the state in the context of hybrid security challenges [7], [8].

Despite the rapid development of alternative means of communication, e-mail remains the primary and most critical channel for the distribution of complex fraudulent messages in the corporate and private sectors. The specificity of e-mail communication lies in detailed, structurally complex texts with high semantic capacity. This allows attackers to carefully mask phishing markers in a mass of legitimate business or personal correspondence, significantly reducing the effectiveness of traditional protection tools based on signature analysis, heuristic rules, or the use of reputational blacklists. In particular, as noted in [3], [9], [10], such inbound traffic filtering systems demonstrate critically low performance when detecting latent financial phishing. Such content is characterized by the absence of obvious technical indicators of compromise (malicious attachments or direct links to destructive web resources) and is based exclusively on multi-level psychological and contextual manipulations. The main reasons for the limited effectiveness of existing solutions are the lack of flexible tools for deep semantic analysis and the limited adaptability of algorithms to the dynamic transformation of fraudulent scenarios.

Considering the capabilities of large language models described in [10], [11], [12], a promising vector for overcoming the outlined scientific and practical problem is the integration into the security contour of tools based on pre-trained large language models. Given their ability to interpret text information taking into account complex semantic-pragmatic connections and contextual features, such models can effectively function as intelligent agents (IA). At the same time, the deployment of such technologies for email monitoring actualizes the need to create formalized methods for their structural organization and interaction, optimized for recognizing complex cognitive threat influences.

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 the proposed judicial multi-agent interaction model, mathematical representation of agent information transformations based on fuzzy logic, and specialized query templates for the intelligent agents; Section 5 presents the proposed financial phishing detection method, including its analytical model and the sequence of information transformations performed through the interaction of intelligent agents; Section 6 presents the experimental evaluation of the proposed method; Section 7 discusses the experimental results, study limitations, and future research directions; the final section concludes the paper.

2. RELATED WORKS

Currently, a wide range of tools based on various approaches to processing and analyzing message data are used to detect financial phishing in e-mail. In particular, a significant group of existing solutions is based on classical intelligent classifiers, such as Support Vector Machines, Naive Bayes, and Random Forest [13], [14]. Despite the ability to effectively identify threats in the presence of explicit markers of compromise (known patterns, keywords, or malicious links), these algorithms are limited by the inevitable need for manual identification of signs and the inability to interpret hidden contextual content, which is most often the basis of modern psychological manipulations by attackers.

In order to overcome these limitations, neural network approaches are increasingly used in the practice of security analysis. Studies prove that the use of CNN and LSTM architectures allows for automatic extraction of complex dependencies in text arrays without the involvement of rigidly fixed rules [2], while the integration of neural network models for estimating structural parameters and URLs [15], [16], [17] demonstrates a high detection capability at the technical level. However, the orientation of such solutions mainly towards the formal and syntactic characteristics of the email message significantly narrows their suitability for detecting complex scenarios of financial phishing, where the pragmatic and semantic load of the text plays a decisive role.

A separate area of modern research in the field of phishing detection is related to the use of large language models designed to analyze contextual relationships in the text. The use of BERT-based architectures allows achieving an accuracy of phishing content recognition of about 98 % [18], and the development of specialized neural network configurations raises this figure to 99.55 % [19], although at the cost of significant computational resources and the complication of infrastructure

integration of systems. The influence of linguistic specificity also acts as a significant destructive factor, since models developed for the English-speaking environment lose their detection ability and increase the level of false positives when analyzing languages with complex morphology [20], [21]. At the same time, the potential of LLM in the tasks of countering social engineering and fake content [22], [23] confirms the feasibility of their development not only as linguistic analysis tools, but also as decision-making support tools in the general cybersecurity system. Furthermore, an analysis of existing publications [24], [25] shows that most modern solutions are limited either to training individual isolated models or to experimentally assessing their accuracy without proper formalization of the processes of practical application. Partial overcoming of the outlined shortcomings was presented in [4], which proposed a method for detecting financial phishing in instant messaging systems based on the ensemble use of dialogical intelligent assistants. In the mentioned study, a classification of financial phishing features was formed according to six dominant attack vectors, a three-component structure of queries was developed, which includes role initialization, contextualization and criterion evaluation, and a mechanism for automated determination of model competence weights for further aggregation of results was implemented. Despite the high classification accuracy at the level of 98.33% and a significant reduction in the probability of missing a threat, the approach proposed in [4] is focused mainly on linear message evaluation and operates with clear numerical values during the generalization of verdicts. This limits its effectiveness in the case of transferring analytical procedures to long text streams of e-mail, where the manipulative influence is multifaceted and requires a comparison of opposing interpretations.

A critical review of scientific works indicates the prospects for the use of IA based on large language models for detecting financial phishing in e-mail. However, known solutions mostly implement linear evaluation of the text of an e-mail, considering it as a static array of features, which does not provide a comparison of opposite interpretations of the same message, reducing their ability to detect hidden social engineering manipulations in the absence of obvious technical indicators of compromise. At the same time, according to the results of [27], [28], [29], the organization of the IA interaction procedure according to the type of judicial (adversarial) interaction enables an increase in the efficiency of the final decision due to the simultaneous consideration of the arguments of the prosecution and defense. In addition, the analyzed works lack a formalized description of methods that would transform the results of the adversarial analysis of several IA into a specific decision on the presence of financial phishing and determine the sequence of stages of such analysis, starting from the formation of the argumentation and ending with the adoption of the final verdict. Moreover, the existing approaches to the aggregation of IA assessments mostly operate with clear numerical values, which does not allow to adequately reflect the uncertainty inherent in expert judgment when comparing opposing arguments, and therefore, justifies the expediency of presenting agent assessments in the form of fuzzy numbers and developing a mechanism for their generalization using the theory of fuzzy logic. Furthermore, the question of the expediency of implementing the IA roles of judicial interaction by one large language model or several independent models remains open, which, by analogy with [4], [27] requires a separate justification. The solution of these issues determined the goal and objectives of this study.

3. RESEARCH AIM AND OBJECTIVES

The aim of the work is to develop a method for detecting financial phishing in e-mail based on a model of forensic interaction of intelligent agents, which will ensure increased recognition efficiency.

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

– develop a model of a multi-agent environment for detecting financial phishing based on the forensic interaction of intelligent agents built on the basis of large language models, with the presentation of evaluation results in the form of fuzzy numbers;

– develop a method for detecting financial phishing in e-mail that implements the proposed model of judicial interaction of intelligent agents in the form of a sequence of stages of analyzing the incoming message and generalizing the fuzzy assessments of agents using the theory of fuzzy logic to form a final decision;

– conduct experimental studies aimed at evaluating the effectiveness of the proposed method.

4. RESEARCH METHODOLOGY

The conceptual basis of the study relies on the adaptation of the principle of adversarial litigation in order to organize IA interaction in a multi-agent environment [28], [29].

According to this principle, the prosecution and defense form reasoned positions on the analyzed electronic message, where the defense responds to the accusation, and the panel of judges independently evaluates the arguments of both parties. Transferring this principle to the problem of detecting financial phishing made it possible to develop a model of a multi-agent environment in which the interaction of the set of IA $\boldsymbol{U}$ is organized by analogy with the judicial process, and the functions of forming arguments, evaluating them, and making a final decision are distributed between agents according to the roles they perform.

At the same time:

$$\boldsymbol{U} = \{U_A, U_D, \boldsymbol{U_J}\}, \quad (1)$$

$$\boldsymbol{U_J} = \{U_1, U_2, U_3\}, \quad (2)$$

where U_A, U_D are Accusation Agent, Defense Agent; $\boldsymbol{U_J}$ **is** Judicial Panel Agents; U_1, U_2, U_3 are Judges Agents.

The preliminary formation of structured arguments by the prosecution and defense allows the panel of judges to evaluate already organized information, which reduces the cognitive load on the agents responsible for the evaluation and lowers the risk of losing relevant context when analyzing large amounts of text data. The result of the work of each agent-judge is an individual linguistic assessment, which is subsequently used for mathematical aggregation and the formation of a final decision on whether the message belongs to the financial phishing class.

Within the proposed model, the Accusation Agent (U_A) analyzes an incoming electronic message from the position of the prosecution in order to detect signs of financial phishing. The agent detects sociotechnical markers, psychological manipulations, signs of artificially creating a sense of urgency or authority, as well as incentives to perform potentially dangerous actions. The result of the agent's operation is a structured conclusion of the prosecution, which contains a reasoned justification for the presence of signs of financial phishing and links to the relevant fragments of the analyzed message.

Defense Agent (U_D) performs analysis of an electronic message from the position of the defense party, using the results of the work of the Prosecutor Agent. The agent conducts a search for arguments confirming the legitimacy of the message, the detection of signs of standard business or personal communication, as well as the formation of structured counterarguments to the position of the prosecution party. The result of the agent's work is a structured conclusion of the defense party, which contains a reasoned justification of the legitimacy of the message and a refutation or clarification of the arguments of the prosecution party.

The Judicial Panel ($\boldsymbol{U_J}$) consists of three independent Judge Agents, each performing the role of a judge. The panel analyzes the formed judicial "case" containing the text of the electronic message, a structured conclusion of the prosecution and a structured conclusion of the defense. Each judge agent independently assesses the weight of the arguments of both sides and forms a response that contains a linguistic assessment of the threat level of the analyzed message, the corresponding triangular fuzzy number and a textual explanation that justifies the decision made. The obtained linguistic assessments, the corresponding triangular fuzzy numbers and textual explanations are used at the following stages to aggregate the results of the judicial assessment, form an integral indicator of the phishing threat and make a final decision on whether the message belongs to the financial phishing class.

The information transformations performed by the specified IA are described as follows:

$$A = f_A(E), \quad (3)$$

$$D = f_D(E, A), \quad (4)$$

$$\langle L_i, M_i, S_i \rangle = f_{J_i}(E, A, D), \quad (5)$$

where E is incoming electronic message adapted for analysis using IA; A is structured conclusion of the Accusation Agent; D is structured conclusion of the Defense Agent; L_i – linguistic assessment formed by the i-th Judge Agent ($i = \overline{1,3}$), selected from a predefined scale of linguistic terms; M_i is triangular fuzzy number corresponding to the linguistic assessment of L_i, received from the *i*-th judge agent; S_i is textual explanations regarding L_i; $f_A(\cdot)$ is function for forming a reasoned conclusion of the Accusation Agent; $f_D(\cdot)$ is function for forming a reasoned conclusion of the Defense Agent; $f_{J_i}(\cdot)$ is function for independently evaluating the arguments of the parties of the *i*-th Judge Agent.

According to the results of [30], the mathematical support of the procedure for transforming and aggregating the formed linguistic estimates L_i or calculating the deterministic indicator of the degree of belonging of the message to the financial phishing class Y is described by expressions (6-11).

In this case, the term set $\boldsymbol{T}$, from which the estimate L_i is selected, is described by the expression:

$$\boldsymbol{T} = \{T_1, T_2, T_3, T_4, T_5\}, \quad (6)$$

where T_1 is very low level of the financial phishing threat; T_2 is low; T_3 is medium; T_4 is high; T_5 is very high.

The procedure for fuzzifying the conclusions of the panel of judges is described by the expressions:

$$M_i = \varphi(L_i) = (a_i, b_i, c_i), \quad (7)$$

$$\mu_{M_i}(x) = \begin{cases} 0, x < a_i \\ \dfrac{x - a_i}{b_i - a_i}, a_i \le x \le b_i \\ \dfrac{c_i - x}{c_i - b_i}, b_i \le x \le c_i \\ 0, x > c_i \end{cases}, \quad (8)$$

where φ is mapping operator; $\mu_{M_i}(x) \in [0,1]$ is membership function, which determines the degree of correspondence of the numerical value x on the segment $X = [0,100]$ to the fuzzy number M_i; b_i is modal value of the fuzzy number M_i; a_i, c_i are respectively, the left and right tolerance limits of the fuzzy number M_i under the condition $a_i \le b_i \le c_i, a_i \le c_i$.

For the boundary terms T_1 and T_5, the branches of (8) with zero denominators are omitted, and the membership function on [0, 100] is defined by the remaining branches, so that it equals 1 at $x = 0$ for T_1 and at $x = 100$ for T_5.

The integration of individual fuzzy estimates of the panel of judges into the resulting triangular fuzzy number $M_\Sigma = (a_\Sigma, b_\Sigma, c_\Sigma)$ is described by the expressions:

$$M_\Sigma = \left(\sum_{i=1}^{3} w_i a_i , \sum_{i=1}^{3} w_i b_i , \sum_{i=1}^{3} w_i c_i \right), \quad (9)$$

$$\sum_{i=1}^{3} w_i = 1, \quad (10)$$

where w_i is the weight coefficient of the competence of the i-th judge.

The calculation of the indicator of the degree of belonging of the message to the financial phishing class $Y \in [0,100]$ is described by the expression:

$$Y = \frac{a_\Sigma + b_\Sigma + c_\Sigma}{3}, \quad (11)$$

To make a final decision on the security status of the message, the calculated indicator Y s compared with the predefined threshold levels δ_1 and δ_2 $(\delta_1 < \delta_2)$ which is described by the expression:

$$R = \begin{cases} Allow, Y < \delta_1 \\ Quarantine, \delta_1 \le Y \le \delta_2, \\ Block, Y > \delta_2 \end{cases} \quad (12)$$

where R is control decision of the security system; $Allow$ is decision to deliver the message to the addressee without restrictions (legitimate letter); $Quarantine$ is decision to send the message to an isolated environment for additional verification (suspicious message); $Block$ is decision to unconditionally block the message (financial phishing); δ_1, δ_2 are threshold values determined by the information system security policy.

The modal values of the terms $T_1 -, T_5$ are evenly spaced with a step of 25 on the interval [0,100], and adjacent membership functions overlap so that their sum equals one at any point [30]. It should be noted that the choice of a triangular profile of membership functions is due to the need to minimize computational complexity when scaling the system to large message flows. The choice of the defuzzification method by the center of mass allows for a smooth change in the output indicator for any variations in the input linguistic estimates. Since aggregation (9) is linear, Y equals the weighted mean of the term centroids (8.33, 25, 50, 75 and 91.67), so Y and R can be reproduced directly from the verdicts of the judges. The introduction of weighting coefficients provides the ability to take into account the difference in architectural complexity and detection ability of the basic large language models on which specific judge agents are built. It is advisable to determine the specific values of the specified coefficients using the results of evaluating the effectiveness of the models obtained in [4]. The choice of the interval [0,100] simplifies the further integration of the obtained indicator into standard security monitoring systems and is convenient for interpreting the results.

Based on the approach proposed in [4], the functioning of IA, described by expressions (3-8), is determined by a set of specialized query templates:

$$\boldsymbol{Q} = \{Q_A, Q_D, Q_J\}, \quad (13)$$

where Q_A, Q_D, Q_J are the query templates for the Accusation Agent, Defense Agent, and Judge Agent, respectively.

Query templates define functional roles, informational constraints, analysis logic, and format for presenting IA results. To form a reasoned position of the prosecution, a Q_A query template to the Agent-Prosecutor, shown in Listing 1, was developed. The query template defines the logic of message analysis, orienting IA to the detection and factual documentation of signs of social engineering, psychological manipulation, and technical anomalies according to the six criteria defined in [4], with the formation of a structured report, which is

subsequently used by the Agent-Lawyer and the Board of Agent-Judges.

Listing 1. ***Query template*** Q_A

ROLE. You act as the Accusation Agent in a judicial multi-agent framework for financial phishing detection.

TASK. Construct an evidence-based accusation supporting the hypothesis that the analyzed email is a financial phishing attempt.

Analyze the message according to the following criteria: (1) brand impersonation; (2) urgency and psychological pressure; (3) suspicious or obfuscated hyperlinks; (4) requests for confidential information; (5) suspicious financial operations; (6) stylistic anomalies. Do not provide arguments supporting the legitimacy of the message.

OUTPUT REQUIREMENTS. Return only a valid JSON object. Do not include any explanations outside the JSON object. All textual values in the JSON object must be written in English.

OUTPUT FORMAT.

{“evidence_strength”: “Weak | Moderate | Strong”, “main_accusation”: “Concise statement describing the primary phishing hypothesis”,

“criteria_analysis”: { “brand_impersonation”: { “detected”: true, “evidence”: “...” },

“urgency”: {“detected”: false, “evidence”: null},

“suspicious_hyperlinks”: {“detected”: true, “evidence”: “...” },

“confidential_data_requests”: {“detected”: false, “evidence”: null },

“financial_operations”: {“detected”: false, “evidence”: null },

“stylistic_anomalies”: {“detected”: true, “evidence”: “...” }}}

INPUT MESSAGE: […]

For a critical analysis of the allegations and the formation of a reasoned position of the defense, the Q_D request template for the Agent-Advocate, shown in Listing 2, has been developed. The request template defines the logic of the message analysis, orienting the IA towards evaluating only those arguments of the prosecution that were formulated by the Agent-Prosecutor, identifying signs of the legitimacy of the message and forming evidentiary counterarguments according to the appropriate criteria. The result of the IA's work is a structured report that is used by the Panel of Agent-Judges when forming individual assessments of agent-judges.

Listing 2. ***Query template*** Q_D

ROLE. You act as the Defense Agent in a judicial multi-agent framework for financial phishing detection.

TASK. Construct an evidence-based rebuttal to the accusation provided by the Accusation Agent. Use the original email message as supporting evidence.

Analyze the accusation according to the following criteria: (1) countering brand impersonation claims; (2) urgency justification; (3) hyperlink legitimacy; (4) justification for confidential data requests; (5) legitimacy of financial operations; (6) stylistic naturalness. Respond only to the accusations presented by the Accusation Agent. Do not generate rebuttals for criteria that are not identified in the accusation. Do not provide arguments supporting the phishing hypothesis.

OUTPUT REQUIREMENTS. Return only a valid JSON object. Do not include any explanations outside the JSON object. All textual values in the JSON object must be written in English.

OUTPUT FORMAT

{“rebuttal_strength”: “Weak | Moderate | Strong”, “main_rebuttal”: “Concise statement summarizing the defense against the accusation”, “criteria_rebuttal”: { “brand_impersonation”: {“rebutted”: true, “rebuttal”: “...”},

“urgency”: {“rebutted”: false, “rebuttal”: null},

“suspicious_hyperlinks”: {“rebutted”: true, “rebuttal”: “...”},

“confidential_data_requests”: {“rebutted”: false, “rebuttal”: null},

“financial_operations”: {“rebutted”: false, “rebuttal”: null},

“stylistic_anomalies”: {“rebutted”: true, “rebuttal”: “...”}}}

PROSECUTOR ACCUSATION: […]

INPUT MESSAGE: […]

For independent evaluation of the arguments of the prosecution and defense, the Q_J query template to the Agent-Judge, shown in Listing 3, has been developed. The query template defines the rules of judicial evaluation, guiding the IA to impartially weigh the arguments of the prosecution and defense, using the incoming message only to verify the presented arguments, without generating new evidence and without performing independent analysis of the message. The result of the IA's work is a structured report that contains an individual linguistic assessment of the threat level, its ordinal

index, the corresponding triangular fuzzy number, and a concise reasoning for the decision made.

Listing 3. ***Query template*** Q_J.

ROLE. You act as a Judge Agent in a judicial multi-agent framework for financial phishing detection.

TASK. Act as an impartial judge. Evaluate the accusation provided by the Accusation Agent and the rebuttal provided by the Defense Agent. Use the original email message only to verify and assess the arguments presented by both parties. Do not introduce new evidence or perform an independent phishing analysis.

Objectively weigh the evidence and rebuttals to determine the final linguistic threat level of the analyzed email.

Select exactly one of the following linguistic threat levels and its corresponding triangular fuzzy number defined on the interval [0,100]:

"Very Low" (Index: 1, Fuzzy: a=0, b=0, c=25) – completely safe, legitimate communication.

"Low" (Index: 2, Fuzzy: a=0, b=25, c=50) – minor anomalies, likely safe.

"Medium" (Index: 3, Fuzzy: a=25, b=50, c=75) – suspicious elements requiring caution.

"High" (Index: 4, Fuzzy: a=50, b=75, c=100) – strong indicators of financial phishing.

"Very High" (Index: 5, Fuzzy: a=75, b=100, c=100) – obvious and critical financial phishing attempt.

Make your decision independently without assuming agreement with other Judge Agents.

OUTPUT REQUIREMENTS. Return only a valid JSON object. Do not include any explanations outside the JSON object. All textual values in the JSON object must be written in English.

OUTPUT FORMAT

{"final_threat_level": "Very Low | Low | Medium | High | Very High",
"term_index": 1 | 2 | 3 | 4 | 5,
"fuzzy_parameters": {"a": 0, "b": 0, "c": 0},
"judicial_reasoning": "Concise justification balancing the accusation and the rebuttal".}

PROSECUTOR ACCUSATION: [...]

DEFENSE REBUTTAL: [...]

INPUT MESSAGE: [...]

5. METHOD DEVELOPMENT

Based on the presented methodology, taking into account the results of the development of neural network information protection tools [31], [32], [33], by analogy with the method described in [3] for detecting financial phishing in instant messengers, the transformation of information implemented by the proposed method for detecting financial phishing in e-mail is generally described by the expressions:

$$\langle Y, R, G\rangle = \Phi(E, \boldsymbol{Q}, \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{W}, \Delta), \quad (14)$$

$$\Delta = \{\delta_1, \delta_2\}, \quad (15)$$

where E is the input email message; $\boldsymbol{Q}$ is the set of query templates; $\boldsymbol{T}$ is the set of linguistic terms; $\boldsymbol{U}$ is the set of intelligent agents; Y is the deterministic indicator of the degree of membership of the message to the financial phishing class; R is the output decision; G is the justification of the decision; $\boldsymbol{W}$ is the set of weighting factors that determine the competence of each judge; Φ is the information transformation operator implementing the proposed method. Information transformations defined by the operator Φ and implemented by the proposed method in the process of IA interaction are illustrated using Fig. 1.

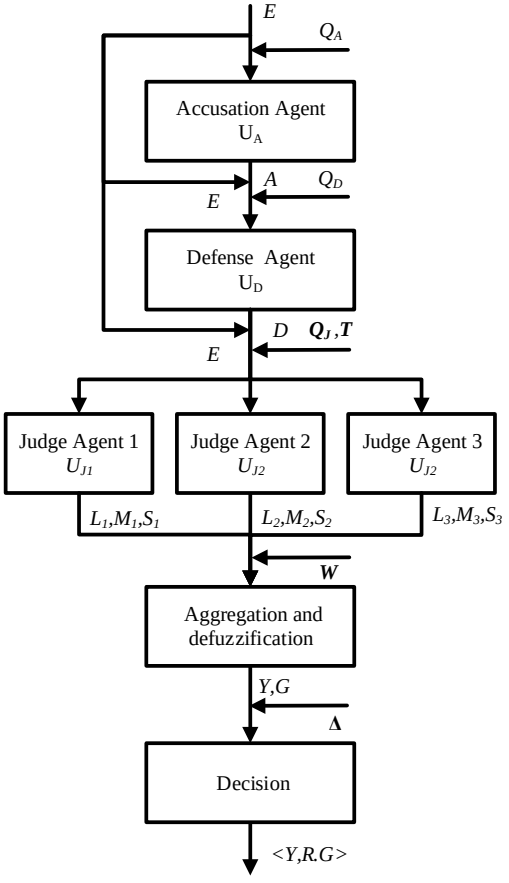


Fig. 1. **Information transformations in the process of interaction of intelligent agents**
***Source*: compiled by the authors**

For the practical implementation of the information transformations shown in Fig. 1, the general operator Φ is decomposed into five sequential stages.

Stage 1. Formation of the prosecution's position. Stage input: E – incoming electronic message adapted for analysis using IA; Q_A – system request template and U_A – specific pre-trained LLM that performs the functions of the Agent-Prosecutor.

Stage output: A – reasoned text conclusion of the prosecution about the presence of signs of financial phishing in the message. Formation of A is implemented by processing U_A of the request formed by integrating E and Q_A.

Stage 2. Formation of the position of the defense. Input of the stage: E, A, Q_D, U_D – a pre-trained LLM that performs the functions of the Agent-Lawyer. Output of the stage: D – a reasoned text conclusion of the defense, which contains counterarguments to the allegations and justification of the legitimacy of the message. Formation of D is implemented by processing the U_D query, formed on the basis of the integration of the message E, A and Q_D.

Stage 3. Judicial evaluation. Input of the stage: E, A, D, Q_J and the term set $\boldsymbol{T}$, which is used to assess the threat of the message, and $\boldsymbol{U_J}$ – a set of specific pre-trained LLMs that perform the functions of three Agent-Judges.

The output of the stage: L_1, L_2, L_3 – linguistic assessments of the threat level of the message from each member of the panel of judges; M_1, M_2, M_3 – triangular fuzzy numbers formed according to the linguistic assessments and S_1, S_2, S_3 – textual explanations of the determined assessments. The formation of these results is implemented by independently processing by each of the models from $\boldsymbol{U_J}$ the query formed on the basis of the integration of E, A, D, $\boldsymbol{T}$ and Q_J.

Stage 4. Calculation of the integral indicator of the phishing threat. The input of the stage: $L_1, L_2, L_3, M_1, M_2, M_3, S_1, S_2, S_3$ and $\boldsymbol{W}$ – a set of weight coefficients of the competence of the Agents-Judges.

Stage output: Y – an indicator of the degree of belonging of the message to the financial phishing class and G – explanations justifying the decision made. Calculation of Y is implemented by sequentially performing the procedures given in (9-12). Formation of G is carried out by concatenating linguistic assessments and explanations of each of the judges.

Stage 5. Formation of initial results. Stage input: Y, G and Δ – a set of predefined threshold values used for the final classification of the message. Stage output: tuple $\langle Y, R, G \rangle$, where R is the control decision of the security system, calculated using expression (12).

6. EXPERIMENTAL RESEARCH

To verify the effectiveness of the proposed method, two series of experiments were conducted.

The first series is a case study illustrating the operation of the proposed method at all stages of processing an incoming message. In the second series of experiments, the effectiveness of the proposed method was evaluated in terms of the accuracy of detecting financial phishing on a test set of electronic messages. When conducting experimental studies, the set $\boldsymbol{U}$ was used, in which the roles U_A and U_D were implemented based on the Claude model, and the roles U_1, U_2, U_3 were implemented based on the ChatGPT, Gemini, and DeepSeek models. The weight coefficients of the judge agents w_1, w_2 and w_3, used in expression (9), were set to 0.35, 0.35 and 0.3 based on [4]. The threshold values δ_1 and δ_2, used in (12), were taken to be equal to 40 and 70, based on the approximately uniform division of the Y scale into three decision-making zones.

When conducting the first series of experiments, an email received by one of the authors was used, the general form of which is shown in Fig. 2.

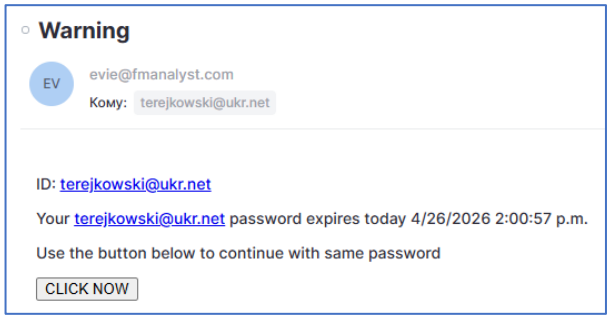


Fig. 2. **Screenshot of the phishing email**
***Source*: compiled by the authors**

The results of information transformations performed at stages 1-3 of the proposed method are shown in Fig. 3, Fig.4, Fig.5, Fig.6 and Fig.7.

In particular, Fig. 3 and Fig. 4 show the responses of U_A and U_D. U_A, having received the incoming message E according to the query template Q_A, formed the conclusion of the prosecution A. U_D, having received the message E and the conclusion A according to the query template Q_D, formed the conclusion of the defense D.

Fig. 5, Fig.6 and Fig.7 shows the results of the evaluation of case materials (E, A, D) performed by three Judge Agents using the query template Q_J.

As shown in Fig. 5, Fig. 6 and Fig.7, the responses of the Agent Judges include linguistic evaluations L_1, L_2, L_3, as well as the corresponding explanations S_1, S_2, S_3.

{"evidence_strength": "Strong", "main_accusation": "The email impersonates an account/password expiration notice to lure the recipient into clicking an embedded link and entering credentials on a page unrelated to any legitimate identity or email provider, consistent with a credential-phishing attempt.", "criteria_analysis": {"brand_impersonation": {"detected": true, "evidence": "The sender name 'EV' and address 'evie@fmanalyst.com' have no relation to any known email or IT service provider, yet the message frames itself as an official password-expiration notice for the recipient's ukr.net account, falsely implying authority from a legitimate service without matching its domain or branding."}, "urgency": {"detected": true, "evidence": "The message states the password 'expires today' with a specific same-day time stamp (4/26/2026 2:00:57 p.m.), creating a false deadline designed to pressure the recipient into immediate action without careful evaluation."}, "suspicious_hyperlinks": {"detected": true, "evidence": "The call-to-action link 'https://economycafe.com/network/index.html#terejkowski@ukr.net' points to an unrelated third-party domain ('economycafe.com') rather than any official ukr.net or email-provider domain, and appends the victim's email address as a URL fragment, a common phishing technique to pre-fill or track credential-harvesting forms."}, "confidential_data_requests": {"detected": true, "evidence": "The message instructs the user to 'continue with same password,' implicitly prompting them to re-enter or confirm their existing email password on an external page, which constitutes a request for confidential authentication credentials."}, "financial_operations": {"detected": false, "evidence": null}, "stylistic_anomalies": {"detected": true, "evidence": "The email contains grammatical and stylistic irregularities typical of phishing (e.g., 'Use the button below to continue with same password', mismatched capitalization in 'CLICK NOW', and a generic non-corporate sender display name 'EV'), along with an inconsistent mix of Russian interface text ('Кому:') and English message content, suggesting an unofficial, non-native, or mass-generated phishing template."}}}

Fig. 3. **Response of the Accusation Agent**
***Source*: compiled by the authors**

{"rebuttal_strength": "Weak",
"main_rebuttal": "While the email does contain several concerning elements, a strict reading of the message content alone does not conclusively prove malicious intent; some elements could plausibly reflect a poorly configured or third-party notification tool rather than deliberate impersonation, though the overall pattern remains highly suspicious.",
"criteria_rebuttal": {
"brand_impersonation": {"rebutted": true, "rebuttal": "The message never explicitly claims to be sent 'from ukr.net' or uses any ukr.net logo, letterhead, or official branding elements within the visible text; it only references the recipient's ukr.net address as the account ID, which could be interpreted as a third-party notification service referencing an external account rather than a direct impersonation of ukr.net itself."},
"urgency": {"rebutted": false, "rebuttal": null},
"suspicious_hyperlinks": {"rebutted": false, "rebuttal": null},
"confidential_data_requests": {"rebutted": true, "rebuttal": "The phrase 'continue with same password' does not explicitly instruct the user to type or submit their password anywhere in the visible message text; it could be read as an instruction to retain the existing password during a routine renewal process rather than an explicit request to re-enter credentials on the linked page."},
"financial_operations": {"rebutted": false, "rebuttal": null},
"stylistic_anomalies": {"rebutted": true, "rebuttal": "The mixed Russian/English interface elements (e.g., 'Кому:') are consistent with the recipient's email client rendering standard interface labels in their local language, rather than being part of the email body itself, so this particular anomaly may reflect client localization rather than an authored stylistic flaw."}

Fig. 4. Response of the Defense Agent
***Source*: compiled by the authors**

{"final_threat_level":"Very High","term_index":5,"fuzzy_parameters":{"a":75,"b":100,"c":100},"judicial_reasoning":"The accusation is supported by strong evidence, particularly the unrelated hyperlink domain, the urgent password-expiration claim, and the implied credential submission workflow. The defense correctly notes that some individual indicators, such as the absence of explicit branding and the localized email client interface, do not independently prove impersonation or malicious intent. However, these rebuttals do not materially weaken the overall case. The combination of urgency, an external non-official link, and an implied password verification process overwhelmingly supports the conclusion of a critical credential-phishing attempt."}

Fig. 5. **Response of the Judge Agent 1**
***Source*: compiled by the authors**

As a result of performing stages 4 and 5 in accordance with expressions (9-12), based on the formed linguistic evaluations L_1, L_2, L_3 and the set of weight coefficients $\boldsymbol{W}$, the indicator of the degree of belonging of the message to the financial phishing class $Y = 80.83$ was calculated and the control decision $R = Block$ was formed. The explanation G was formed by concatenating $L_1, L_2, L_3, S_1, S_2, S_3$, shown in Fig. 5, Fig.6 and Fig.7. The results of the first series of experiments demonstrated an example of correct implementation of the proposed method across all stages of message processing.

1. Final Threat Assessment

- **Threat Level: High**
- **Term Index: 4**
- **Fuzzy Parameters:** $a = 50, b = 75, c = 100$

2. Judicial Reasoning & Balance of Arguments

Upon reviewing the evidence presented by the Prosecution and the arguments raised by the Defense:

- **Strength of Prosecution:** The Prosecution provided strong evidence of a credential phishing attempt, identifying key indicators:
  - **Manufactured Urgency:** The same-day password expiration warning accompanied by a specific timestamp pressures the user to act immediately without verifying the claim.
  - **Suspicious Link:** The action button directs the user to an unrelated third-party domain (`economycafe.com`), rather than an official `ukr.net` service domain. Appending the user's email address as a URL fragment is a standard tactic used to pre-fill or track credential-harvesting forms.
- **Weakness of Defense:** The Defense's counterarguments are unpersuasive:
  - The Defense argues that the message does not explicitly request credentials. However, the phrase *"continue with same password"* in the context of a link to an external renewal portal logically implies entering or confirming existing credentials.
  - Most importantly, the Defense left the critical elements—the urgency lure and the suspicious external URL—completely **unrebutted**.

3. Final Conclusion

The combination of spoofing an account renewal workflow, using an unrelated third-party domain, and imposing artificial time pressure demonstrates clear phishing intent.

Decision: A **High** threat level is assigned.

Fig. 6. Response of the Judge Agent 2
***Source*: compiled by the authors**

{"final_threat_level": "High", "term_index": 4, "fuzzy_parameters": {"a": 50, "b": 75, "c": 100}, "judicial_reasoning": "The prosecution's evidence of a mismatched sender domain, urgent deadline, off-domain hyperlink with embedded victim email, and implicit credential request is strong. The defense's rebuttals that the message does not explicitly claim to be from ukr.net or explicitly ask for password entry are technically true but do not undermine the overall phishing pattern, as the link destination and urgency remain unrebutted. The stylistic anomalies are partially explained by client localization, but the core indicators of credential phishing are compelling. Thus, the threat level is High."}

Fig. 7. **Response of the Judge Agent 3**
***Source*: compiled by the authors**

For the second series of experiments, an experimental corpus of 1,000 messages was compiled by the authors, comprising 500 phishing and 500 legitimate samples. Of these, 850 samples (85 %) were author-composed full-length email messages representing corporate contexts, extended textual structures, and complex social-engineering scenarios, with some messages containing masked hyperlinks and others containing no explicit technical indicators of compromise. The remaining 150 samples (15 %) were selected from the SMS Spam Collection [34] as a supplementary source of concise textual messages to diversify the evaluated linguistic patterns.

In this study, financial phishing is defined in accordance with Section 1 (based on [4], [5]) as a form of cyber fraud that uses social engineering to induce the recipient to transfer funds or disclose payment and authentication data. The task is treated as binary classification: messages meeting this

definition were labeled as phishing, and all others as legitimate. Labels of author-composed emails were determined by their creation scenario, while specific phishing SMS messages were manually selected from the SMS Spam Collection [34] to diversify the evaluated textual patterns. All labels were assigned before the evaluation and independently of the criteria used in the query templates. The legitimate class included generic spam and other types of unwanted messages, because separating financial phishing from them was outside the scope of this research. The general indicators in the query templates reflect common phishing patterns, while financial specificity is addressed by the criteria on confidential data requests and financial operations.

Message preprocessing was limited to normalizing text encoding, inserting each message into the query template as a separate data block, and checking its length against the input limits of the models. The messages were provided to the agents as text, and URLs were treated as text strings without following them. The contents of email attachments were not analyzed. The method relies on the natural-language processing capabilities of the underlying large language models and can therefore potentially be applied to messages in other languages, although its performance across languages was not specifically evaluated in this study. The corpus contains no duplicate messages, and the messages are written in several languages, predominantly Ukrainian and English.

The experimental evaluation of the proposed framework was carried out in June–July 2026, following preparatory sample collection initiated in April 2026. To reflect the operational conditions of publicly accessible conversational AI tools, the agents were deployed through their primary web interfaces under standard platform configurations: the Accusation Agent and Defense Agent were executed using Anthropic's Claude (labeled as Sonnet 5 · Medium in the interface), while the Judicial Panel comprised OpenAI's ChatGPT using the default model available in the free-tier interface at the time of testing, Google's Gemini (labeled as Flash), and DeepSeek (Instant mode). Because commercial web platforms do not disclose backend model snapshot identifiers or permit manual control of the temperature parameter, generation relied on the platform-default decoding configuration. Potential vendor-side model updates during the evaluation period therefore cannot be completely excluded and are acknowledged as an experimental limitation.

To simulate real-time email gateway operation, each message was evaluated using a single-pass processing procedure without repeated trials or selective filtering. The analysis was performed in sequential session batches, with conversational context cleared between batches to minimize contextual carryover across evaluated samples. In rare instances where an agent returned conversational commentary instead of the required JSON structure, an immediate clarifying prompt was issued to obtain a structured response. This procedure was applied consistently to all 1,000 evaluated messages. Regarding potential data leakage, 85% of the corpus consisted of author-composed emails prepared in 2025-2026, which substantially reduces the likelihood of their direct presence in the models' pre-training data. However, absolute absence of prior exposure cannot be independently guaranteed for closed commercial LLMs.

The results of the analysis of the test set messages by the proposed method are given in Table 1. For the calculation of the binary classification metrics reported in Table 1, the same rule was applied to all evaluated configurations. Messages assigned to the Allow zone were treated as legitimate, whereas messages assigned to the Quarantine or Block zones were treated as phishing detections.

In addition, Table 1 shows the results of the analysis of the same sample by separate language models Gemini, ChatGPT, and DeepSeek, as well as by the weighted integration method [4] for comparison purposes, which applies identical competence weights for individual IAs, but does not incorporate the procedure of judicial interaction.

Table 1. **Recognition results**

Indicator	e_1	e_2	e_3	e_W	e_Σ
True Positives (Phishing), samples	465	475	455	480	490
True Negatives (Legitimate), samples	475	470	480	475	494
False Positive Rate, %	5.0	6.0	4.0	5.0	1.2
False Negative Rate, %	7.0	5.0	9.0	4.0	2.0
Error Rate, %	6.0	5.5	6.5	4.5	1.6

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

For a concise presentation of the results in Table 1, the following notations are used: $e_1 - e_3$ are recognition results by Gemini, ChatGPT and DeepSeek models; e_W is by the weighted integration method; e_Σ – by the proposed method.

As shown in Table 1, the proposed method provides the lowest classification error among the

solutions directly evaluated in this study under the same experimental protocol, with an error rate of 1.6%.

Nonetheless, the implementation of the proposed method requires an increase in the amount of computing resources, since it involves the use of five IAs and the implementation of procedures related to the implementation of judicial interaction.

7. DISCUSSION

Experimental studies have shown that the proposed method reduces the total classification error rate to 1.6 %, which corresponds to a 3.4-4.1–fold reduction compared with the individual LLM-based solutions and to an approximately 2.8-fold reduction compared with the weighted integration method. Thus, under the experimental conditions and comparative protocol adopted in this study, the proposed method achieved the lowest classification error among the evaluated solutions. The obtained error rate is also comparable to the results reported for modern multi-agent LLM-based approaches to phishing detection, being lower than the Error Rate of approximately 2.11 % reported in [35] and the approximately 6.1 % reported in [36], while remaining slightly higher than the best configurations reported in [27], with Error Rates ranging from 0.6 % to 1.3 %. Since these results were obtained under different experimental conditions and on different datasets, they are considered as reference points rather than direct comparative baselines.

For the proposed method, precision, recall, specificity, and F1-score are 0.988, 0.980, 0.988, and 0.984, respectively. The false negative rate of 2.0 % and the false positive rate of 1.2 % reported in Table 1 characterize the two error types separately. In phishing detection, the false negative rate is of particular practical relevance because missed phishing messages may result in security incidents, whereas false positives cause legitimate messages to be quarantined or blocked, which also entails operational costs, although their impact is typically less severe. A formal cost-sensitive analysis was not performed, since the relative costs of the two error types depend on the deployment context. The error rate of 1.6 % for the proposed method corresponds to a 95% Wilson confidence interval of 1.0-2.6 %, while the 4.5 % error rate of the weighted integration method corresponds to a 95 % confidence interval of 3.4-6.0 %. These intervals characterize sampling variability within the evaluated corpus and do not account for possible variability across repeated LLM runs, since repeated-run analysis was not performed. Because per-message outputs were not stored and only aggregate confusion counts were recorded, paired statistical tests such as McNemar's test could not be performed. Therefore, the observed difference in error rates is interpreted as an empirical result under the adopted experimental conditions rather than as statistically confirmed superiority. ROC-AUC was not calculated because the continuous aggregated scores were not stored for individual messages, and the evaluation was performed on the final threshold-based decisions.

This result is consistent with the implementation in the proposed method of the judicial interaction procedure, which combines an architecturally homogeneous pair of agents that implement the prosecution and defense parties with an architecturally heterogeneous composition of the panel of judges. Meanwhile, the use of a single LLM to implement the roles of Accusation Agent and Defense Agent is due to the fact that the competition between the parties is ensured by the opposite construction of the corresponding query templates, and not by the architectural diversity of the basic models. Instead, the use of LLMs different in architecture and supplier in the composition of the panel of judges is aimed at ensuring the independence of individual assessments, since the use of a single model for all judges creates the prerequisites for increasing the correlation of individual errors. Such an organization of IA interaction allows for separating the process of forming reasoned positions and the process of their independent assessment.

The operation of such interaction is illustrated by the case study, in which the panel of judges demonstrated the discrepancy of individual assessments of the threat level: one of the judges defined it as Very High, while the other two independently of each other came to the conclusion High. In this case, the final decision was formed via weighted fuzzy aggregation procedure, which ensures the coordination of the judges' individual assessments into a single final decision. The effectiveness of the judicial interaction procedure is further supported by comparing the proposed method with the weighted integration method [4].

Since both methods use the same composition of LLMs and identical competence weights, the observed improvement in recognition accuracy is consistent with the contribution of judicial interaction, which is absent in the specified method.

To ensure consistent experimental conditions, all configurations under comparison were evaluated

using the same dataset, and the judge weights and threshold values remained identical across all instances of their application. The judge weights were adopted from [4] without recalibration on email-phishing data, and the threshold values were treated as predefined security policy parameters rather than values tuned to a specific sample. Consequently, the choice of these parameters cannot account for the observed performance differences between the proposed method and the weighted integration method. Calibrating the weights and thresholds using a validation set was not performed, and the sensitivity of the results to their values was not analyzed. The comparison of the adopted aggregation scheme with majority voting and calibrated probability estimates in terms of classification accuracy was not performed either. These aspects remain subjects for future research.

However, the implementation of the proposed method requires an increase in the amount of computational resources. The number of calls to the LLM increases fivefold compared to the use of a separate IA and by a factor of 1.67 compared to the weighted integration method, which does not involve judicial interaction. In this regard, the proposed method is focused primarily on tasks in which the accuracy of decision-making takes priority over resource consumption.

Latency, cost per message, throughput, and token limits were not measured, since the agents were run through the web interfaces of the platforms, and integration with email security gateways or SIEM systems was not considered.

The direct experimental comparison in this study was limited to the LLM-based and integration-based solutions evaluated under the adopted experimental protocol. Specialized fine-tuned phishing classifiers, transformer-based (BERT-type) models, conventional machine-learning approaches, and other debate-based multi-agent methods were not included, since the proposed method operates without training on labeled data, whereas such approaches require a separate training and validation protocol, and the available results of debate-based methods were obtained under different experimental conditions. Their comparison on the same corpus is a direction for further research.

The study does not include an ablation study of the judicial interaction. Configurations without the Defense Agent, without the Accusation Agent, and with majority voting, self-consistency, or debate prompting were not evaluated, so the contributions of individual components remain unquantified. The independence of the prosecution and the defense is relative, since the defense is formed after the accusation and both agents are based on the same language model, which may correlate their errors. The judges receive the original message, and their query template directs them to verify the arguments of the parties and does not require a new analysis of the message. An additional limitation of the current implementation of the method is the analysis of only the text content of an electronic message, which leaves out of consideration multimodal threats that can also be used by attackers to implement phishing attacks. A further limitation concerns the robustness of the method to prompt injection and adversarial emails. The text of an incoming message is controlled by the sender and is passed to all agents of the framework, so embedded instructions or specially crafted text could distort the arguments of the prosecution and defense agents and the verdicts of the judges, leading to an underestimated threat level. The evaluated corpus did not contain such messages, and the query templates did not include protective measures against them. The robustness of the method to these attacks was therefore not evaluated and remains an open risk. The use of different LLMs in the panel of judges and the weighted aggregation may limit the effect of a single manipulated agent, but this cannot be assumed to ensure resilience, since the same text is supplied to all agents. Experimental assessment on adversarial messages and the development of input isolation and output validation measures are directions for further research.

The use of externally hosted models also raises data protection concerns. In the experiments, the agents were accessed through the web interfaces of external platforms, and real corporate or personal correspondence was not used. Operational use on real email may require models operated within the controlled environment of the organization, and the method was not evaluated with such models.

Consequently, future research should be directed at an ablation study of the components of judicial interaction and at optimizing the architecture of the proposed method in order to reduce resource consumption and message processing delay, in particular by parallelizing individual stages of judicial interaction. Latency and cost per message under deployment conditions should also be assessed. Another promising direction is to expand the functional capabilities of the method by integrating tools for analyzing the graphic and multimedia content of electronic messages to detect multimodal phishing attacks.

8. CONCLUSIONS

As a result of the research, a method for detecting financial phishing in e-mail was developed based on a model of judicial interaction of intelligent agents, built using large language models. The key feature of the proposed method lies in the adaptation of the principle of adversarial litigation to the organization of interaction of intelligent agents, which ensures the formation of reasoned positions of the prosecution and defense parties with their subsequent evaluation by a panel of agent judges and aggregation of the results obtained using fuzzy set theory. In the process of developing the method, a model of judicial interaction of intelligent agents was built, which formalizes information transformations between the participants in the process, the structure of query templates, the procedure for independent judicial evaluation and the mechanism for integrating individual conclusions into the final decision on whether the message belongs to the class of financial phishing.

Experimental studies have shown that the proposed method provides a total classification error rate of 1.6 %, which is 3.4-4.1 times lower than that of the individual large language models evaluated in this study and approximately 2.8 times lower than that of the weighted integration method without the use of judicial interaction. Thus, under the experimental conditions and comparative protocol adopted in this study, the proposed method demonstrated the lowest classification error among the evaluated solutions. These results suggest that the increase in recognition accuracy is associated not only with the use of an ensemble of large language models, but also with the organization of their interaction in accordance with the proposed model of the judicial process. The practical value of the proposed method lies in the possibility of increasing the accuracy of financial phishing recognition without additional training of large language models, which may facilitate the adaptation of the method to new phishing scenarios and its use in email protection tools where accuracy takes priority over resource consumption.

Conflict of interests: Authors declare that they do not have a conflict of interest, in particular financial, personal, copyright, or of any other nature, which could have influenced the research or the results published in this article

Financing: The research was carried out without financial support

Accessibility data: The query templates, parameters of the experiment, and aggregated results supporting the findings of this study are presented in the main text of the manuscript. The experimental corpus is not publicly available because the research is ongoing

Using artificial intelligence tools: During the preparation of the manuscript, the Claude Sonnet 5.5 model (Anthropic) was used for language editing, translation, and assistance in the wording of individual fragments of the text. All fragments processed with this tool were checked and edited by the authors. The scientific provisions, experimental results, and conclusions are the intellectual contribution of the authors. The large language models used as intelligent agents in the experiments (Claude, ChatGPT, Gemini, and DeepSeek) are the object of the study and are described in Section 6

REFERENCES

1. 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. IIP Series*. 2024; 4: 92–101. DOI: https://doi.org/10.58532/nbennurdtch9.

2. 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.

3. 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.

4. 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; 9 (1): 71–84, https://www.scopus.com/pages/publications/105031963942. DOI: https://doi.org/10.15276/hait.09.2026.06.

5. “Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive)”.

*Official Journal of the European Union, L*333, 80–152. – Available from: https://eur-lex.europa.eu/eli/dir/2022/2555/oj.

6. "The White House". *National cybersecurity strategy*. – Available from: https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Cybersecurity-Strategy-2023.pdf.

7. "On the decision of the National Security and Defense Council of Ukraine of September 14, 2020. On the National Security Strategy of Ukraine". *Decree No. 392/2020*. 2020. – Available from: https://zakon.rada.gov.ua/laws/show/392/2020?lang=en.

8. "On the decision of the National Security and Defense Council of Ukraine of May 14, 2021 "On the Cybersecurity Strategy of Ukraine*". Decree No. 447/2021*. 2021. – Available from: https://zakon.rada.gov.ua/laws/show/447/2021?lang=en.

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

10. 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, https://www.scopus.com/pages/publications/85159346775. DOI: https://doi.org/10.3390/s23094403.

11. 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, https://www.scopus.com/pages/publications/85190278063. DOI: https://doi.org/10.3390/s24072077.

12. 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, https://www.scopus.com/pages/publications/85107530292. DOI: https://doi.org/10.1016/j.cose.2021.102328.

13. Firman, T. & 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.

14. 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.

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, https://www.scopus.com/pages/publications/85146629037.DOI: https://doi.org/10.1016/j.jksuci.2023.01.004.

16. Komosny, D. "Phishing detection on webpages in European non-English languages based on machine learning". *Scientific Reports*. 2025; 15: 37472, https://www.scopus.com/pages/publications/105019775362. DOI: https://doi.org/10.1038/s41598-025-21384-w.

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

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. 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 *2nd International Conference on Advanced Trends in Information Theory (ATIT)*. 2020. p. 276–280. DOI: https://doi.org/10.1109/ATIT50783.2020.9349302.

21. Toliupa, S., Tereikovskiy, I., Dychka, I., Tereikovska, L. & Trush, A. "The method of using production rules in neural network recognition of emotions by facial geometry". In *2019 3rd International*

Conference on Advanced Information and Communications Technologies (AICT). 2019. p. 323–327. DOI: https://doi.org/10.1109/AIACT.2019.8847847.

22. 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.

23. 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/2507.04426v1.

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

25. 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.

26. 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.

27. Nguyen, N. T. V., Childress, F. D. & Yin, Y. "Debate-driven multi-agent LLMs for phishing email detection". *arXiv*. 2025. DOI: https://doi.org/10.48550/arXiv.2503.22038.

28. Chun, J., Elkins, K. & Lee, Y. S. "AgenticSimLaw: A juvenile courtroom multi-agent debate simulation for explainable high-stakes tabular decision making". *arXiv*. 2026. DOI: https://doi.org/10.48550/arXiv.2601.21936.

29. Devadiga, P. & Agarwal, P. "SAMVAD: A multi-agent system for simulating judicial deliberation dynamics in India". *arXiv*. 2025. DOI: https://doi.org/10.48550/arXiv.2509.03793.

30. Ross, T. J. "Fuzzy logic with engineering applications (4th ed.)". *John Wiley & Sons*. 2016.

31. 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.

32. 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.

33. 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.

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

35. Xue, Y., Spero, E., Koh, Y. S. & Russello, G. "MultiPhishGuard: An explainable and adaptive multi-agent LLM system for phishing email detection". *arXiv*. 2025. DOI: https://doi.org/10.48550/arXiv.2505.23803.

36. Li, W., Manickam, S., Chong, Y.-W. & Karuppayah, S. "PhishDebate: An LLM-based multi-agent framework for phishing website detection". *arXiv*. 2025. DOI: https://doi.org/10.48550/arXiv.2506.15656.

Received 03.09.2026
Received after revision 26.10.2026
Accepted 02.11.2026

DOI: https://doi.org/10.15276/hait.09.2026.42
УДК 004.056.5:004.89:004.77

Метод виявлення фінансового фішингу в електронній пошті на основі моделі судової взаємодії інтелектуальних агентів

Терейковська Людмила Олексіївна[1)]
ORCID: https://orcid.org/0000-0002-8830-0790; tereikovskal@gmail.com. Scopus Author ID: 57198815503
Крак Юрій Васильович[2)]
ORCID: https://orcid.org/ 0000-0002-8043-0785; yuri.krak@gmail.com. Scopus Author ID: 6602577533
Корченко Анна Олександрівна[3)]
ORCID: https://orcid.org/0000-0003-0016-1966; anna.korchenko@uken.krakow.pl. Scopus Author ID: 56029291400
Баліна Олена Іванівна[1)]
ORCID: https://orcid.org/ 0000-0003-4621-9668; elena.i.balina@gmail.com. Scopus Author ID: 57223105709
Терейковський Олег Ігоревич[2)]
ORCID: https://orcid.org/0000-0001-5045-0163; tereikovskyio@gmail.com. Scopus Author ID: 57216153388
[1)] Київський національний університет будівництва і архітектури, пр. Повітряних Сил, 31. Київ, 03037, Україна
[2)] Київський національний університет імені Тараса Шевченка, вул. Володимирська, 60. Київ, 01033, Україна
[3)] Університет національної комісії з питань освіти, вул. Підхорунжих, 2. Краків, 30-084, Польща

АНОТАЦІЯ

Актуальність: зумовлена потребою вдосконалення виявлення фінансового фішингу в електронній пошті, зокрема коли шахрайські повідомлення використовують складні прийоми соціальної інженерії та не містять явних технічних індикаторів компрометації. Традиційні засоби фільтрації та рішення на основі окремих великих мовних моделей можуть демонструвати недостатню ефективність у виявленні такого фішингу, оскільки, як правило, здійснюють ізольоване оцінювання тексту й не передбачають зіставлення різних інтерпретацій його змісту. **Метою статті** є розроблення методу виявлення фінансового фішингу в електронній пошті на основі моделі судової взаємодії інтелектуальних агентів, що забезпечить підвищення ефективності його розпізнавання. **Завдання:** розроблення моделі багатоагентного середовища судової взаємодії інтелектуальних агентів із поданням оцінок у вигляді нечітких чисел, розроблення методу виявлення фінансового фішингу в електронній пошті, що реалізує запропоновану модель у вигляді послідовних етапів аналізу вхідного повідомлення та узагальнення нечітких оцінок агентів, а також експериментальне оцінювання ефективності методу. **Методи:** у дослідженні застосовано багатоагентну модель судової взаємодії інтелектуальних агентів на основі великих мовних моделей та апарат нечіткої логіки. Індивідуальні вердикти виражаються лінгвістичними оцінками, які відображаються трикутними нечіткими числами, і агрегуються із застосуванням вагових коефіцієнтів компетентності. **Наукова новизна** полягає в адаптації процедури змагальності судового процесу до організації взаємодії інтелектуальних агентів, яка передбачає формування протилежних аргументованих позицій обвинувачення і захисту, їх оцінювання колегією суддів та агрегування вердиктів суддів на основі нечіткої логіки, що забезпечує можливість підвищення точності розпізнавання фінансового фішингу. **Практична значимість:** запропонований метод може бути використаний у засобах захисту електронної пошти, де точність має пріоритет над споживанням ресурсів, надаючи текстове обґрунтування рішення щодо класифікації повідомлення. **Результати:** експериментальні дослідження на корпусі, сформованому авторами, із використанням Claude, ChatGPT, Gemini і DeepSeek показали нижчий рівень помилки класифікації для запропонованого методу, ніж для окремих великих мовних моделей і для методу зваженої інтеграції без судової взаємодії, за того самого експериментального протоколу. **Висновки:** запропонований метод забезпечує можливість підвищення точності розпізнавання фінансового фішингу в електронній пошті на основі судової взаємодії інтелектуальних агентів, реалізованих на базі великих мовних моделей, без додаткового навчання цих моделей.

Ключові слова: фінансовий фішинг; електронна пошта; великі мовні моделі; інтелектуальні агенти; соціальна інженерія; кібербезпека

ABOUT THE AUTHORS



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, проспект Повітряних Сил, 31. Київ, 03037, Україна
Галузь досліджень: нейромережевий аналіз біометричних параметрів; розробка інструментів біометричної автентифікації; розпізнавання емоцій; інтелектуальні системи захисту інформації



Iurii V. Krak – Corresponding Member of National Academy of Sciences of Ukraine, Doctor of Physical and Mathematical Sciences, Professor, Head of Department of Theoretical Cybernetics. Taras Shevchenko National University of Kyiv, 60, Volodymyrska Str. Kyiv, 01033, Ukraine
ORCID: https://orcid.org/ 0000-0002-8043-0785; yuri.krak@gmail.com. Scopus Author ID: 6602577533
***Research field*:** Artificial Intelligence; Robotics; Communicative Information; Neurocomputing; Sign Languages; Natural Language Processing; Pattern Recognition; Human-Computer Interfaces; Data Classification and Clustering; Emotion Research, Information Technology

Крак Юрій Васильович - член-кореспондент НАН України, доктор фізико-математичних наук, професор, завідувач кафедри Теоретичної кібернетики. Київський національний університет імені Тараса Шевченка, вул. Володимирська, 60. Київ, 01033, Україна
Галузь досліджень: робототехніка; комунікативна інформація; нейрокомп'ютери; жестові мови; обробка природної мови; розпізнавання образів; інтерфейси людина-комп'ютер; класифікація та кластеризація даних; дослідження емоцій, інформаційні технології



Anna O. 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/ 0000-0003-0016-1966; anna.korchenko@uken.krakow.pl. Scopus Author ID: 56029291400
***Research field*:** Information Security, Computer Security, Artificial Immune Systems, Business Continuity Management, Psychophysiological Studies Using Polygraph

Корченко Анна Олександрівна - доктор технічних наук, професор, професор кафедри комп'ютерної інженерії та кібербезпеки. Університет національної комісії з питань освіти, Підхорунжих, 2. Краків, 30-084, Польща
Галузь досліджень: інформаційна безпека, комп'ютерна безпека, штучні імунні системи, управління безперервністю бізнесу, психофізіологічні дослідження з використанням поліграфа



Olena I. Balina - Candidate of Engineering Sciences, Associate 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/ 0000-0003-4621-9668; elena.i.balina@gmail.com. Scopus Author ID: 57223105709
***Research field*:** Mathematical Analysis; Probability Theory and Mathematical Statistics; Intelligent Information Systems

Баліна Олена Іванівна - кандидат технічних наук, доцент, доцент кафедри інформаційних технологій проєктування та прикладної математики. Київський національний університет будівництва і архітектури, проспект Повітряних Сил, 31. Київ, 03037, Україна
Галузь досліджень: математичний аналіз; теорія ймовірностей та математична статистика; інтелектуальні інформаційні системи



Oleh I. Tereikovskyi - Assistant, Department of Theoretical Cybernetics. Taras Shevchenko National University of Kyiv, 60, Volodymyrska Str. Kyiv, 01033, 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 Ігоревич – асистент кафедри Теоретичної кібернетики. Київський національний університет імені Тараса Шевченка, вул. Володимирська, 60. Київ, 01033, Україна
Галузь досліджень: системи біометричної автентифікації; розробка програмного забезпечення для інформаційної безпеки; машинне навчання