




PhishIntentionLLM: Uncovering Phishing Website Intentions through Multi-Agent Retrieval-Augmented Generation

WENHAO LI¹ , SELVAKUMAR MANICKAM¹✉ , YUNG-WEY CHONG²  and SHANKAR KARUPPAYAH¹ 

¹ Cybersecurity Research Centre, Universiti Sains Malaysia, Pulau Pinang, Malaysia

² School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia
wenhaoli@ieee.org, {selva, chong, kshankar}@usm.my

Abstract. Phishing websites remain a major cybersecurity threat, yet existing methods primarily focus on detection, while the recognition of underlying malicious intentions remains largely unexplored. To address this gap, we propose *PhishIntentionLLM*, a multi-agent retrieval-augmented generation (RAG) framework that uncovers phishing intentions from website screenshots. Leveraging the visual-language capabilities of large language models (LLMs), our framework identifies four key phishing objectives: Credential Theft, Financial Fraud, Malware Distribution, and Personal Information Harvesting. We construct and release the first phishing intention ground truth dataset (~2K samples) and evaluate the framework using four commercial LLMs. Experimental results show that *PhishIntentionLLM* achieves a micro-precision of 0.7895 with GPT-4o and significantly outperforms the single-agent baseline with a ~95% improvement in micro-precision. Compared to the previous work, it achieves 0.8545 precision for credential theft, marking a ~4% improvement. Additionally, we generate a larger dataset of ~9K samples for large-scale phishing intention profiling across sectors. This work provides a scalable and interpretable solution for intention-aware phishing analysis.

Keywords: Cybercrime · Large Language Models (LLMs) · Phishing Website · Multi-Agent Retrieval-Augmented Generation (RAG) System.

1 Introduction

Phishing is a dominant form of cybercrime that exploits both system vulnerabilities and human psychology to deceive users into disclosing sensitive information [1,2]. Among its various forms, phishing websites pose one of the most critical threats, using visually deceptive interfaces to impersonate trusted entities [3]. The prevalence of such attacks has continued to grow, with the Anti-Phishing Working Group (APWG) reporting 989,123 unique phishing websites in the fourth quarter of 2024, compared to 888,585 in the same period of 2021 [4]. This rise is fueled by phishing-as-a-service platforms, phishing toolkits, and affordable infrastructure as well as advanced evasive techniques phishers used [5,6],

making it easier for attackers to launch more long-lasting large-scale phishing campaigns. The impacts of these attacks go far beyond financial losses, including intellectual property theft and significant reputational damage [7].

To counter phishing website, a large body of research has focused on phishing website detection using heuristic-based, machine learning, and deep learning techniques that analyze features such as URLs, HTML structures, and domain metadata [8,9]. Although these studies effectively detect phishing websites, none have focused specifically on identifying the malicious intentions behind these phishing websites.

Understanding the underlying intentions behind phishing websites provides deeper insights into attacker strategies and motivations. For instance, a phishing site targeting the banking sector with a credential theft intent presents a different threat profile than one designed to harvest healthcare data. Even phishing websites impersonating the same brand can exhibit diverse malicious goals, as illustrated in Fig. 1. This level of granularity enables more precise threat intelligence, tailored defense strategies, and informed regulatory responses.

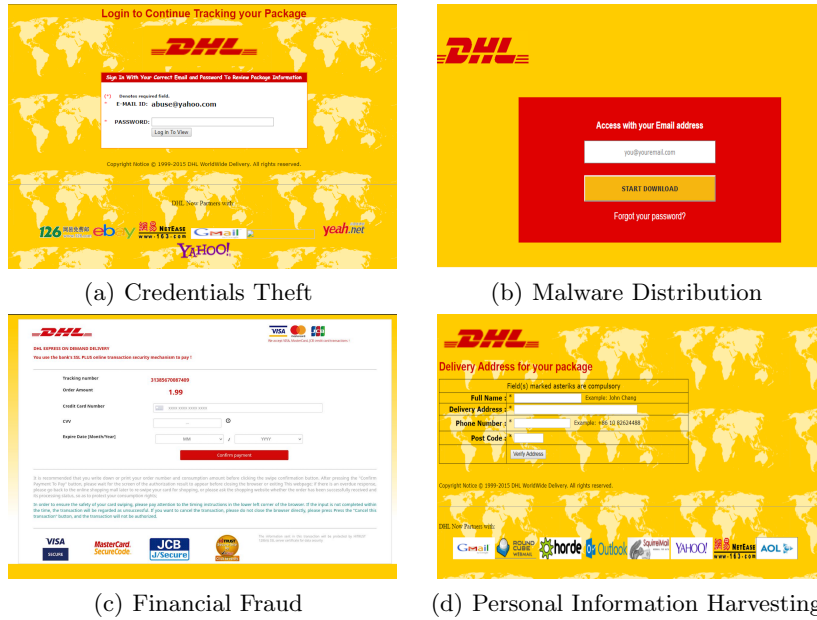


Fig. 1. The Phishing DHL Websites with varied Intentions.

To bridge this gap, we propose a novel multi-agent retrieval-augmented generation (RAG) framework for identifying malicious intentions behind phishing websites through visual screenshot analysis. Our approach leverages the visual-language capabilities of large language models (LLMs) in combination with a domain-specific retrieval module to detect four primary phishing threat categories: Credential Theft, Financial Fraud, Malware Distribution, and Personal Information Harvesting.

The key contributions of this paper are as follows:

- We manually construct and publicly release the first phishing intention ground truth dataset containing $\sim 2K$ phishing website samples with screenshots ¹, labeled intention categories, and sectoral information.
- We propose *PhishIntentionLLM*, a novel multi-agent RAG framework that synergizes general and expert agents with knowledge bases for phishing intention detection.
- We comprehensively evaluate our framework using four commercial LLMs across multiple performance metrics, and benchmark it against a single-agent baseline and prior work focused on credential theft detection.
- We leverage our framework to detect the intention of a larger-scale phishing samples using GPT-4o and generate and publicly release a $\sim 9K$ phishing intention dataset ¹, analyzing the distribution of phishing intentions and their associated sectors to uncover empirical patterns in attacker behaviors.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 introduces four primary phishing intentions. Section 4 describes the architecture of the proposed multi-agent framework. Section 5 outlines the models and metrics used for evaluation as well as the ground truth dataset construction. Section 6 presents experimental results and comparative analysis. Section 7 discusses a larger-scale empirical phishing intention analysis with proposed framework, while Section 8 concludes the paper with future works.

2 Related Work

Phishing detection has been widely studied using heuristics [10], machine learning and deep learning approaches that analyze features such as URLs [11], HTML content [12], domain metadata [13] or hybrid features [14]. While these methods have proven effective in identifying phishing websites, many of these techniques rely on code-level or metadata features that can be easily obfuscated or manipulated, making them particularly vulnerable to cloaking techniques that hide malicious content from automated scanners while displaying convincing visuals to human victims [6].

To address limitations of feature-based methods, recent studies have explored visual and multimodal analysis [15], using screenshots, layout structure, and Optical Character Recognition (OCR)-extracted text for phishing detection in addition to existing approaches [16,17]. These approaches better mimic how human users perceive websites and are naturally resilient to cloaking techniques that hide malicious content in code while keeping visual appearance intact. By focusing on visual elements, such models can generalize across phishing pages that differ at the source-code level but share deceptive front-end appearances.

However, they offer limited insights into the nature or intent of the attack. Once a phishing website is detected, understanding its specific malicious goal

¹ Dataset: <https://github.com/v1ct0r133/PhishIntentionLLM>

(e.g., credential theft vs. malware distribution) is often overlooked, leaving a critical gap in threat profiling and response strategies.

Understanding attackers’ intentions has gained attention in broader cybersecurity contexts especially in the field of cybercrime [18], such as classifying types of email phishing (e.g., business email compromise vs. generic scams) [19], ransomware behavior [20], and scams and fraud campaign strategies [21]. These studies demonstrate that identifying intent is crucial for targeted mitigation, forensic analysis, and policy-making. However, such intent-focused analysis has rarely been applied to phishing websites, especially in a structured, automated manner.

To date, only one known study has attempted to detect phishing website intentions, focusing specifically on credential theft [22]. While this work represents an important step forward and demonstrates that identifying intent can aid in detecting previously unknown phishing websites, it is limited to credential theft and does not address other potential malicious phishing intention. Therefore, a more scalable and generalizable approach is required to capture the full spectrum of phishing intentions.

Recent advancements in LLMs and RAG have enabled more context-aware and interpretable solutions across various natural language processing and security tasks [23,24]. These models integrate external knowledge retrieval with language understanding, offering robust performance on complex, multi-stage tasks.

In summary, although phishing detection has advanced through traditional and deep learning methods, most approaches focus solely on binary classification, neglecting the identification of underlying malicious objectives. While visual and multimodal analysis helps address obfuscation, it has yet to be applied to phishing intention detection. Existing intent recognition efforts in cybersecurity are limited, with only one study on phishing intent that targets credential theft and lacks support for multiple intentions. Moreover, powerful tools like RAG and LLMs remain underexplored in this context. This highlights the need for a comprehensive, scalable framework that leverages visual and contextual cues to identify diverse phishing intentions.

3 Background

Phishing websites employ various deceptive strategies, each designed with specific malicious intentions. As illustrated in Fig. 1, our analysis of real-world phishing campaigns reveals four predominant categories of malicious intent: credential theft, malware distribution, financial fraud, and personal information harvesting. These intentions represent the primary objectives that drive phishing attacks in the current threat landscape. This section provides essential background on each category to establish a foundation for understanding the methodology presented in this study.

Credential theft, as shown in the Fig.1(a), perhaps the most common phishing objective, involves attackers creating counterfeit websites that mimic legitimate

platforms to capture user authentication credentials. These attacks typically impersonate trusted entities such as financial institutions, email providers, or corporate platforms, employing visual and structural similarities to the original sites. Once obtained, these credentials facilitate account takeovers, enabling attackers to access sensitive information, conduct unauthorized transactions, or establish footholds for further attacks.

Malware distribution phishing leverages deceptive interfaces to induce users to download malicious software, as shown in the Fig. 1(b). Such attacks frequently masquerade as software updates, security scans, media players, or document viewers. The distributed malware may include ransomware, information stealers, remote access trojans, or other malicious payloads that compromise system integrity and user privacy. These attacks typically feature prominent download buttons, alarming security warnings, or counterfeit system notifications.

Financial fraud phishing specifically targets monetary exploitation through various schemes designed to manipulate victims into financial transactions, as shown in the Fig. 1(c). These attacks employ deceptive narratives including fake investment opportunities, fraudulent merchandise sales, technical support scams, and counterfeit financial alerts. Distinguished by their emphasis on payment information collection or direct transfer solicitation, these attacks often create artificial urgency to circumvent rational decision-making processes.

Personal information harvesting, as shown in the Fig. 1(d), aims to collect comprehensive personally identifiable information beyond mere credentials. These attacks solicit sensitive data including government identification numbers, home addresses, employment details, financial information, and healthcare data, often through illegitimate forms, surveys, or registration pages. This information enables identity theft, sophisticated social engineering, or sale on underground markets for subsequent exploitation.

4 Methodology

This section presents the proposed methodology of this study. We describe the system’s hierarchical agent architecture, knowledge retrieval mechanisms, processing pipeline. The formalized algorithm for this framework is proposed to demonstrate how these components interact to produce accurate threat classifications with supporting evidence chains, addressing the challenging task of multi-category phishing intention identification.

4.1 System Architecture

PhishIntentionLLM represents a novel multi-agent RAG framework for identifying malicious intentions behind phishing websites through screenshot analysis. Our approach leverages the visual understanding capabilities of LLMs combined with specialized RAG to detect up to four primary threat categories: Credential Theft, Financial Fraud, Malware Distribution, and Personal Information Harvesting.

The Fig. 2 presents an overview of the proposed framework. The system employs a hierarchical multi-agent architecture comprising five specialized layers, each with distinct cognitive responsibilities, working in conjunction with domain-specific knowledge bases. The Vision Analysis Agent serves as the perception layer, responsible for extracting raw data from phishing website screenshots. Using vision-language models, it identifies and organizes visual elements including textual content, interface components, page layout, and domain information when available.

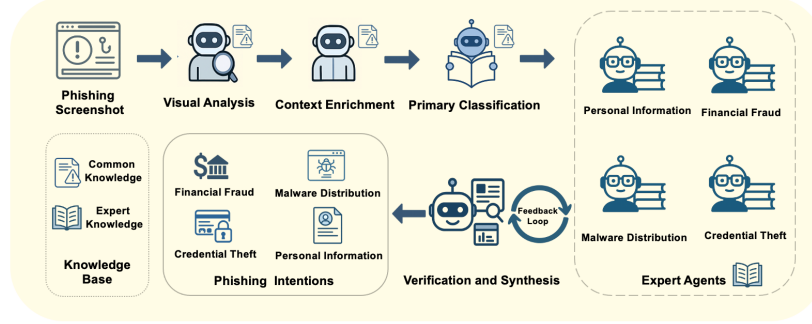


Fig. 2. Proposed PhishIntentionLLM Framework.

The Context Enrichment Agent functions as a semantic layer that enhances extracted elements with security-relevant context. This agent retrieves basic threat patterns from the knowledge base, tags suspicious elements with contextual information, maps visual elements to potential security implications, and generates preliminary threat hypotheses based on established patterns in phishing detection.

The Classification Agent performs multi-label classification to identify the most likely threat categories. It calculates confidence scores for each threat category, selects one to three primary threats for deeper analysis, associates evidence with each identified type, and creates an initial classification hypothesis that guides subsequent specialist analysis. The Specialist Analysis Layer contains four expert agents, each dedicated to a specific threat category. When activated based on initial classification, these specialists conduct in-depth analysis within their domains. The Credential Theft Agent evaluates login form characteristics and domain spoofing patterns, while the Financial Fraud Agent analyzes payment solicitation and unrealistic financial promises. Similarly, the Malware Distribution Agent examines download prompts and software update impersonation, and the Personal Information Agent assesses excessive data collection and privacy policy issues.

The Validation Agent integrates all previous analyses to form a comprehensive assessment. This agent combines specialist findings, resolves potential conflicts between competing hypotheses, weighs evidence based on reliability and relevance, and produces a final classification with complete evidence chains and associated confidence scores.

4.2 Knowledge Retrieval Architecture

PhishIntentionLLM incorporates a dual-layer knowledge architecture, as shown in the Definition 1, that augments agent reasoning through retrieval-augmented generation. The Basic Threat Pattern Repository contains domain-agnostic phishing indicators, including common deception patterns, visual deception elements, suspicious text patterns, and URL red flags. This knowledge supports initial detection and context enrichment phases, providing foundational patterns that transcend specific threat categories.

The Category-Specific Knowledge Repository is organized by threat type and provides detailed domain knowledge. For Credential Theft, it includes common targets, obfuscation techniques, and form submission patterns. The Financial Fraud section contains scam typologies, pressure tactics, and payment anomalies. Malware Distribution knowledge encompasses malware types, download mechanisms, and system access requests, while Personal Information resources detail data collection patterns and privacy indicators. This specialized knowledge enables expert agents to conduct nuanced analyses within their respective domains.

Definition 1 (Knowledge Base Structures). *The PhishIntentionLLM framework employs two complementary knowledge repositories:*

1. Basic Threat Pattern Repository (K_B):

$$K_B = \{P_c, P_v, P_t\} \quad (1)$$

where:

- $P_c = \{p_1, p_2, \dots, p_n\}$ is the set of common phishing patterns
- $P_v = \{v_1, v_2, \dots, v_m\}$ is the set of visual deception techniques
- $P_t = \{t_1, t_2, \dots, t_k\}$ is the set of text-based manipulation patterns

2. Specialist Knowledge Repository (K_C):

For each threat category $c \in \{\text{Credential Theft, Financial Fraud, Malware Distribution, Personal Information Harvesting}\}$, we maintain:

$$K_C^c = \{F^c, D^c\} \quad (2)$$

where:

- $F^c = \{f_1^c, f_2^c, \dots, f_p^c\}$ is the set of primary features for category c
- $D^c = \{T^c, M^c, I^c\}$ is the detailed knowledge structure, containing:
 - $T^c = \{t_1^c, t_2^c, \dots, t_q^c\}$: common targets
 - $M^c = \{m_1^c, m_2^c, \dots, m_r^c\}$: specialized techniques
 - $I^c = \{i_1^c, i_2^c, \dots, i_s^c\}$: distinctive indicators

4.3 Processing Pipeline

The *PhishIntentionLLM* processing workflow begins with image input, where the system ingests website screenshots as primary data. The Visual Analysis phase extracts text, interface elements, and page structure using vision-language capabilities. During Context Enhancement, the extracted elements are enriched

with security context from the basic threat repository, establishing preliminary security implications for observed elements. Initial Classification identifies one to three primary threat categories and calculates confidence scores, determining which specialist agents to activate. In the Specialist Analysis phase, these activated agents perform in-depth examination within their respective threat domains, applying category-specific knowledge to the enriched context. Evidence Synthesis integrates all analyses to produce a cohesive final classification with complete evidence chains. When confidence falls below established thresholds, a feedback loop activates additional specialist analyses to improve certainty. This adaptive mechanism allows the system to handle ambiguous cases by gathering additional perspectives. Finally, Result Generation outputs identified threats with confidence scores and supporting evidence, providing a comprehensive assessment of phishing intentions.

Fig. 3 demonstrates the *PhishIntentionLLM* approach through precise computational stages and decision procedures. The algorithm emphasizes the mathematical relationship between input screenshot analysis and threat categorization using set operations and conditional branches. Notable features include the top-k selection function for candidate threat categories (line 11), the union operation for specialist analysis aggregation (line 19), and the argmax function for determining the highest confidence category when necessary (line 39).

```

Require: Phishing website screenshot image  $I$ 
Require: Basic threat pattern knowledge base  $K_B$ 
Require: Category-specific knowledge base  $K_C$ 
Require: Confidence threshold  $\tau$ 
Ensure: Identified threat categories with evidence and confidence scores
1: function PHISHINTENTIONLLM( $I, K_B, K_C, \tau$ )
2:    $T \leftarrow \emptyset$  ▷ Set of identified threat categories
3:   Layer 1: Vision Analysis
4:    $E \leftarrow \text{VISIONANALYSISAGENT}(I)$  ▷ Extract visual elements
5:   Layer 2: Context Enrichment
6:    $patterns \leftarrow \text{RETRIEVEPATTERNS}(K_B)$  ▷ Get relevant threat patterns
7:    $E_{enriched} \leftarrow \text{CONTEXTENRICHMENTAGENT}(E, patterns)$  ▷ Add security context
8:   Layer 3: Initial Classification
9:    $features \leftarrow \text{RETRIEVECATEGORYFEATURES}(K_C)$  ▷ Get category features
10:   $C, S \leftarrow \text{CLASSIFICATIONAGENT}(E_{enriched}, features)$  ▷  $C$ : categories,  $S$ : scores
11:   $P \leftarrow \{c_i \in C \mid \text{Top-}k(S, k = 3)\}$  ▷ Select top 1-3 categories
12:  Layer 4: Specialist Analysis
13:   $A \leftarrow \emptyset$  ▷ Specialist analysis results
14:  for each category  $c \in P$  do
15:     $knowledge \leftarrow \text{RETRIEVESPECIALISTKNOWLEDGE}(K_C, c)$ 
16:     $A_c \leftarrow \text{SPECIALISTAGENT}(E_{enriched}, knowledge, c)$ 
17:     $A \leftarrow A \cup \{(c, A_c)\}$ 
18:  end for
19:  Layer 5: Validation and Synthesis
20:   $R, conf \leftarrow \text{VALIDATIONAGENT}(E, E_{enriched}, C, S, A)$ 
21:  if  $conf < \tau$  then ▷ Feedback loop for low confidence
22:     $categories \leftarrow \{"Credential Theft", "Financial Fraud", "Malware Distribution", "Personal Information Harvesting"\}$ 
23:    for each category  $c \in categories \setminus P$  do
24:       $knowledge \leftarrow \text{RETRIEVESPECIALISTKNOWLEDGE}(K_C, c)$ 
25:       $A_c \leftarrow \text{SPECIALISTAGENT}(E_{enriched}, knowledge, c)$ 
26:       $A \leftarrow A \cup \{(c, A_c)\}$ 
27:    end for
28:     $R, conf \leftarrow \text{VALIDATIONAGENT}(E, E_{enriched}, C, S, A)$ 
29:  end if
30:  Result Formatting
31:  for each threat category  $t \in R$  do
32:     $evidence \leftarrow \text{Extract evidence from } A \text{ for category } t$ 
33:     $confidence \leftarrow \text{Extract confidence from } A \text{ for category } t$ 
34:    if  $confidence \geq \tau$  then
35:       $T \leftarrow T \cup \{(t, evidence, confidence)\}$ 
36:    end if
37:  end for
38:  if  $|T| = 0$  then ▷ Ensure at least one category is returned
39:     $t^* \leftarrow \arg \max_{t \in R} confidence(t)$ 
40:     $evidence \leftarrow \text{Extract evidence from } A \text{ for category } t^*$ 
41:     $confidence \leftarrow \text{Extract confidence from } A \text{ for category } t^*$ 
42:     $T \leftarrow \{(t^*, evidence, confidence)\}$ 
43:  end if
44:  return  $T$ 
45: end function

```

Fig. 3. The Proposed Algorithm for PhishIntentionLLM Framework.

5 Evaluation

5.1 Model Selection

To evaluate the effectiveness of our framework, we selected four state-of-the-art multimodal LLMs with diverse architectures and capabilities. Qwen2.5-VL-72B-Instruct [25] is a 72B-parameter vision-language model by Alibaba Cloud, pre-trained on multilingual and multimodal datasets and known for strong visual reasoning performance. Gemini-2.0-Flash-001 [26], developed by Google DeepMind, balances efficient inference and advanced multimodal processing. GPT-4o [27], OpenAI's flagship model, integrates high-level visual understanding with robust reasoning, while GPT-4o-mini [28] offers a more lightweight alternative that retains strong multimodal capabilities. All models support API integration, ensuring compatibility with our framework and enabling comparative evaluation across computational and architectural dimensions.

5.2 Ground Truth Dataset

To construct the ground truth dataset for the evaluation, we randomly selected phishing website samples from three existing datasets that contain screenshots for each phishing instance and manually removed the samples with poor quality of screenshots [15,29,30]. The labeling process involved three cybersecurity engineers, each with a minimum of three years of professional experience. Two engineers independently labeled the intentions of these phishing samples based on the screenshots, while the third engineer reviewed all labeled samples to ensure consistency and accuracy. The engineers assigned one or multiple intentions to each phishing sample based on their visual content analysis.

For instance, if a phishing website solely requested username and password credentials, it was labeled as "Credential Theft." However, if the website additionally solicited telephone numbers or address information, it received dual labels of "Credential Theft" and "Personal Information Harvesting."

Through this rigorous labeling process, we constructed a multi-intention phishing dataset comprising 2,063 phishing samples from domains such as e-commerce, finance, social networking, telecommunications, and delivery services. This is one of the first phishing datasets to include explicit phishing intentions. The ground truth dataset has been publicly released.

Table 1. Distribution of Phishing Intentions and Number of Intentions per Record.

Category	Type	Count
Phishing Intention Type	Credentials Theft	1696
	Malware Distribution	68
	Financial Fraud	222
	Personal Information Harvesting	408
Number of Intentions per Record	One Intention	1757
	Two Intentions	281
	Three Intentions	25

Table 1 presents the distribution of phishing intentions identified in ground truth dataset, along with the number of records containing one, two, or three distinct intentions. The majority of records are associated with Credentials Theft (1,696), followed by Personal Information Harvesting (408), Financial Fraud (222), and Malware Distribution (68). Additionally, most records contain only a single phishing intention (1,757), while fewer records exhibit two (281) or three (25) co-occurring intentions. No record contains four intentions.

5.3 Evaluation Metrics

To comprehensively evaluate this framework, various evaluation metrics are used in this study. The overall accuracy measures the proportion of correctly classified website screenshots:

$$\text{Accuracy} = \frac{|\{s \in S : \hat{Y}_s = Y_s\}|}{|S|} \quad (3)$$

where S represents all samples, Y_s the true intention labels, and \hat{Y}_s the predicted labels for sample s .

Since phishing websites often exhibit multiple intentions simultaneously and some malicious phishing intentions (e.g., credential theft) are likely to occur more frequently than others (e.g., malware distribution) which leads to imbalanced classes, we employ micro-averaged metrics that aggregate contributions from all classes and reflect the actual data distribution, which is crucial in real-world scenarios:

$$\text{Precision}_{\text{micro}} = \frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C (TP_c + FP_c)} \quad (4)$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C (TP_c + FN_c)} \quad (5)$$

$$\text{F1}_{\text{micro}} = \frac{2 \times \text{Precision}_{\text{micro}} \times \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}} \quad (6)$$

where $C = 4$ represents our intention classes, with TP_c , FP_c , and FN_c denoting true positives, false positives, and false negatives for class c , respectively.

We also introduce Accuracy by Complexity (Acc_{comp}) to address the nuanced nature of multi-intention phishing websites:

$$\text{Acc}_{\text{comp}}(k) = \frac{|\{s \in S_k : \text{match}(Y_s, \hat{Y}_s) \geq t_k\}|}{|S_k|} \quad (7)$$

where S_k represents the set of samples with exactly k intentions, $\text{match}(Y_s, \hat{Y}_s)$ counts the number of correctly matched intentions, and t_k is the threshold for samples with k intentions, defined as:

$$t_k = \begin{cases} 1, & \text{if } k = 1 \\ 1, & \text{if } k = 2 \\ 2, & \text{if } k = 3 \end{cases} \quad (8)$$

This lies in the fact that for many phishing websites, even experienced cybersecurity experts can have slightly different understandings of the underlying intentions. Therefore, we believe partial matches provide valuable insights when evaluating multi-intention scenarios. We applied this to all above evaluations to ensure fair comparison against selected models.

For individual intention analysis, we calculate standard metrics for each class c in our four phishing intention categories:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (9)$$

$$F1_c = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad \text{Accuracy}_c = \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c} \quad (10)$$

where TN_c represents samples correctly identified as not belonging to class c .

6 Results

6.1 Model Performance

Table 2 and Fig. 4 present the general performance metrics of different LLMs integrated into our *PhishIntentionLLM* framework. GPT-4o achieved the highest precision (0.7895) among all evaluated models, making it particularly valuable for phishing intention detection. High precision in this context means the model correctly identifies specific phishing intentions with minimal misclassifications, providing security analysts with more accurate understanding of attackers' objectives.

Table 2. General Performance Metrics of Selected LLMs with PhishIntentionLLM.

Model	Precision _{micro}	Recall _{micro}	F1 _{micro}	Accuracy _{micro}
GPT-4o	0.7895	0.8544	0.8207	0.8915
Gemini 2.0	0.7843	0.8976	0.8371	0.8987
GPT-4o-mini	0.6149	0.9744	0.7540	0.8149
Qwen2.5-VL-72B	0.4520	0.9428	0.6111	0.6518

While GPT-4o excels in precision, it maintains strong performance across other metrics with a micro-averaged F1 score of 0.8207 and accuracy of 0.8915. This balanced performance demonstrates GPT-4o's effectiveness as a foundation model for identifying the underlying intentions of phishing campaigns. Gemini 2.0 follows closely with slightly lower precision (0.7843) but higher recall (0.8976), resulting in the highest overall F1 score (0.8371) and accuracy (0.8987).

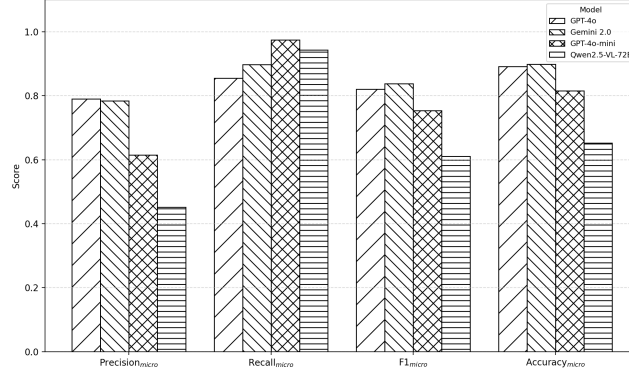


Fig. 4. Comparison of Selected LLMs on General Performance Metrics.

This indicates Gemini 2.0 identifies more actual phishing intentions at the expense of slightly more misattributed intentions compared to GPT-4o. GPT-4o-mini and Qwen2.5-VL-72B demonstrate significantly different performance characteristics, with extremely high recall values (0.9744 and 0.9428 respectively) but considerably lower precision scores (0.6149 and 0.4520). This suggests these models excel at capturing all potential intentions behind phishing websites but more frequently assign incorrect intentions, potentially complicating threat intelligence and response prioritization.

Table 3 provides insights into model performance across different phishing intention complexity levels. Interestingly, the performance ranking shifts when evaluated using Accuracy by Complexity (Acc_{comp}) metrics.

Table 3. Comparison of Selected LLMs on (Acc_{comp}) and Overall Accuracy

Model	Acc_{comp} (1 Intention)	Acc_{comp} (2 Intentions)	Acc_{comp} (3 Intentions)	Overall Accuracy
GPT-4o-mini	0.8980	1.0000	0.9600	0.9130
GPT-4o	0.8996	0.9324	0.8800	0.9039
Qwen2.5-VL-72B	0.8953	0.9644	0.8800	0.9045
Gemini 2.0	0.8750	0.9927	0.8800	0.8910

GPT-4o-mini demonstrated the highest overall accuracy (0.9130) across all complexity levels. The model achieved strong performance on single-intention phishing sites (0.8980) and notably high scores for multi-intention websites (1.0000 for two intentions and 0.9600 for three intentions).

GPT-4o maintains strong performance across complexity levels with high scores for single-intention phishing sites (0.8996), two-intention sites (0.9324), and three-intention scenarios (0.8800). This consistent performance further validates its reliability for accurately identifying diverse phishing strategies.

The Acc_{comp} metrics for websites with three intentions show consistent performance across most models (0.8800), with only GPT-4o-mini achieving a higher score (0.9600). This suggests that the *PhishIntentionLLM* framework provides

reliable intention identification even in the most complex phishing scenarios with multiple malicious objectives.

These results demonstrate that while GPT-4o provides the most precise identification of specific phishing intentions, all evaluated models show strong capabilities in identifying phishing intentions across various complexity levels when integrated with our *PhishIntentionLLM* framework.

6.2 PhishIntentionLLM vs. Single-Agent

To further validate the effectiveness of our proposed framework, we conducted a comparative analysis between *PhishIntentionLLM* and a single-agent baseline using identical foundation model (Gemini 2.0). As shown in Table 4 and Fig. 5, the multi-agent RAG approach substantially outperforms the single-agent scenario across nearly all metrics.

Table 4. Comparison of PhishIntentionLLM and Single-Agent Scenario

Metric	PhishIntentionLLM (Gemini 2.0)	Single-Agent (Gemini 2.0)
Precision _{micro}	0.7843	0.4014
Recall _{micro}	0.8976	0.6341
F1 _{micro}	0.8371	0.4916
Accuracy _{micro}	0.8987	0.6195
Overall Accuracy	0.8910	0.5870
Acc _{comp} (1 Intention)	0.8750	0.5287
Acc _{comp} (2 Intentions)	0.9927	0.9253
Acc _{comp} (3 Intentions)	0.8800	0.8800

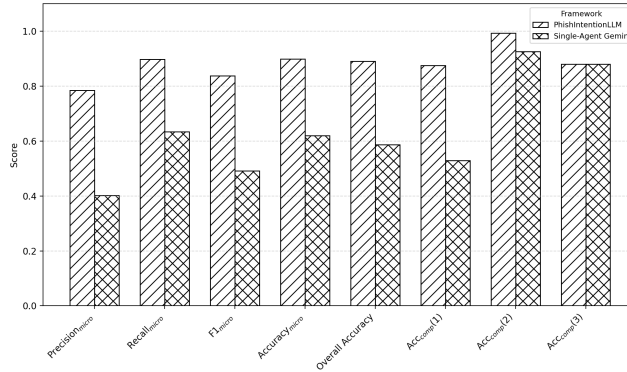


Fig. 5. Comparison of PhishIntentionLLM and Single-Agent Scenario.

The most striking improvement appears in precision, where *PhishIntentionLLM* achieves 0.7843 compared to just 0.4014 for the single-agent approach which represents a ~95% improvement. This dramatic enhancement in precision demonstrates that our hierarchical agent architecture with specialized knowledge bases significantly reduces false positive classifications, enabling much more accurate identification of specific phishing intentions.

Similarly, recall improved from 0.6341 to 0.8976 ($\sim 42\%$ increase), indicating the multi-agent system’s superior ability to identify all relevant intentions present in phishing websites. The combined improvements in precision and recall culminate in a $\sim 70\%$ increase in F1 score (0.8371 vs. 0.4916) and a $\sim 45\%$ enhancement in micro-accuracy (0.8987 vs. 0.6195).

The Accuracy by Complexity (Acc_{comp}) metrics reveal that our approach demonstrates particularly high effectiveness against phishing websites with single and two malicious intentions. For single-intention websites, *PhishIntentionLLM* achieves 0.8750 accuracy compared to 0.5287 for the single-agent approach with a $\sim 65\%$ improvement.

6.3 PhishIntentionLLM vs. PhishIntention

To benchmark our framework against existing methods, we compared *PhishIntentionLLM* (GPT-4o) with PhishIntention [22], the only prior work focused on phishing intention analysis. Since PhishIntention is limited to detecting credential theft intentions only, we conducted a focused comparison on this specific intention type regarding test precision, test accuracy, F1 and recall, following PhishIntention’s methodology with the same 9:1 (training:testing) ratio on our ground truth dataset.

As shown in Table 5 and Fig. 6, *PhishIntentionLLM* demonstrates superior performance across all metrics in credential theft detection. Our framework achieved a precision of 0.8545 compared to PhishIntention’s 0.8206, representing a $\sim 4\%$ improvement in correctly identifying genuine credential theft attempts while reducing false positives.

Table 5. Comparison of PhishIntention and PhishIntentionLLM

Model	Precision	Recall	F1 Score	Accuracy
PhishIntention [22]	0.8206	0.7471	0.7821	0.6578
PhishIntentionLLM (Ours)	0.8545	0.9946	0.9193	0.8602

The most dramatic improvement appears in recall, where *PhishIntentionLLM* achieved 0.9946 compared to PhishIntention’s 0.7471 with a $\sim 33\%$ increase. This substantial enhancement in recall indicates that our multi-agent RAG framework can identify virtually all credential theft attempts (99.5%), whereas the existing approach misses approximately one-quarter of such phishing attempts.

The combined improvements in precision and recall result in a significant enhancement in F1 score (0.9193 vs. 0.7821), representing an $\sim 18\%$ increase over the existing approach. Additionally, overall accuracy improved from 0.6578 to 0.8602, a $\sim 31\%$ enhancement that demonstrates the superior classification capabilities of our framework. PhishIntentionLLM outperforms PhishIntention by fusing visual-text understanding with RAG-driven knowledge, unlike the latter’s static visual approach.

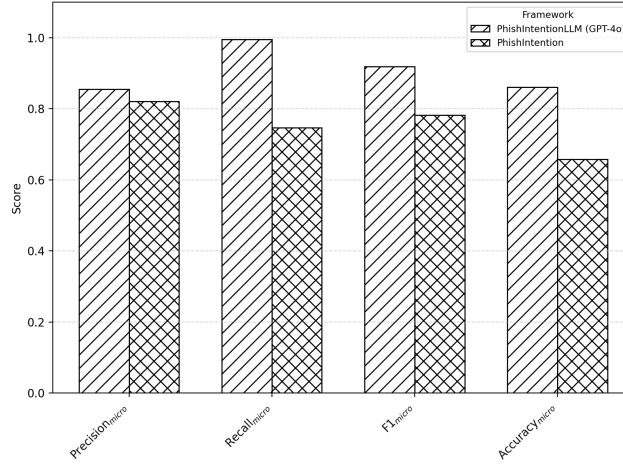


Fig. 6. Comparison of PhishIntention and PhishIntentionLLM on Credentials Theft Intention.

7 Discussion

To understand the phishing intentions in a larger scale, we use our proposed framework with GPT-4o to further evaluate 6K more phishing samples to profile their phishing intentions which results in ~9K samples including our ground truth dataset. A sector-intention frequency matrix was constructed (see Fig. 7(a)). The matrix maps the occurrence of four major phishing intentions: Credentials Theft, Financial Fraud, Malware Distribution, and Personal Information Harvesting—across identified sectors such as financial, e-commerce, telecommunications, and government. The analysis reveals that the financial sector is the most frequently targeted, with 2,882 instances of credentials theft and 2,032 cases of personal information harvesting. Other highly targeted sectors include online/cloud service, email provider, and social networking. Notably, credentials theft and personal information harvesting followed by financial fraud are the dominant intentions across most sectors, while malware distribution remains relatively less frequent.

We further examined records containing multiple phishing intentions to explore the strategic complexity embedded within these campaigns. Fig. 7(b) shows the frequency and sectoral distribution of co-occurring phishing intentions found in samples with either two or three distinct objectives. Among the two-intention combinations, Credentials Theft + Personal Information Harvesting was by far the most prevalent, appearing in 3,368 instances, with the financial sector being the most frequently targeted (1,185 cases). This combination highlights a dual objective wherein adversaries aim not only to compromise login credentials but also to capture accompanying personal data, thereby increasing the potential for downstream exploitation. The next most common two-intention pair, Credentials Theft + Financial Fraud, appeared in 141 instances, again dominated by

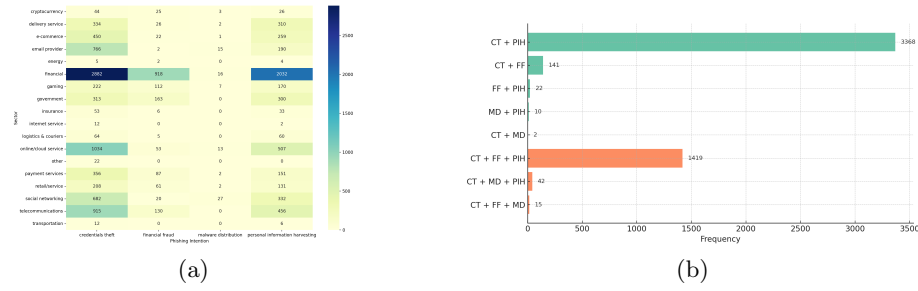


Fig. 7. (a) shows phishing intentions across different sectors; (b) displays frequencies of two- and three-intention combinations. Two-intention combinations are in green; three-intention combinations are in orange. **Note:** CT = Credentials Theft, PIH = Personal Information Harvesting, FF = Financial Fraud, MD = Malware Distribution.

attacks on the financial sector (91 cases), suggesting a strong correlation between credential compromise and direct financial gain.

In records exhibiting three distinct phishing intentions, the most frequent combination was Credentials Theft + Financial Fraud + Personal Information Harvesting, found in 1,419 samples, with the financial sector again serving as the primary target (821 occurrences). This triad reflects the layered objectives of modern phishing schemes, where attackers simultaneously seek unauthorized access, financial exploitation, and user profiling. Less frequent but notable three-intention combinations included Credentials Theft + Malware Distribution + Personal Information Harvesting (42 instances, led by the social networking sector) and Credentials Theft + Financial Fraud + Malware Distribution (15 instances, with the gaming sector most affected), demonstrating that more complex phishing strategies often align with the sector-specific threat surface and user value.

The prevalence of such multi-intention combinations likely reflects an adaptive response by attackers to increasingly robust verification mechanisms used by modern systems. As single data points such as passwords or email addresses are often insufficient to gain full access to target systems—particularly those with multi-factor authentication or behavioral risk scoring—phishing campaigns have evolved to collect multiple types of sensitive information concurrently. This multi-vector approach significantly enhances the likelihood of bypassing layered defenses and achieving the attackers’ ultimate goals.

8 Conclusion

This study presents *PhishIntentionLLM*, a novel multi-agent RAG framework for uncovering phishing website intentions through screenshot analysis. While our results demonstrate strong performance and scalability, future work can explore real-time deployment to profile phishing intentions in the wild. Additionally, the proposed approach holds potential for broader application, such as recognizing attacker intentions in phishing emails and other social engineering vectors.

References

1. Alkhalil, Z., Hewage, C., Nawaf, L., Khan, I.: Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science* **Volume 3 - 2021** (2021). <https://doi.org/10.3389/fcomp.2021.563060>
2. Khonji, M., Iraqi, Y., Jones, A.: Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials* **15**(4), 2091–2121 (2013). <https://doi.org/10.1109/SURV.2013.032213.00009>
3. Lim, K., Park, J., Kim, D.: Phishing vs. legit: Comparative analysis of client-side resources of phishing and target brand websites. In: *Proceedings of the ACM Web Conference 2024*. p. 1756–1767. WWW '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3589334.3645535>
4. Anti-Phishing Working Group (APWG): <https://apwg.org/>
5. LI, W., LAGHARI, S.U.A., MANICKAM, S., CHONG, Y.W.: Exploration and evaluation of human-centric cloaking techniques in phishing websites. *KSII Transactions on Internet and Information Systems* **19**(1), 232–258 (January 2025). <https://doi.org/10.3837/tiis.2025.01.011>
6. Li, W., Manickam, S., Laghari, S.U.A., Chong, Y.W.: Uncovering the cloak: A systematic review of techniques used to conceal phishing websites. *IEEE Access* **11**, 71925–71939 (2023). <https://doi.org/10.1109/ACCESS.2023.3293063>
7. Zieni, R., Massari, L., Calzarossa, M.C.: Phishing or not phishing? a survey on the detection of phishing websites. *IEEE Access* **11**, 18499–18519 (2023). <https://doi.org/10.1109/ACCESS.2023.3247135>
8. Li, W., Manickam, S., Chong, Y.W., Leng, W., Nanda, P.: A state-of-the-art review on phishing website detection techniques. *IEEE Access* **12**, 187976–188012 (2024). <https://doi.org/10.1109/ACCESS.2024.3514972>
9. Kulkarni, A., Balachandran, V., Das, T.: Phishing webpage detection: Unveiling the threat landscape and investigating detection techniques. *IEEE Communications Surveys & Tutorials* **27**(2), 974–1007 (2025). <https://doi.org/10.1109/COMST.2024.3441752>
10. Nguyen, L.A.T., To, B.L., Nguyen, H.K., Nguyen, M.H.: A novel approach for phishing detection using url-based heuristic. In: *2014 International Conference on Computing, Management and Telecommunications (ComManTel)*. pp. 298–303 (2014). <https://doi.org/10.1109/ComManTel.2014.6825621>
11. Sahingoz, O.K., Buber, E., Demir, O., Diri, B.: Machine learning based phishing detection from urls. *Expert Systems with Applications* **117**, 345–357 (2019). <https://doi.org/10.1016/j.eswa.2018.09.029>
12. Opara, C., Wei, B., Chen, Y.: Htmlphish: Enabling phishing web page detection by applying deep learning techniques on html analysis. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–8 (2020)
13. Shirazi, H., Bezawada, B., Ray, I.: "kn0w thy domaIn name": Unbiased phishing detection using domain name based features. In: *Proceedings of the 23rd ACM on Symposium on Access Control Models and Technologies*. p. 69–75. SACMAT '18, Association for Computing Machinery, New York, NY, USA (2018)
14. Aljofey, A., Bello, S.A., Lu, J., Xu, C.: Comprehensive phishing detection: A multi-channel approach with variants tcn fusion leveraging url and html features. *Journal of Network and Computer Applications* **238**, 104170 (2025). <https://doi.org/10.1016/j.jnca.2025.104170>
15. Lin, Y., Liu, R., Divakaran, D.M., Ng, J.Y., Chan, Q.Z., Lu, Y., Si, Y., Zhang, F., Dong, J.S.: Phishpedia: A hybrid deep learning based approach to visually identify

- phishing webpages. In: 30th USENIX Security Symposium (USENIX Security 21). pp. 3793–3810. USENIX Association (Aug 2021)
16. Liu, D.J., Lee, J.H.: A cnn-based sia screenshot method to visually identify phishing websites. *Journal of Network and Systems Management* **32**(1), 8 (2024)
 17. Koide, T., Nakano, H., Chiba, D.: Chatphishdetector: Detecting phishing sites using large language models. *IEEE Access* **12**, 154381–154400 (2024)
 18. Kassa, Y.W., James, J.I., Belay, E.G.: Cybercrime intention recognition: A systematic literature review. *Information* **15**(5) (2024). <https://doi.org/10.3390/info15050263>
 19. Stojnic, T., Vatsalan, D., Arachchilage, N.A.G.: Phishing email strategies: Understanding cybercriminals’ strategies of crafting phishing emails. *SECURITY AND PRIVACY* **4**(5), e165 (2021). <https://doi.org/10.1002/spy2.165>
 20. Walton, B.J., Khatun, M.E., Ghawaly, J.M., Ali-Gombe, A.: Exploring large language models for semantic analysis and categorization of android malware. In: 2024 Annual Computer Security Applications Conference Workshops (ACSAC Workshops). pp. 248–254 (2024). <https://doi.org/10.1109/ACSACW65225.2024.00035>
 21. Kolupuri, S.V.J., Paul, A., Bhowmick, R.S., Ganguli, I.: Scams and frauds in the digital age: ML-based detection and prevention strategies. In: Proceedings of the 26th International Conference on Distributed Computing and Networking. p. 340–345. ICDCN ’25, Association for Computing Machinery, New York, NY, USA (2025). <https://doi.org/10.1145/3700838.3703672>
 22. Liu, R., Lin, Y., Yang, X., Ng, S.H., Divakaran, D.M., Dong, J.S.: Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 1633–1650. USENIX Association, Boston, MA (Aug 2022)
 23. Rahman, M., Piryani, K.O., Sanchez, A.M., Munikoti, S., De La Torre, L., Levin, M.S., Akbar, M., Hossain, M., Hasan, M., Halappanavar, M.: Retrieval augmented generation for robust cyber defense. Tech. rep., Pacific Northwest National Laboratory (PNNL), Richland, WA (United States) (09 2024). <https://doi.org/10.2172/2474934>
 24. Arikkat, D.R., M., A., Binu, N., M., P., Biju, N., Arunima, K.S., P, V., Rehiman K. A., R., Conti, M.: Intellbot: Retrieval augmented llm chatbot for cyber threat knowledge delivery. In: 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN). pp. 644–651 (2024)
 25. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
 26. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
 27. OpenAI Team: Hello gpt-4o. Tech. rep., OpenAI (05 2024), <https://openai.com/index/hello-gpt-4o/>
 28. OpenAI Team: Gpt-4o mini: advancing cost-efficient intelligence. Tech. rep., OpenAI (07 2024), <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>
 29. Putra, I.K.A.A.: Phishing website dataset (Jul 2023). <https://doi.org/10.5281/zenodo.8041387>
 30. Dalgic, F., Bozkir, A., Aydos, M.: Phish-iris: A new approach for vision based brand prediction of phishing web pages via compact visual descriptors. In: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). pp. 1–8 (2018). <https://doi.org/10.1109/ISMSIT.2018.8567299>