A Taxonomy and Multi-Layered Defense Framework for Generative AI-Powered Phishing Campaigns on Social Media Platforms

Cyber Security Engineering
Master's Degree Programme in Information and Communication Technology
Department of Computing, Faculty of Technology
Master of Science in Technology Thesis

Author:
Tanzina Akter

Supervisors:
Seppo Virtanen (University of Turku)
Ismayil Hasanov (University of Turku)

July 2025

Master of Science in Technology Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Cyber Security
Programme: Master's Degree Programme in Information and Communication Technology
Author: Tanzina Akter
Title: A Taxonomy and Multi-Layered Defense Framework for Generative AI-Powered Phishing Campaigns on Social Media Platforms
Number of pages: 73 pages
Date: July 2025

Abstract

This research aims to analyze the increasing threat of generative AI-powered phishing campaigns on social media platforms. Social media platforms are the prime targets for attackers because it has a huge amount of user information. Attackers try to mislead users and steal information by utilizing advanced AI technologies, such as Large Language Models (LLMs), deepfakes, and automated bots. The primary goal of this research is to design a structured taxonomy and multi-layered defense framework. The proposed taxonomy helps to classify the attacks, and the defense framework will detect and mitigate phishing. The study adopts a mixed-method approach to validate the proposed taxonomy. It includes a literature review to identify existing gaps, a survey to understand user awareness, and interviews with experts to collect real-world experiences. Thematic analysis of interview data is done by using Braun and Clarke's six-phase framework to extract recurring patterns and validate the proposed taxonomy.

This study examines the five core dimensions of the taxonomy, such as generative modality, phishing vector, platform feature exploited, target profile, and delivery/automation pattern. This dimension helps to understand how the phishing content is created, how it is delivered, which platform features are exploited, who is being targeted, and how automation is used. It helps researchers, professionals, and security teams to identify patterns, understand attacker behavior, and adapt response strategies. The validation of the dimension is done by describing recent cases, a social media users survey, and conducting interviews. The results reveal that AI-generated phishing campaigns are difficult to detect due to their fluency, personalization, and use of trusted channels. Based on these findings, a defense framework is introduced that consists of four interconnected layers: Integration, Detection, Response, and Feedback & Learning.

The study concludes that existing solutions are becoming outdated due to the fast growth of AI-generated phishing. The research emphasizes the need for better user education and stronger platform controls to stay updated. It proposes to create an open-access community resource based on the taxonomy and defense framework as a key recommendation. This resource would help users, researchers, and cyber security professionals to share real experiences, track new phishing techniques, and keep the framework updated. Overall, this research provides a strong foundation for improving protection against AI-generated phishing threats.

Keywords: AI-powered phishing, social media, generative Artificial Intelligence, phishing taxonomy, multi-layered defense, thematic analysis, taxonomy dimensions, human factor in cyber security

Table of contents

1 Introduction 1

1.1 Problem Statement 4

1.2 Research Purpose 5

1.3 Research Objectives 6

2 Literature Review 7

2.1 Phishing Attack 7

2.1.1 AI-based Phishing Attacks 7

2.1.2 AI-based Phishing Attacks on Social Media 10

2.2 Taxonomy in Cyber security 11

2.3 Existing Taxonomies of Phishing Attacks 12

2.4 Existing Defense Framework and Countermeasures 15

2.5 Gaps in existing Taxonomies 16

3 Research Methodology 18

3.1 Research Design 18

3.2 Data Collection 19

3.3 Survey Design 19

3.3.1 Survey Structure 20

3.3.2 Survey Collection and Tools 20

3.3.3 Survey Questionnaire 21

3.4 Interview Design 26

3.4.1 Interview Purpose 27

3.4.2 Interview Format 27

3.4.3 Interview Questionnaires 28

3.4.4 Interview Analysis 29

3.5 Case Study Selection 31

3.6 Taxonomy Development Process 31

3.7 Defense Framework Development Process 32

4 Taxonomy design for AI-based Phishing Attacks on social media 34

4.1 Proposed Taxonomy Design 34

4.1.1 Dimension 1: Generative Model Modality 36

4.1.2 Dimension 2: Phishing Vector 37
4.1.3 Dimension 3: Social Platform Feature Exploited 37
4.1.4 Dimension 4: Target Profile 38
4.1.5 Dimension 5: Delivery & Automation Pattern 39
4.2 Taxonomy Validation 40
4.2.1 Validation Through Case Studies 40
4.2.2 Validation Through Survey Data 44
4.2.3 Validation Through Expert Interviews 46
5 Defense Framework Architecture 48
5.1 AI-powered Phishing Tools and Techniques 48
5.2 Design Objectives and Requirements 49
5.3 Framework Design 50
5.3.1 Integration Layer 52
5.3.2 Detection Layer 53
5.3.3 Response & Mitigation Layer 54
5.3.4 Feedback & Learning Loop 54
6 Results and Discussion 56
6.1 Survey Findings 56
6.2 Interview Insights 63
6.2.1 Participant 1 (Cyber Security Specialist) 63
6.2.2 Participant 2 (Engineer and Instagram Content Creator) 63
6.2.3 Participant 3 (A General User) 64
6.2.4 Participant 4 (Cyber security Student) 64
6.2.5 Participant 5 (Facebook Content Creator) 64
7 Conclusion 66
7.1 Summary of the Key Contributions 66
7.2 Proposal for a Community Resource 67
7.3 Future Research 69
References 70

List of Figures

FIGURE 1 PHISHING ATTACK WORKFLOW 1
FIGURE 2 PERCENTAGE OF USERS AGED 18–34 (OR 18–44 FOR FACEBOOK) ON SOCIAL MEDIA PLATFORM 2
FIGURE 3 FOUR PILLARS OF THE WORMGPT TOOL 8
FIGURE 4 THE IMPACT OF BEHAVIORAL PHISHING TRAINING OVER TIME 9
FIGURE 5 PERCENTAGE BREAKDOWN OF AI-GENERATED PHISHING METHODS 10
FIGURE 6 TAXONOMY DESIGN OF SOCIAL ENGINEERING ATTACKS 13
FIGURE 7 SIX PHASES OF EMAIL-BASED PHISHING ATTACKS 14
FIGURE 8 THE RESEARCH METHODOLOGY OVERVIEW 18
FIGURE 9 DISCLAIMER OF THE PARTICIPANTS' RIGHTS 21
FIGURE 10 DEMOGRAPHIC DATA COLLECTION 22
FIGURE 11 USERS TECHNICAL EXPERTISE 22
FIGURE 12 DURATION OF USERS' ONLINE ACTIVITY 23
FIGURE 13 PLATFORM-BASED DATA COLLECTION 23
FIGURE 14 AI TOOLS LITERACY 24
FIGURE 15 USERS' BELIEFS REGARDING AI-GENERATED CONTENT 24
FIGURE 16 USERS' EXPERIENCE WITH PHISHING ATTACKS 25
FIGURE 17 DATA COLLECTION REGARDING DIFFERENT SOCIAL MEDIA FEATURES 25
FIGURE 18 USER´S CONFIDENCE LEVEL DETECTION 26
FIGURE 19 THE POSSIBILITY OF CLICKING ON MALICIOUS CONTENT 26
FIGURE 20 THE TAXONOMY DEVELOPMENT PROCESS OF AI-BASED PHISHING ATTACKS ON SOCIAL MEDIA 32
FIGURE 21 THE DEVELOPMENT PROCESS OF THE DEFENSE FRAMEWORK 33
FIGURE 22 THE PROPOSED TAXONOMY OF AI-POWERED PHISHING ATTACKS ON SOCIAL MEDIA 35
FIGURE 23 MULTI-LAYERED DEFENSE FRAMEWORK ARCHITECTURE 51
FIGURE 24 THE DETECTION PROCESS OF AI CONTENT AS A PART OF A DEFENSE FRAMEWORK 52
FIGURE 25 DISTRIBUTION OF RESPONDENTS BY AGE GROUP 56
FIGURE 26 RESPONDENTS’ TECHNICAL EXPERTISE LEVELS 57
FIGURE 27 AMOUNT OF TIME SPENT ON SOCIAL MEDIA PLATFORMS DAILY 57
FIGURE 28 MOST FREQUENTLY USED SOCIAL MEDIA PLATFORMS 58
FIGURE 29 AWARENESS OF GENERATIVE AI TOOLS AMONG RESPONDENTS 59
FIGURE 30 USERS’ VIEWS ON GENERATIVE AI’S CONTENT FOR PHISHING 59
FIGURE 31 FREQUENCY OF FACING MALICIOUS CONTENT 60
FIGURE 32 MOST COMMON FORMS OF SUSPECTED AI-DRIVEN PHISHING CONTENT 61
FIGURE 33 CONFIDENCE LEVELS IN DISTINGUISHING AI-GENERATED MESSAGES 61
FIGURE 34 THE POSSIBILITY OF CLICKING SUSPICIOUS MESSAGES FROM A PERSON PRETENDING TO BE FAMILIAR 62
FIGURE 35 PROPOSAL FOR A COMMUNITY RESOURCE 68

List of Tables

TABLE 1 DIFFERENT TYPES OF PHISHING 7
TABLE 2 GAPS IN EXISTING TAXONOMIES AND DEFENSE FRAMEWORKS 17
TABLE 3 INTERVIEW PARTICIPANTS DETAILS 28
TABLE 4 THEMATIC ANALYSIS OF INTERVIEW DATA 30
TABLE 5 FEATURES AND EXPLANATION OF DIFFERENT GENERATIVE MODALITIES 36
TABLE 6 DIFFERENT PHISHING VECTORS 37
TABLE 7 DIFFERENT MEDIA FOR SENDING MALICIOUS CONTENT 38
TABLE 8 DIFFERENT TARGET PROFILES BY THE ATTACKERS 39
TABLE 9 DELIVERY AND AUTOMATION PATTERN OF PHISHING ATTACKS 39
TABLE 10 DIMENSION ANALYSIS OF AI-POWERED ROMANCE SCAMS 41
TABLE 11 DIMENSION ANALYSIS OF DEEPFAKE SCAMS 41
TABLE 12 DIMENSION ANALYSIS OF SOCIAL MEDIA BOTS 42
TABLE 13 DIMENSION ANALYSIS OF E-MAIL AND ADS PHISHING 43
TABLE 14 DIMENSION ANALYSIS OF AI CHATBOTS 43
TABLE 15 DIMENSION ANALYSIS OF PUMP-AND-DUMP VIA ASTROTURFING 44
TABLE 16 THE VALIDATION OF THE PROPOSED TAXONOMY THROUGH SURVEY QUESTIONS 45
TABLE 17 VALIDATION OF THE PROPOSED TAXONOMY BY CONDUCTING EXPERT INTERVIEWS 46
TABLE 18 AI-BASED TOOLS AND TECHNIQUES THAT ARE USED IN DESIGNING PHISHING CONTENT 48

1 Introduction

In the era of digital technology, phishing has become a rapidly evolving cyber threat. With the rise of Artificial Intelligence (AI), these attacks are becoming increasingly sophisticated and more difficult to detect. AI offers both significant opportunities and major challenges. Cybercriminals use it to launch more advanced cyberattacks, but it also enhances security methods to detect them. AI-generated phishing is more realistic, adaptive, and convincing than traditional phishing methods. Traditional phishing depends on manually created content. The malicious contents are sent via email with minimal personalization [1]. Its scale is limited, and it is easier to detect due to low language quality. On the other hand, AI-based phishing utilizes tools such as ChatGPT, deepfake audio, and video technology to create highly customized content that mimics human behavior. Attackers can use this content to target multiple individuals across various social media platforms. AI-generated phishing content is challenging to detect due to its sophisticated language, automation, and advanced AI techniques. Figure 1 shows the process of phishing campaigns.

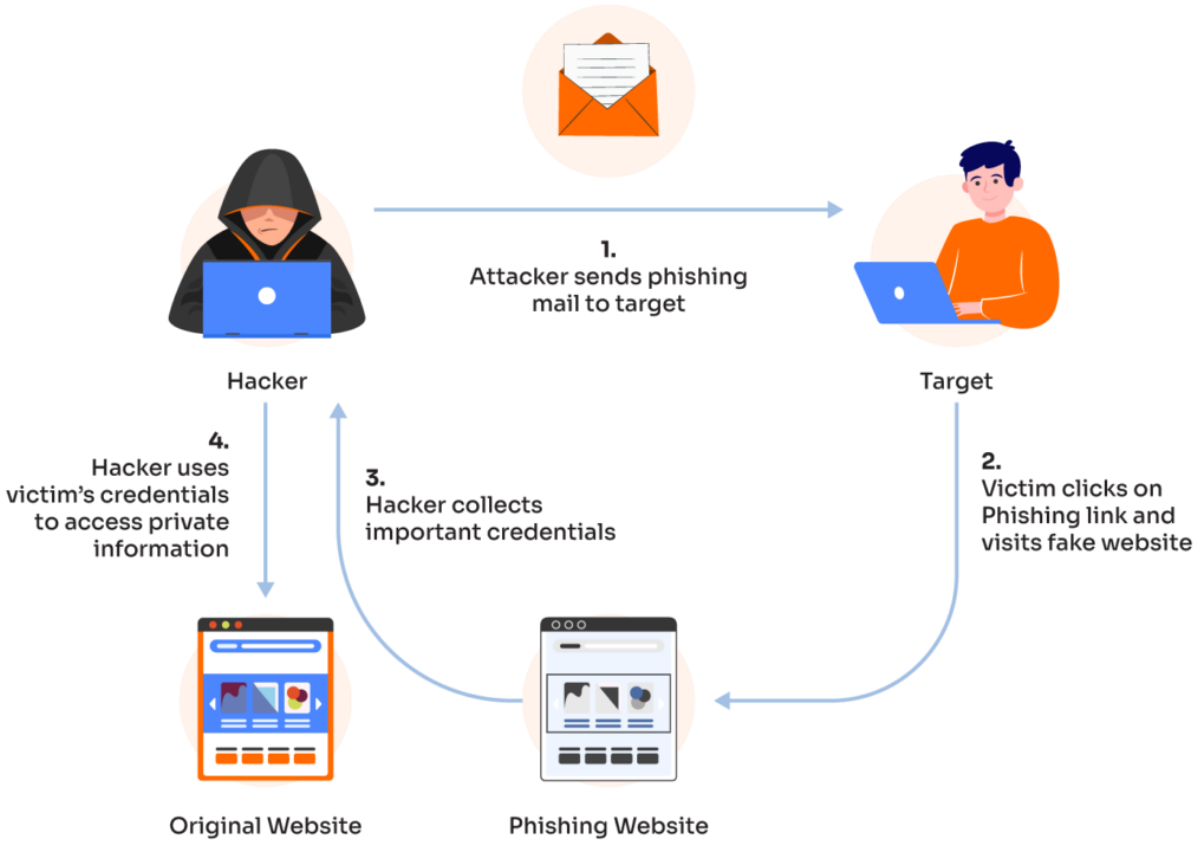

Figure 1 Phishing attack workflow

Social media has connected with over half of the world's population since its launch in 1996. Between 2010 and January 2025, the number of users on social media platforms increased from 970 million to 5.24 billion [3]. As social media has become an important part of society, attackers target social media users to attack them using different AI tools [4]. Previously, attackers sent malicious content via email to many users at a time. But, as social media became more popular, they took advantage of different social media features like direct messages, public posts, and ads to mislead users [1]. Figure 2 illustrates the percentage of young users of Instagram, TikTok, Facebook, and X (formerly Twitter) [6] [7] [8]. 68.3% of users aged 18–34 use TikTok, and Facebook has 70.8% followers aged 18–44. The result shows continued relevance among both young and middle-aged users. Instagram has 61.1% in the 18–34 group, which confirms its popularity among younger users. X has the lowest rate, with only 37% aged 27–42, which suggests a more age-diverse or older user base.

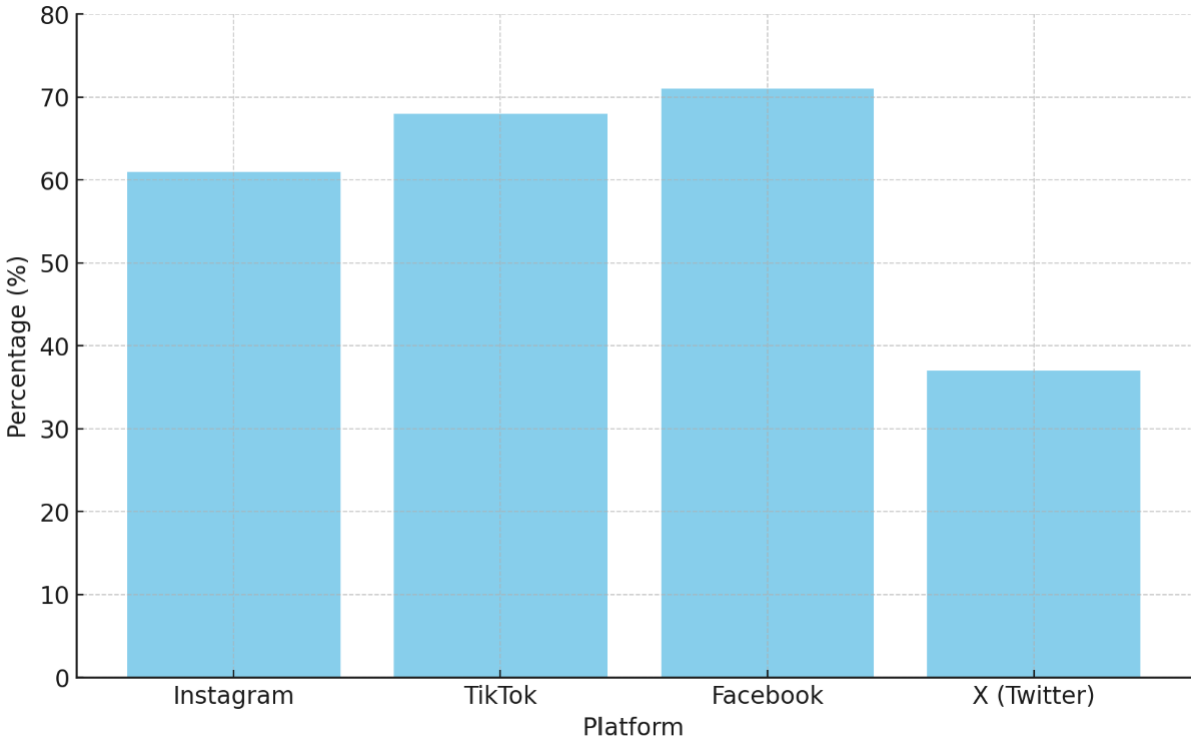

Figure 2 Percentage of users aged 18–34 (Or 18–44 For Facebook) on social media platform

The main targets for AI-based phishing attacks on social media are TikTok, Facebook, and Instagram. Young adults are more active on social media and more likely to engage with messages and content. They are less aware of suspicious links or profiles, which makes it easy for attackers to conduct a successful phishing campaign. As people share their personal information on social media, it becomes easier for attackers to create personalized content or

fake profiles using the shared data. There are some common techniques used by attackers to steal information from users, such as fake profiles, urgent alerts, malicious links in posts or comments, malicious quizzes or surveys, and phishing via direct messages. These scams are designed to mislead users into revealing sensitive information or clicking harmful links. Scammers improve the success rate of the campaigns by using social engineering techniques. These techniques are used to pretend to be someone familiar, such as a friend, influencers, or trusted businesses.

One of the main reasons behind the increasing number of AI-driven phishing attacks is human errors rather than technological failures. Many social media users remain unaware of the risks. The users frequently ignore basic safety practices such as verifying friend requests or avoiding suspicious links. The need for a structured taxonomy and a dedicated defense framework becomes crucial. A taxonomy gives a systematic and structured understanding of AI-generated phishing attacks. This helps to classify phishing attacks based on how they are generated, delivered, and who they target. The main core dimensions of phishing attacks are generative model modality, phishing vector, social platform feature exploited, target profile, and delivery and automation pattern. In addition, a defense framework with multi-layered protection strategies can combine user education, automated detection, and platform-level safety.

In this research, a taxonomy and defense framework for AI-driven phishing attacks on social media platforms is proposed to mitigate the evolving cyber threats. The proposed model focused on five core dimensions, such as generative model modality, phishing vector, social platform feature exploited, target profile, and delivery and automation pattern. In the next step, all the dimension is properly described with the necessary tables. The generative model modality helps to identify AI techniques, such as LLMs, deepfake videos, or synthetic voice, that are used to design malicious content. The phishing vector identifies the method of delivering a phishing message where a social platform features exploited dimensions help to detect specific platform functionalities, such as Instagram stories, Facebook sponsored ads, and Twitter threads. The target profile categorizes who is being targeted based on the demographic data, scrolling time, or technical knowledge. Finally, the delivery and automation patterns detect whether the attack is manual, semi-automated, or fully automated via bots, scripts, or AI. Then, the validation process of these five core dimensions is done by conducting surveys, expert interviews, and analysing real-world case studies. A total of 37 responses were gathered from different social media platforms using a Google Form survey. The interviews were conducted via Zoom with one cyber security specialist, two social media content creator, one cyber security student and

one general user. The surveys are analyzed using pie and bar charts. On the other hand, thematic analysis is done using Braun and Clarke's [9] six-phase methods for the expert interviews. In addition, six real scams from 2025 were analyzed to validate the five dimensions of the proposed taxonomy. The third step is to design a defense architecture using a multi-layered framework. The defense framework shows the process of how data is delivered, and actions are taken.

The document is structured into six chapters. Firstly, the introduction part gives an overview of phishing attacks and the current scenario of phishing on social media. In this chapter, the problem statement, research objective, and research purposes are described. The literature review chapter provides an overview of traditional phishing attacks and AI-powered phishing attacks. This chapter introduces the concept of taxonomy in cyber security and the existing taxonomies and defense framework. The chapter concludes by highlighting the key gaps and limitations in existing frameworks. Chapter three describes the overall research design process. The data collection part consists of the overall process of surveys and expert interviews, and real-world case studies. The purpose of selecting six real-world examples is described in the last part of the data collection. Finally, the third and fourth section provides the details about the development process of the proposed taxonomy and defense framework. Chapter six includes the survey results with proper explanations, a summary of the expert interviews and their experience with social media phishing attacks. Finally, the result from the real-world case analysis is described in the last part. The conclusion part describes the key contributions of the thesis. Finally, a proposal for community resources and the future scope for this research is provided in this part.

1.1 Problem Statement

Phishing attacks have evolved from traditional email-based schemes to sophisticated AI-based attacks. Traditional phishing research and defense methods were mainly focused on attacks that are delivered via email and SMS. The methods can classify static, manually crafted messages only. However, the AI-driven phishing introduces more complex risks by using AI tools such as LLMs, deepfake technology, and automated social engineering. Now, attackers can mimic human behavior and make more realistic content for the targeted individuals or companies. While the threats continue to increase, the existing taxonomies and defense framework remain outdated and insufficient. The existing methods often fail to classify the dynamic, generative, and behavioral nature of AI-based phishing, particularly in the context of social media. Social

media has become an ideal target for attackers due to its huge user base, casual environment, and trust-based interactions. These platforms offer multiple phishing vectors that include direct messaging, fake profiles, and sponsored content. These vectors are actively exploited using AI techniques to mislead users more effectively.

The complexity of phishing attacks is increasing, but there is a lack of structured frameworks to categorize or mitigate them. This presents a critical gap in research focusing on social media platforms. With over 4.95 billion users worldwide who share their daily activity on different platforms. The large number of users is primarily comprised of younger people and middle-aged adults. This behavioral vulnerability makes social media users prime targets for AI-based phishing attacks; however, defense strategies have not evolved accordingly. The current methods overlook the platform-specific features and mainly focus on traditional attacks. They are unable to distinguish human-written and AI-generated content. Most of the existing taxonomies are validated through email datasets or technical protocol logs. They do not examine the real-world user experiences from different platforms like Instagram, Facebook, or TikTok.

Therefore, there is an urgent need to develop a dedicated taxonomy that reflects the complexity of AI-driven phishing on social media. In addition, a defense framework is also needed to focus on platform-specific features, user behavior, and evolving generative attack methods. A new taxonomy must be multi-dimensional, platform-specific, and AI-based to provide effective analysis in both academic and practical cyber security responses.

1.2 Research Purpose

The goal of this research is to understand the complexity of AI-powered phishing attacks on social media platforms, such as Facebook, Instagram, TikTok, YouTube, and LinkedIn. Attackers use different AI techniques to exploit users by mimicking real users, creating convincing messages, or influencing users' behavior. Most of the existing taxonomies and defense frameworks are designed to analyze traditional phishing attacks. These primarily focus on traditional email or SMS-based threats and fail to identify AI-generated attacks. This research tries to fill the gap by developing a comprehensive taxonomy that analyses the multidimensional nature of AI-generated phishing.

In addition, the study aims to collect better insight into users' experience by conducting surveys and interviews. These insights will help to validate the taxonomy and reveal human

vulnerabilities, such as over-trusting behavior and low detection confidence. The research proposes a multi-layered, non-technical defense framework based on the findings. The defense framework integrates user awareness and platform-level protections to mitigate these evolving threats. Finally, the study promotes a user-centered cyber security approach and provides a foundation for future research and platform design improvements that respond effectively to AI-driven social engineering tools.

1.3 Research Objectives

The research objectives are given as follows:

- Design a taxonomy to classify AI-powered phishing attacks on social media platforms based on five core dimensions, such as delivery methods, generative model types, target user profiles, platform features exploited, and automation patterns.
- To validate the proposed taxonomy, real-world case studies of AI-generated phishing incidents are analysed that ensure the taxonomy classification accurately reflects recent attack strategies.
- To conduct user surveys and expert interviews for gathering qualitative and quantitative insights to analyze user awareness and behavioral vulnerabilities of AI-based phishing risks.
- To evaluate survey and interview responses by describing users' experience with phishing attacks, how well they can detect them, and which platform-specific features are most exploited.
- To identify existing gaps and limitations in current phishing taxonomies and defense framework methods of AI-based phishing attacks.
- To propose a multi-layered, non-technical defense framework that aligns with the proposed taxonomy and addresses all the key dimensions of AI phishing.
- To recommend actionable, data-driven strategies to improve online safety that focus on education, policy guidance, and adaptive social media defenses designed for vulnerable user groups.

2 Literature Review

This chapter provides an overview of existing literature related to traditional and AI-based phishing attacks. It also analyzes existing taxonomies in cyber security to categorize different phishing attacks. The review begins by examining the broader landscape of phishing attacks and discussing AI-powered phishing, including its evolution on social media platforms. Then, it examines how phishing attacks have been classified in previous taxonomies by different authors. The existing defense frameworks that were proposed by the other authors to mitigate phishing attacks are analyzed in this chapter. Finally, it identifies the gaps in the existing taxonomies and defense frameworks.

2.1 Phishing Attack

Phishing is a type of social engineering where attackers try to mislead people into sharing their sensitive information by pretending to be a trustworthy individual [10]. This process is done through emails or fake websites. Technical deception and psychological manipulation are also used to steal the data. The main aim is to steal confidential data like passwords or financial details. There are several types of phishing attacks. Each attack uses different techniques to steal information from users. The types of phishing attacks are discussed in Table 1.

Table 1 Different Types of Phishing

Types of Phishing	Description
Deceptive Phishing	This type of phishing tries to pretend as a trusted organization to steal personal or financial information.
Spear Phishing	Targets specific individuals using customized details to make the scam more convincing.
Clone Phishing	Duplicates a real email and replaces links or attachments with malicious ones.
Whaling	This phishing attack targets high-profile individuals like executives using modified phishing messages.
Link Manipulation	Mislead users into clicking on malicious URLs that will lead to fake websites.
Voice Phishing	Uses phone calls to mislead victims into disclosing their personal information.

2.1.1 AI-based Phishing Attacks

AI phishing is a type of cyberattack where attackers use AI to create more realistic and convincing phishing messages. The attackers collect data from social media and public sources.

These data are used to generate malicious content by including personal details to make the scam more realistic. These contents are difficult to identify since they mimic real conversations and reliable accounts. AI can also create a fake account by using real information. AI Phishing is becoming a powerful and evolving cyber threat by using advanced data analysis and automation. The usage of AI in phishing has increased dramatically since late 2022, resulting in a 1,265% rise in phishing emails and losses of almost $2 billion in 2022 [11]. AI-powered phishing uses advanced technology to enhance the effectiveness of phishing attacks. Various types of AI techniques are used in phishing attacks. The basic processes of all the techniques are the same, except for some boundaries. In this section, the key pillars of WormGPT-powered phishing attacks are demonstrated. Data analysis is the first step in the process, where AI gathers and evaluates personal data from public sources [11]. The second step is personalization, where messages are created according to the target's interests and behavior. In the content creation stage, AI generates realistic messages using information from trusted individuals or organizations. Finally, scaling and automation enable attackers to send large numbers of these generated messages. By combining all the elements, AI-generated phishing attacks become more difficult to identify. Figure 3 shows the four pillars of AI-generated phishing attacks by using the wormGPT tools [46].

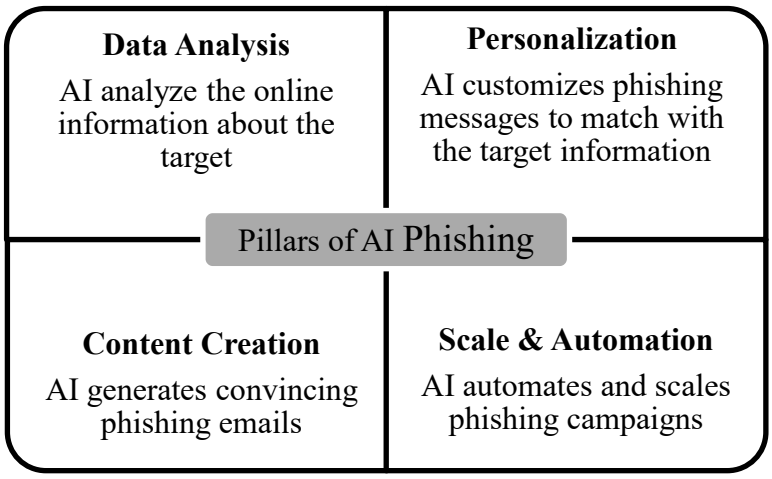

Figure 3 Four pillars of the WormGPT tool

According to the lasted trend report 2025, phishing attacks have significantly increased. Around 68% of data breaches involve human error, where 80-95% of these attacks begin with phishing [12]. ChatGPT and other AI tools are mainly responsible for the rise. Phishing volume increased by 4,151% since ChatGPT's launch in 2022. It highlights that technical defenses frameworks

are not sufficient to solve all the issues. The report shows that behavior-based cybers ecurity training can minimize phishing incidents by up to 86% in just six months. The report's main point is that human activity poses the most risk and can provide the best defense. This line graph in Figure 4 shows the impact of behavioral phishing training over time:

Success Rate (Yellow line): The percentage of users reporting real threats increased significantly. It's rising from 13% to 71% over 24 months.

Failure Rate (orange dashed line): The percentage of users clicking on phishing simulations steadily declined from 20% to just 2%.

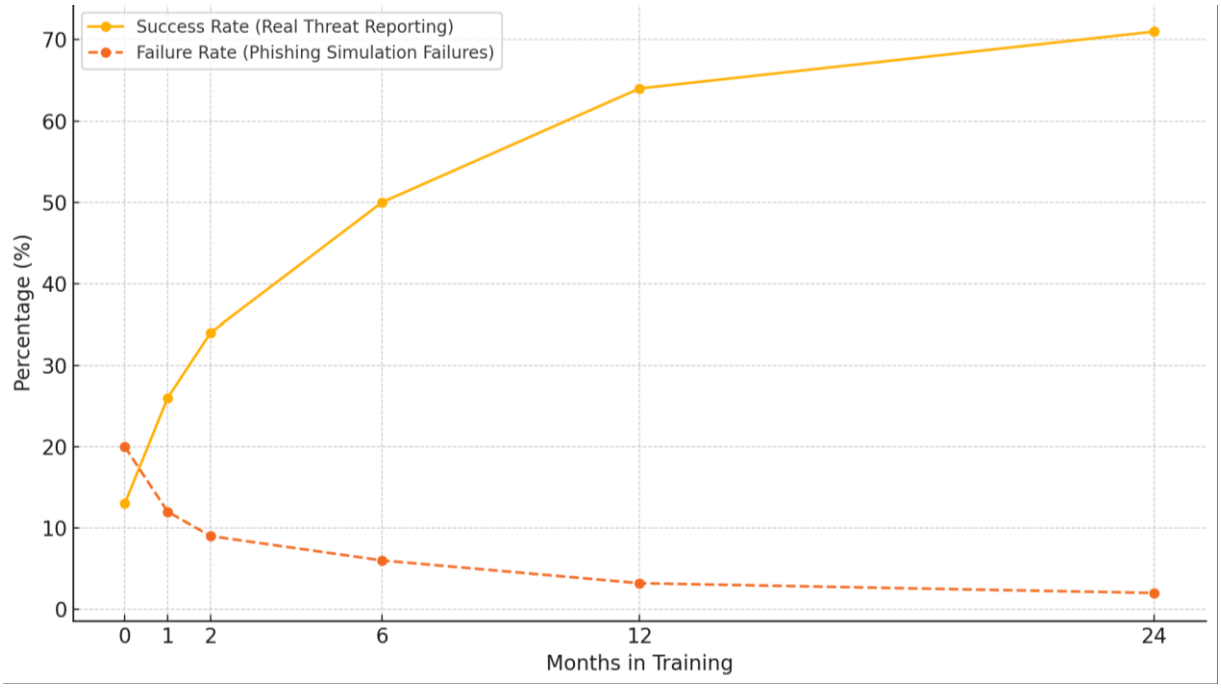

Figure 4 The impact of behavioral phishing training over time

In AI-driven phishing attacks identification, a clear and efficient taxonomy is essential for minimizing cyber threats. It improves users' ability to identify and detect scams by providing behavior-based strategies. This results in more accurate reporting and a significant decrease in phishing incidents. Gupta et al. [13] categorize phishing attacks by type, impact, and method, which helps to develop more focused security policies and compliance measures that directly mitigate human-centric cyber risks. Rabitti et al. [14] describe the importance of understanding the taxonomy of any phishing attacks. It helps organizations to identify potential risks by providing proactive measures such as targeted employee awareness programs and policy enhancements. According to the existing research, a taxonomy of AI phishing attacks helps to

improve detection, training, and response. To defend against more sophisticated attacks, we must move beyond technological tools by developing an effective, knowledgeable workforce.

2.1.2 AI-based Phishing Attacks on Social Media

Social media platforms are common targets for attacks because they contain vast amounts of personal data. People often share their daily activities and private information on these platforms. Cyber criminals use AI tools such as LLMs, deepfake audio and video, and AI-powered bots to create malicious content. Users are often unable to detect AI-generated phishing messages because they are more convincing and realistic. The scammers send this personalized and believable content to targeted users at once. For example, deepfake technology can make it look like a trusted person is speaking in a video call, which makes it easier to trick victims [15]. Similarly, AI-powered social media bots can smoothly interact with users through direct messages and comments. Some recent studies show that the social media features like direct messaging, stories, live streams, feed posts, and sponsored ads are vulnerable to AI-based phishing attacks [16]. These features give attackers many ways to gain users' trust. That malicious content cannot be detected by traditional phishing detection tools. The pie chart in figure 5 shows the most common AI methods that are used to create phishing scams on social media. It helps us to understand which parts of these methods need more defense strategies to protect users. The data presented in the pie chart is gathered from different academic and industry sources. It reflects general trends in AI-powered phishing techniques observed during 2023–2024. Figure 5 shows the different AI-generated phishing Methods on social media.

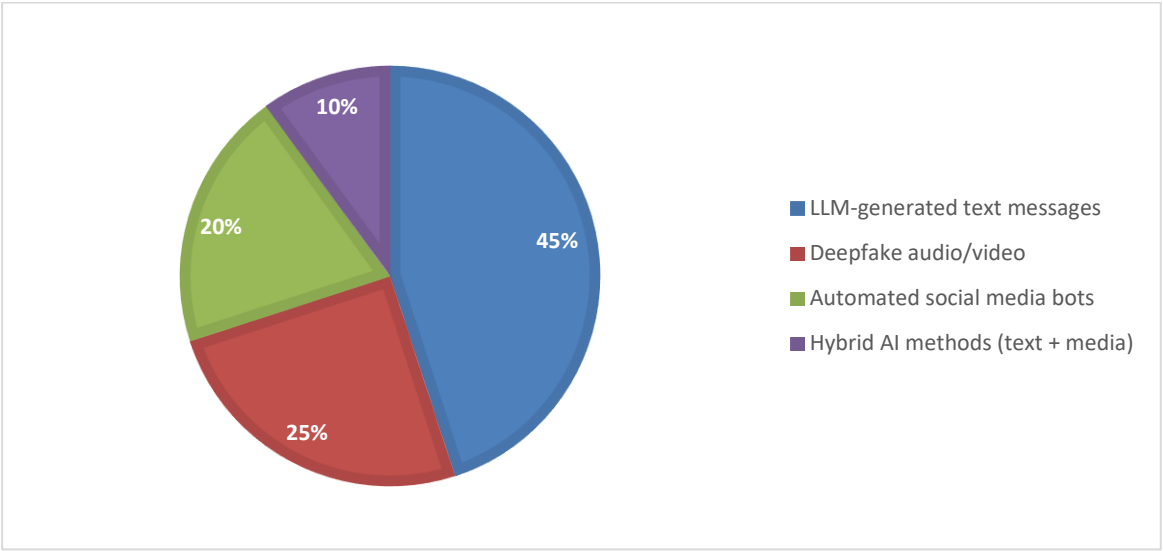

Figure 5 Percentage breakdown of AI-generated phishing methods

2.2 Taxonomy in Cyber security

A taxonomy in cyber security provides a structured framework for systematically describing the characteristics of cyber security threats [44]. This taxonomy helps cyber security professionals and researchers to better analyse, detect, and respond the cyber threats. The cyber threats on different platforms are significantly increasing, but the research gap remains. The effective and well-structured taxonomy helps to develop security solutions against any cyber threats. Many researchers and industries have proposed different taxonomy-based systems focusing on different aspects, such as attack methods, attack vectors, and targeted vulnerabilities, to classify cyber threats.

Loukas et al. [17] propose a taxonomy to classify both cyberattacks and corresponding defense mechanisms for Emergency Management Systems (EMS). As the EMS are digitally interconnected with other systems, they become more vulnerable to cyberattacks. The proposed taxonomy is designed to categorize current threats and defenses. In addition, this system can also anticipate emerging challenges in a field where technology is evolving rapidly. The authors classify the taxonomy in three main ways for attacks, and for the defense, they use three main dimensions. It classifies attacks based on the type of network targeted, the specific function within the emergency response process that is affected, and the method or vector used to carry out the attack. Similarly, it organizes defenses by their nature, whether preventive or reactive, their level of distribution across the system, and the role of organizational factors such as human behavior, policies, and procedures. The proposed taxonomy provides a comprehensive way to analyze any cyber threats and cyber security strategies for EMS.

Rizwan et al. [18] present a comparative study of existing cyber security taxonomies by highlighting the impact of emerging AI-powered threats and defenses. The paper mentioned that the traditional taxonomies are unable to provide a proper way to analyze both AI-based attacks and defense frameworks. As AI continues to transform the cyber security landscape, there is an urgent need to develop new taxonomies that can capture its evolving impact. So, the authors design an enhanced and comparative taxonomy that directly combines AI-powered threats and defenses into the cyber security classification landscape. This system is more advanced than traditional taxonomies because it combines AI as both a threat vector and a defensive tool. Rather than providing a technical answer, the authors give a conceptual study for researchers, cyber security experts, and professionals. It suggests an organized method for

understanding how AI is changing the environment of cyber threats. The method also supports the evolution of defense strategies. Through a systematic comparison, the taxonomy helps to evaluate and refine cyber security approaches in a rapidly evolving technological environment.

Rabitti et al. [14] suggest organizing the existing cyber risk taxonomies for understanding the concept properly. It proposes a meta-taxonomy that categorizes current methods into four primary groups. The groups are described as follows:

- Attack-Based Taxonomies: This taxonomy focuses on the way of carrying out cyberattacks, such as through malware and phishing.
- Harm-Based Taxonomies: This taxonomy categorizes risks depending on the type and extent of damage, like financial loss or reputational harm.
- Operational Risk-Based Taxonomies: This taxonomy integrates cyber risks into broader operational risk frameworks.
- Holistic Taxonomies: This taxonomy aims to provide a complete view of cyber risk in a multi-dimensional structure.

According to the authors, an effective taxonomy should be comprehensive, simple to understand, well-divided, and flexible. They examine the existing taxonomies using these criteria and demonstrate how the selection of a taxonomy affects risk analysis, data quality, insurance planning, and decision-making. Finally, the study concluded by emphasizing the necessity of straightforward and uniform taxonomies. It also suggests that data and machine learning may be used in future research to create more effective solutions.

2.3 Existing Taxonomies of Phishing Attacks

Numerous researchers have contributed to developing taxonomies for phishing attacks. They aimed to understand the evolving nature of phishing and help in designing effective detection and prevention mechanisms. The existing taxonomies have been proposed to categorize attacks based on delivery method, content sophistication, and psychological manipulation method. According to Nuiaa et al. [19], the study proposed a multi-dimensional taxonomy that classifies phishing attacks into five major dimensions. Each dimension has a subcategory. The new taxonomy provides a clear view of phishing attacks by covering every step from creating fake content to stealing information. It also includes modern threats like mobile malware and fake

social media profiles, which makes it more comprehensive than older models. The taxonomy shows how attackers use different methods and platforms. Using the proposed taxonomy, cyber security experts can identify and defend phishing attacks because it provides a more flexible and clear perspective.

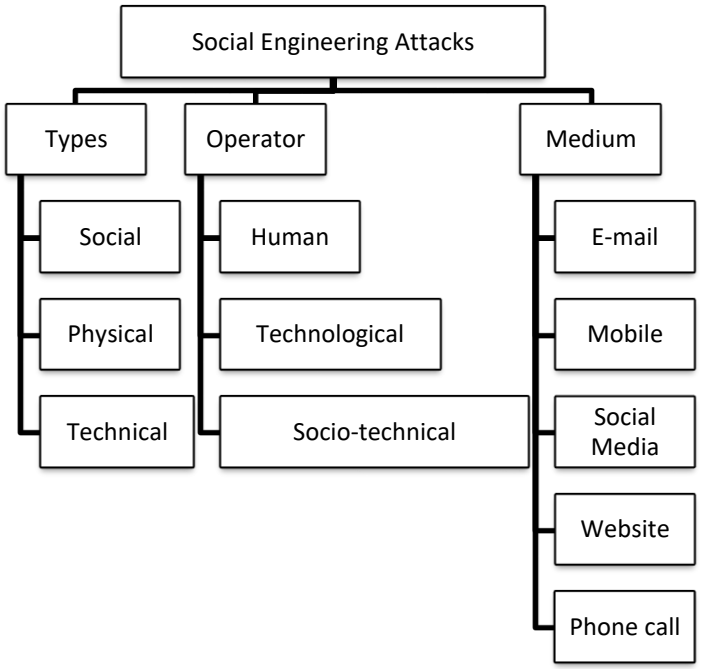

Figure 6 Taxonomy design of social engineering attacks

In this digital and internet-driven era, Aldawood et al. [20] focused on the growing importance of social engineering attacks and the need for both human and technical countermeasures. They proposed a taxonomy that classifies social engineering attacks based on the type, operator, and medium of the attack. The proposed frameworks describe the different techniques used by the attackers, such as social, technical, socio-technical, and physical approaches. It also mentioned the use of in-person and technology-based interactions. Figure 6 shows the taxonomy design of social engineering attacks.

The six phases of email-based phishing attacks taxonomy are described in the figure 7. Each phase includes different methods that attackers use. This model is more detailed and flexible than earlier ones, providing a better understanding of phishing attacks. The taxonomy was validated through integration into a real-world incident management system. It gives a clear improvement in the quality and consistency of phishing attack documentation.

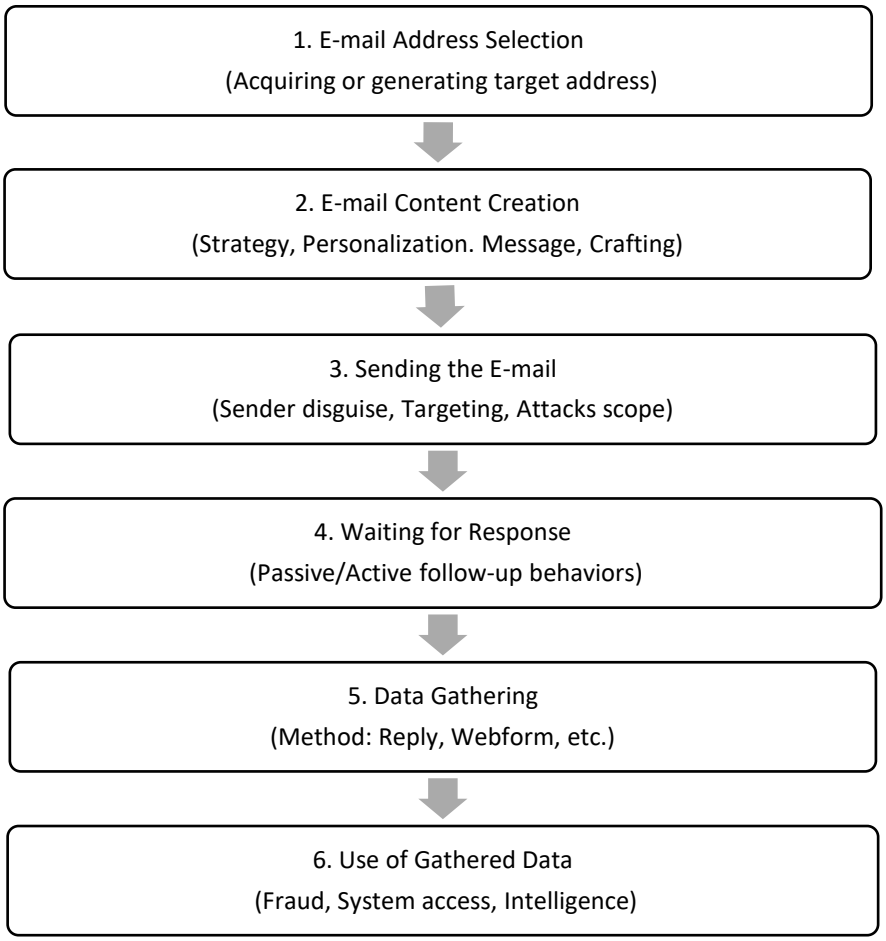

Figure 7 Six phases of email-based phishing attacks

Ivaturi et al. [22] proposed a detailed taxonomy for the classification of social engineering attacks. The authors categorize social engineering methods based on the medium and nature of interaction between the attacker and the victim. Social engineering exploits human vulnerabilities more than technological ones, according to the study. It is necessary to understand the context of human interaction for developing effective countermeasures. The proposed taxonomy includes four primary attacks, such as direct person-to-person interaction, communication through digital media (such as emails or messages), voice manipulation (like phone scams), and video deception (such as video calls or pre-recorded content). Attackers use different methods, like messages, phone calls, or video chats, to mislead people by taking advantage of their trust. They create a sense of urgency or pretend to be someone important. These incidents especially happen on different social media platforms, where people often interact with each other casually. Social media users are more likely to believe what others say without any doubt.

Similarly, Tulkarm et al. [23] present a detailed taxonomy that differentiates human-based from computer-based social engineering attacks. This taxonomy categorizes attacks across three dimensions:

1. Operator (human-based vs. computer-based)
2. Method of deception (social, technical, and physical)
3. Nature of communication (direct vs. indirect)

The study provides some real-world attack examples where social engineering caused major financial losses. The attackers make the scam more effective because of human trust rather than technical methods. While various tools like intrusion detection, biometrics, AI, and machine learning can help to mitigate the risks, prevention depends on user knowledge and awareness.

2.4 Existing Defense Framework and Countermeasures

Many researchers and industries have proposed different taxonomy-based systems focusing on different aspects, such as attack methods, attack vectors, and targeted vulnerabilities, to classify cyber threats. Many researchers have suggested different defense frameworks to protect against AI-powered phishing attacks. Since most of the phishing attacks occur because of human error, it is more important to focus on user awareness and knowledge than on technical solutions. AI can help to detect and mitigate phishing attacks, but it will be more effective when users understand the risks. By creating or improving defense frameworks, user awareness can be improved. This will help to minimize the attacks. This research discusses several existing AI-based defense frameworks that aim to protect users in real time from phishing threats.

Odeh et al. [24] proposed an AI-based defense framework to detect phishing attacks in real-time. Traditional methods of phishing detection depend on manual analysis or static signature-based systems, which struggle to detect the rapidly evolving methods. To solve this issue, the study proposed an AI-driven method that combines machine learning, deep learning, and behavioral analysis to identify phishing attacks. The information from users is collected from different sources, such as web browsing, email interactions, and link clicks. This continuous collection process is done in the detection layer. Then, the collected data is analyzed by AI methods that evaluate patterns, URL structures, and website content for indicators of phishing. These models have large datasets so that they can detect sophisticated malicious content. The

AI continuously monitors user behavior to detect unusual activity, like visiting suspicious websites or entering login details on fake pages.

Study by Alabdan et al. [30] present the application of AI for improving the detection of phishing attacks on the web. This study mainly focused on Machine Learning (ML) and Deep Learning (DL). Traditional detection systems are slow and often ineffective against evolving phishing attacks. AI enables real-time detection by analyzing user behavior, website features, and content using ML, DL, and Natural Language Processing (NLP). It identifies phishing patterns by analyzing malicious content. The proposed model used three different detection techniques, such as Supervised Learning, Unsupervised Learning, and Reinforcement Learning. AI-powered tools such as web filters, anti-phishing software, and browser extensions help to protect users' data by continuously monitoring web traffic and email activity. The tools help to block and stop scams immediately. While AI improves detection speed and accuracy, some challenges remain, including managing false positives, evolving attacker methods. Overall, the research demonstrates the vital role of AI in transforming phishing detection from a reactive to a proactive process by enhancing the security strategies of individuals and organizations.

2.5 Gaps in existing Taxonomies

Phishing attacks on social media are increasing day by day. Attackers are using AI methods to improve their attack techniques. AI makes the attacks more difficult to detect and recognize. The complicated nature of platform-specific attacks and AI-powered phishing is not effectively captured by most of the current taxonomies. To identify these gaps, this research investigates several existing phishing taxonomies. Table 2 shows key taxonomies proposed by the authors with a comparison of their limitations, scope, and social media involvement. This demonstrates the need for a new taxonomy focused on AI-driven phishing on social media platforms and helps to identify research gaps.

Table 2 Gaps in existing taxonomies and defense frameworks

Taxonomy	Authors (Year)	Scope	Social media	Gaps/Limitations
E-mail-Based Phishing Attack Taxonomy	[21] Rastenis et al. (2020)	Six phases of email-based phishing	No	Excludes social platforms and AI-based threats.
Phishing Environments & Countermeasures	[25] Aleroud & Zhou (2017)	Media (email, web, social networks), devices, techniques	Yes	Fails to differentiate platform-specific vulnerabilities or detailed phases.
Phishing Attack Defense Taxonomy	[13] Gupta, Tiwari & Jain (2017)	Taxonomy of technical and user-centric defenses	Yes	Focuses on High-level defensive methods: lacking detailed analysis of attack types
Social Engineering Attack Taxonomy	[22] Ivaturi & Janczewski (2011)	Person–person vs. media (text/email/SMS, voice, video)	Yes	Addresses media in general, but does not mention AI-based attacks on social media platforms
Social Media Phishing and Countermeasures	[30] Chung, Koay & Leau (2021)	Social-media-specific attack factors and defenses	Yes	Focuses on only one platform (Twitter)
Social Engineering Attacks Survey & Taxonomy	[27] Odeh, Eleyan & Eleyan (2021)	Classification of social engineering by operator, method, and channel	Yes	Focuses on social engineering but does not provide the details classification of phishing.
Taxonomy of Phishing Attacks	[19] Nuiaa & Manickam (2023)	Phishing phases, attacker types, targets, media, and strategies	Yes	Does not provide detailed coverage of social media platforms or the latest AI-driven phishing methods.
AI-driven Phishing Detection Systems	[28] Ogundairo & Broklyn (2024)	AI methodologies for phishing detection (ML, DL, NLP, Feedback)	Yes	Focuses on detection; lacks integration into a taxonomy of attack characteristics and gaps in coverage.
Real-Time AI-powered Phishing Detection	[26] Asiri, Xiao, Alzahrani, & Li (2024)	AI-driven, behavior-based phishing detection	Yes	Does not formalize a taxonomy.

3 Research Methodology

This chapter consists of four sections. The first section describes the process of designing this research. The second part is about data collection. The data collection part is divided into three sections such as survey design, expert interviews, and recent case studies. The survey process, design, and questions are included in the survey design section. The interview section consists of three parts, which are the interview purpose, the interview structure, and the interview questionnaires. With the surveys and interviews, this research also analyzed the six real-world cases to validate the taxonomy. The final two parts of this section are about the development process. All the step of developing taxonomy and defense for AI-powered phishing attacks on social media is discussed. Figure 8 shows the overview of the methodology section.

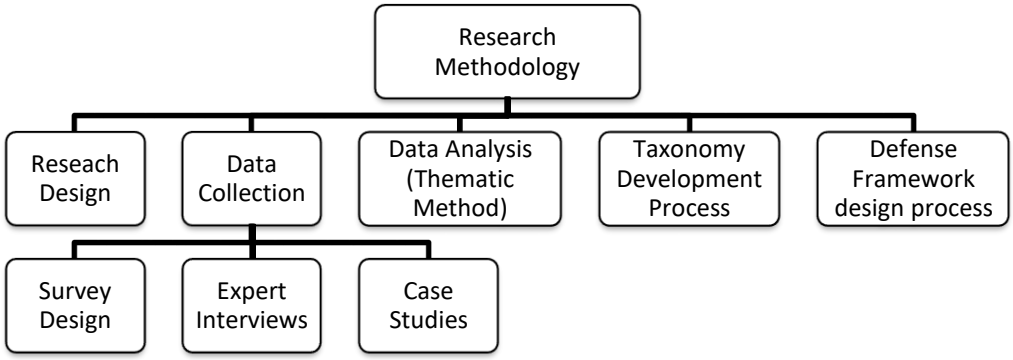

Figure 8 The research methodology overview

3.1 Research Design

A mixed methods approach (qualitative and quantitative) is adopted for this research to develop an AI-powered Phishing attack on social media platforms. The methodology follows a structured, multi-phase method combining literature review, content analysis, taxonomy design, and defense framework design. The proposed taxonomy is validated by analysing survey findings, interview insights, and case studies. This taxonomy and defense framework aims to mitigate the AI-powered phishing attacks on social media platforms.

The qualitative aspect focuses on the design and development of the taxonomy. It involves the following steps.

- A systematic literature review of academic papers, the current situation according to different websites, and cyber security reports.
- The content analysis of attack vectors, platform-specific methods, and real-world phishing case studies.

The proposed taxonomy and framework are validated and evaluated with the help of the mixed-methods approach. It involves the following steps.

- Surveys and interviews are structured to get feedback from cyber security professionals to assess the taxonomy's relevance, clarity, and coverage.
- Surveys are examined using pie and bar charts, and interviews are analyzed using thematic methods.
- A comparison between the new taxonomy and current models.

3.2 Data Collection

To validate the proposed framework, data is collected for this research in three different ways. The first method is surveying with ten common questions that help assess user awareness and experience with AI-powered phishing attacks on social media platforms. The second method involves interviews with four individuals: a cyber security expert, two content creators, a cyber security student, and a general social media user. Finally, some recent phishing incidents are analyzed according to the five core dimensions of the proposed model. The methods are briefly discussed in the following sections.

3.3 Survey Design

The survey was conducted to collect real data from social media users to validate the relevance and completeness of the proposed taxonomy of AI-powered phishing attacks. This taxonomy includes five main dimensions, such as Generative Model Modality, Target Profile, Phishing Vector, Social Platform Feature Exploited, and Delivery & Automation Pattern. The goal of this survey is to understand how users respond to phishing attacks in each of these areas.

3.3.1 Survey Structure

The survey had ten required multiple-choice questions and optional open comments. Each question was linked to at least one part of the proposed framework. The questions were divided into four thematic categories as follows.

1. Demographics and User Profile: The first two questions were related to age group of the participants, followed by their technical background.

2. Social Media Platforms Patterns: One question was related to which platforms are most used. This helps to identify which social media platforms are more vulnerable. To collect the information about the amount of time spent online, participants were asked specific questions regarding their usage.

3. Phishing Awareness and Experience: To collect data regarding each participant experience with phishing attacks, some questions were asked about how they encountered the suspicious content, the type of message received, and the delivery methods used.

4. Perception of AI-generated Content: Some survey questions aimed to collect data about participant's understanding of phishing attacks, and their confidence level of detecting them.

3.3.2 Survey Collection and Tools

The survey was prepared using Google Forms. The Google Forms tools are more convenient for monitoring and exporting results from the responses. The Google Form link was distributed on LinkedIn, Twitter, Facebook, and sent to cyber security-related communities to reach a wide and diverse audience. The survey was conducted over two weeks. To improve data quality, the survey included input checks and required all ten multiple-choice questions to be answered. Optional comments were allowed at the end. All the participants stayed anonymous, so their names, emails, or IP addresses were not collected. The purpose of the research and the participants' rights were explained at the beginning of the form, as shown in the figure 9.

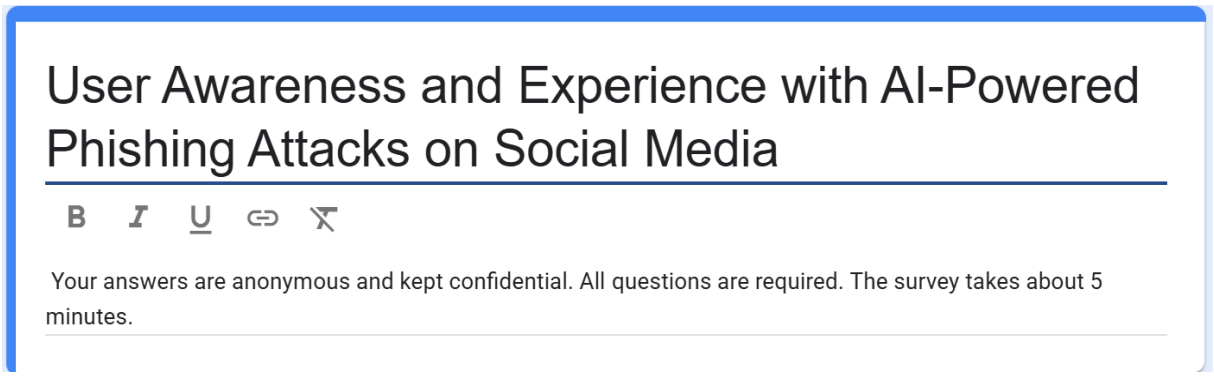

Figure 9 Disclaimer of the participants' rights

After the survey ended, the data was analyzed using Excel and different charts to show the results. Responses were reviewed according to the five dimensions of the proposed phishing taxonomy. The analysis helped to see how well the model matched real user experiences. This approach helped to ensure the survey was ethical, accessible, and useful for testing the framework.

3.3.3 Survey Questionnaire

All the survey questions are prepared to align with the five core dimensions of the proposed taxonomy: Generative Model Modality, Target Profile, Phishing Vector, Social Platform Feature Exploited, and Delivery & Automation Pattern. All the survey questions from the Google form are listed below:

1. The following question collects demographic information to determine the possibility of getting scammed according to the age group. By dividing users according to age, the taxonomy illustrates how risks may vary among age groups. Figure 10 shows the first question of the survey about the different age groups.

Figure 10 Demographic data collection

2. The second question is about the technical expertise of the user. This helps to assess the users' digital knowledge levels. People with different technical backgrounds may respond differently to AI-generated phishing. It can help to determine how detailed the target profile category should be in the taxonomy. Figure 11 shows the second question of the survey related to technical expertise.

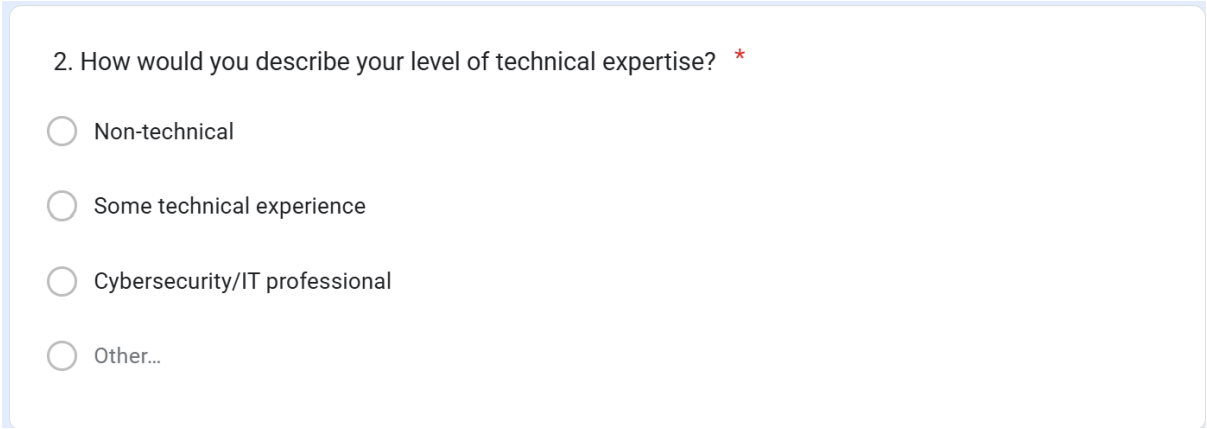

Figure 11 Users technical expertise

3. The third question collects data regarding the amount of time users spend on social media per day. Users who are online more often may be targeted more frequently. This helps evaluate the relevance of platform-based dimensions in real-world scenarios. Figure 12 shows the third question of the survey that provides information about users' online activity level.

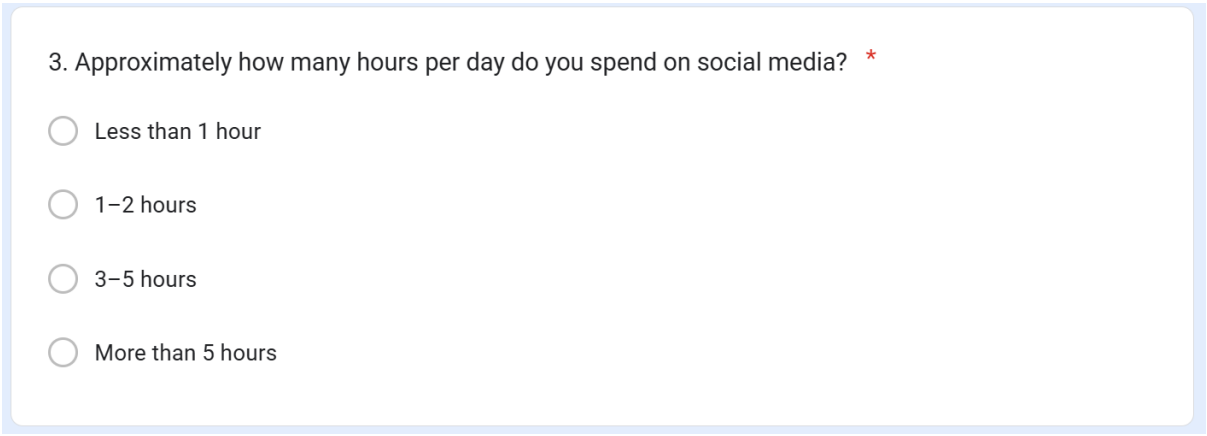

Figure 12 Duration of users' online activity

4. The following question identifies which social media platforms are frequently used. It ensures relevance by highlighting where users are most active. It helps to validate which social media features should be emphasized. Figure 13 shows the fourth question of the survey.

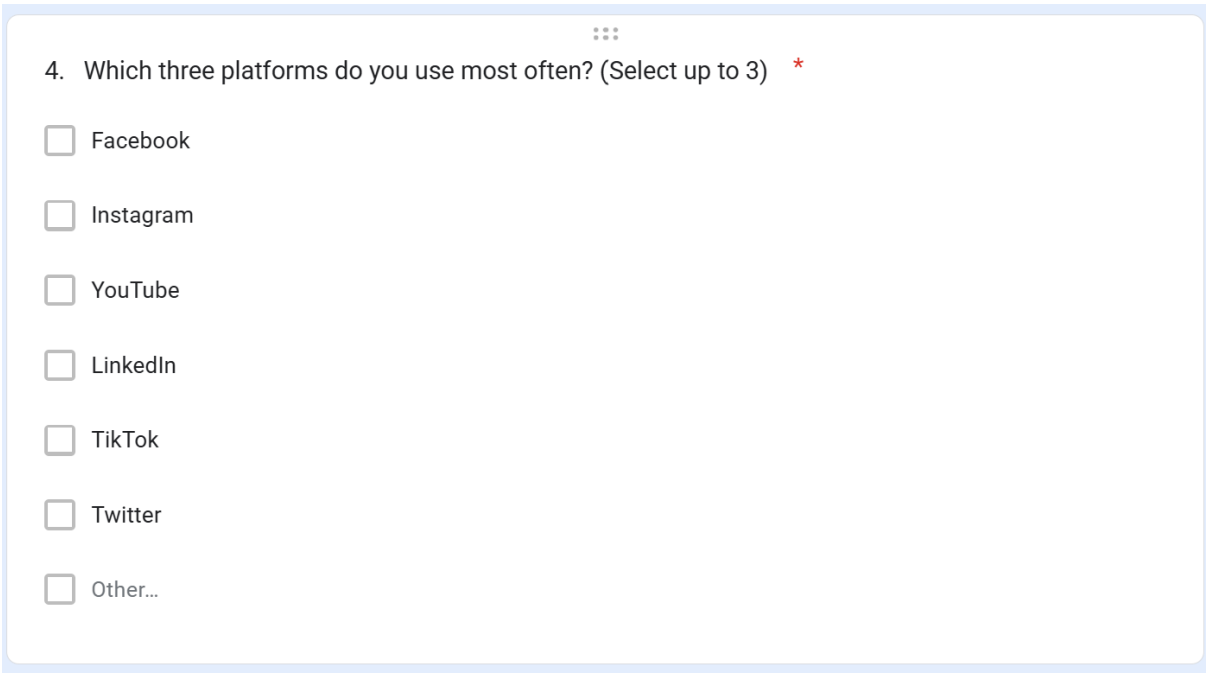

Figure 13 Platform-based data collection

5. The fifth question of the survey helps to check participants' awareness of AI tools. These tools are commonly used in phishing attacks. Figure 14 displays the fifth survey question.

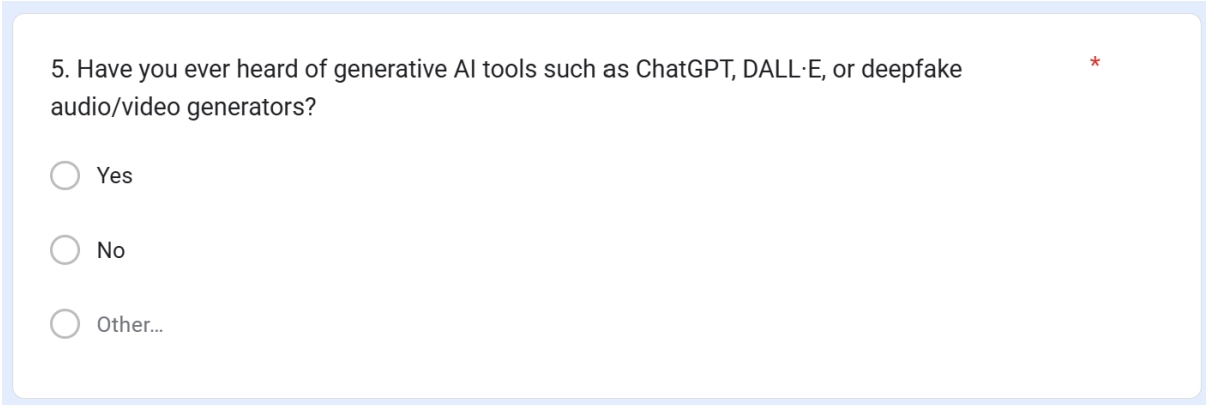

Figure 14 AI tools literacy

6. The next question illustrates whether people believe generative AI tools can create convincing phishing messages on social media. Figure 15 shows the sixth question of the survey.

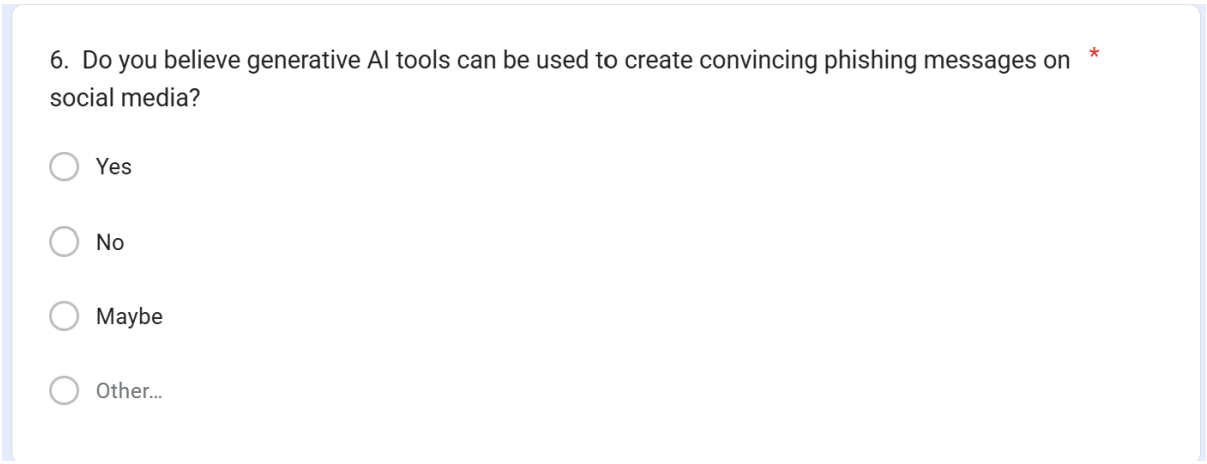

Figure 15 Users' beliefs regarding AI-generated content

7. The following question is shown in Figure 16, which gathers information about users' real experience with AI-generated phishing attacks on social media.

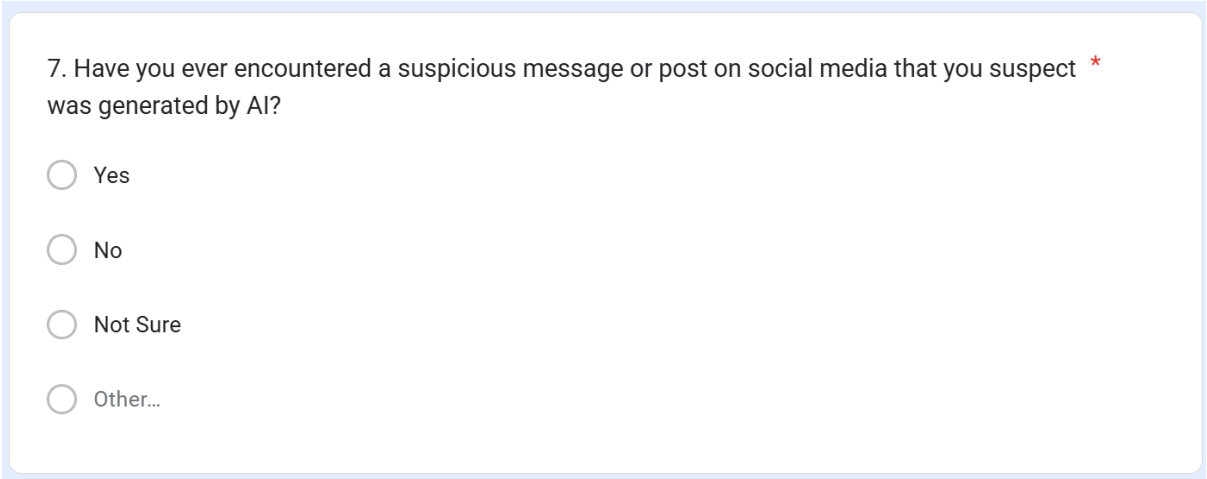

Figure 16 Users' experience with phishing attacks

8. This question identifies how phishing attacks are delivered. There are different social media features, such as DMs, comments, ads, or videos, that are used to deliver the malicious content. It validates the taxonomy's categorization of vectors and platform features. Figure 17 shows a question about different social media features.

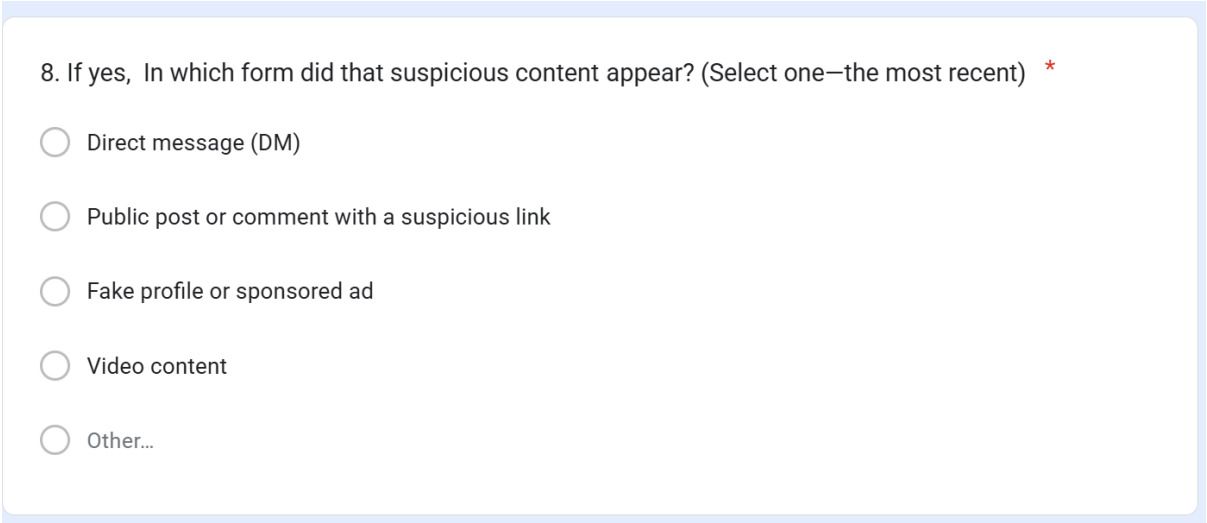

Figure 17 Data collection regarding different social media features

9. This evaluates the users' ability to detect phishing attacks on social media. This also shows how well users believe they can differentiate the human-written and AI-generated content. The ninth question is shown in Figure 18, which describes users' confidence level in detecting AI content.

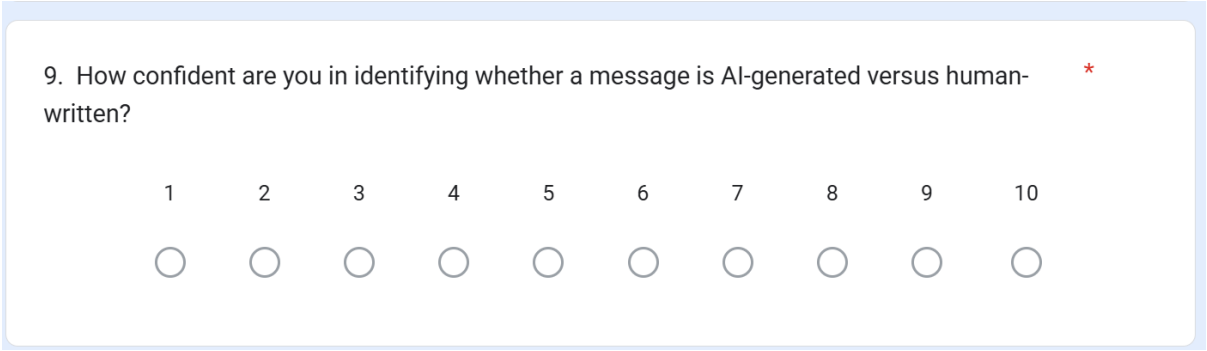

Figure 18 User´s confidence level detection

10. The question tests whether users are likely to be tricked by the attacker, as shown in Figure 19.

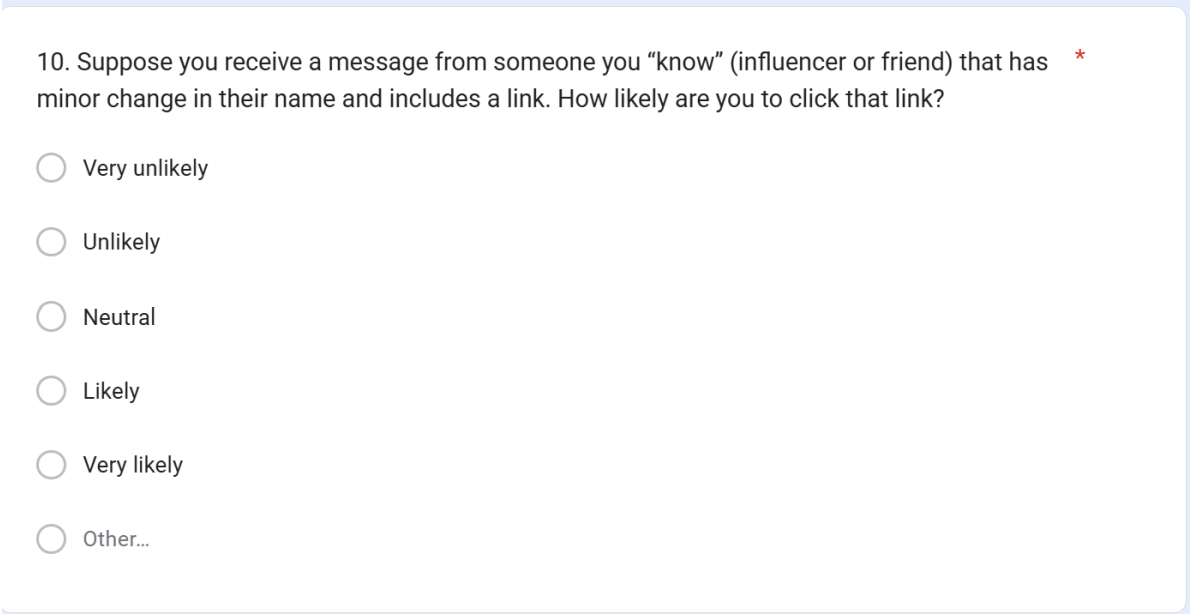

Figure 19 The possibility of clicking on malicious content

3.4 Interview Design

While conducting surveys gives the quantitative validation of the proposed taxonomy, the interviews give the qualitative validation. The interviews were conducted with five different participants with different backgrounds. Interviewee selection criteria, interview place, data, analysis method, and questions are discussed in this section.

3.4.1 Interview Purpose

The main purpose of conducting interviews in this research is to validate the relevance, clarity, and effectiveness of the proposed taxonomy for AI-driven phishing attacks on social media platforms. The interviews help to evaluate deeper thoughts and feelings that may be overlooked by simple questions. The purpose of the interviews is to determine the significance and relevance of the five dimensions by comparing them with real-life cases. The interviews specifically aim to determine whether these dimensions can provide the following insights.

- Real experiences of phishing attacks by gathering information about how they faced the phishing attack.
- Compare authentic responses with online threats and determine how they recognize, understand, and react to malicious content.
- A practical method to explain and classify new phishing techniques that use AI technologies.

3.4.2 Interview Format

A set of semi-structured interviews was conducted with five participants from different fields. Each chose to represent their perspectives and experiences about AI-powered phishing threats on social media platforms. The key features of the interviews are described below:

- **Medium:** Online (Zoom)
- **Interview Style:** Semi-structured (ten core questions + follow-up questions for flexibility).
- **Duration:** 25–30 minutes per participant
- **Selection Criteria:** Participants were selected to represent a mix of technical expertise, platform insight, and general user behavior that ensures effective feedback for the taxonomy.

Table 3 shows the participants' details given with their full consent.

Table 3 Interview Participants Details

Name	Age	Occupation	Company
K. M. Mehedi Hassan	27Y	Cyber security Specialist	British Council, Bangladesh
Faysal Hossain Durjoy	29Y	Engineer and Content Creator	JinKwang Construction & Engineering Co. Ltd. and Instagram page: faysaldurjoy.life
Md. Junaid Ahmmed	27Y	General Social Media User	University of Turku
Rejwan Arefin Miraj	28Y	Cyber security Student	Bangladesh University of Professionals - BUP
Tahmina Rahsan	30Y	Content Creator	Facebook page: Hello Butterfly

3.4.3 Interview Questionnaires

All the interview questions have designed to learn about participants' real-life experiences and their views on social media phishing attacks. It was also focused on how these attacks relate to the five dimensions of the proposed taxonomy. Each question was designed to gather qualitative insights that help evaluate the taxonomy's completeness, practicality, and relevance. The validation of interview questions for the five dimensions is described in the taxonomy validation section. All participants were asked the following ten questions during their interviews.

Q1: Can you share any recent phishing attacks on social media that you've experienced or heard about?

Q2: Did the content seem human-written or AI-generated? What made you think so?

Q3: Through what channel or feature did the phishing message reach you (e.g., DM, comment, ad, video)?

Q4: Which part of the message seemed most suspicious to you?

Q5: Who do you think the attackers were targeting, and why?

Q6: Did the phishing attack appear personalized to you in any way? How?

Q7: Was there any indication the message was automated or part of a bot-like interaction?

Q8: Which social media features (e.g., stories, polls, live streams) do you think are easiest to exploit?

Q9: Do you believe AI-generated phishing is difficult to detect than traditional phishing? Why or why not?

Q10: If you were to classify the phishing attacks, what terms or criteria would you use (such as delivery method, target audience, or intended outcome)?

3.4.4 Interview Analysis

For analysing the data from the interview, this research adopts thematic analysis methods. One of the most popular techniques for examining qualitative data is thematic analysis, which provides a structured framework for identifying and analyzing key patterns in datasets. The goal of conducting interviews was to gather different views of AI-generated phishing on social media, validate the proposed taxonomy, and understand behavioral patterns. This research follows Braun and Clarke's (2006) six-phase method for thematic analysis [9]. The six phases are Familiarization with Data, Generating Initial Codes, Searching for Themes, Reviewing Themes, Defining and Naming Themes, and Producing the Report.

Phase 1: Familiarization with Data

Each interview transcript was reviewed multiple times to get a proper understanding of the interviewers' insights. Several common patterns were identified, including impersonation and platform-specific features such as direct messages, comments, and sponsored ads.

Phase 2: Generating Initial Codes

Initial codes were generated to capture repeated ideas from the raw data. Some codes are polished AI-generated grammar, suspicious messages or comments, automated responses, and bot activity. These codes were applied systematically through the entire dataset to ensure consistency and coverage.

Phase 3: Searching for Themes

The codes were organized to search for the themes. The themes help to monitor the deeper patterns of the interview data. For example, codes related to persuasive language and perfect grammar were grouped under "AI-Enhanced Persuasion." In addition, phishing through direct messages, comments, or ads was categorized as "Delivery via Trust Channels." Similarly, rapid replies and template-driven interactions were added to the theme "Bot-Like Automation." A total of eight themes were identified that aligned closely with both participant experiences and the core dimensions of the proposed taxonomy.

Phase 4: Reviewing Themes

All themes were reviewed against the full dataset. This step ensures that the data accurately reflects participant perspectives. Some themes were refined and merged.

Phase 5: Defining and Naming Themes

Themes were clearly defined and named based on their content. These are also relevant to the proposed taxonomy. A total of nine themes were identified based on the interviews. Each theme was analyzed according to the proposed taxonomy dimensions to ensure alignment with the theoretical framework.

Phase 6: Producing the Report

The final analysis was organized, shown in Table 4. It shows each theme with a short description, related codes, and the matching parts of the taxonomy. The results showed a connection between users' experiences and the ideas in the taxonomy.

Table 4 Thematic analysis of interview data

Theme	Description	Example codes	Validated Taxonomy dimension
AI-Enhanced Persuasion	AI-generated content uses flawless language and realistic tone to mimic legitimate sources.	'Perfect grammar', 'too polished', 'no broken English'	Generative Modality, Automation Pattern
Delivery via Trust Channel	Messages often arrive via DMs, comments, or ads, appearing to come from trusted accounts.	'Instagram DM', 'sponsored ad', 'comment under post'	Phishing Vector, Platform Features
Impersonation and Identity Abuse	Phishing campaigns frequently involve impersonating trusted individuals, brands, or service teams.	'Fake Meta support', 'crypto trader impersonation'	Platform Feature Exploited, Target Profile
Exploiting Urgency and Reward Triggers	Scams exploit emotions like urgency to pressure people into quick or impulsive decisions.	'Limited time offer', 'act now', 'urgent appeal'	Delivery Pattern
Bot-Like or Automated Behavior	Indicators of automation include identical messages, quick replies, and a lack of personalized engagement.	'Fast replies', 'template responses', bot activity'	Automation Pattern
Personalization (Minimal to Moderate)	Some attacks reference user bios or public content for moderate personalization.	'Mentioned deadline', 'related to recent post'	Target Profile

Theme	Description	Example codes	Validated Taxonomy dimension
Targeted Victim Groups	Specific groups like students, gamers, and crypto enthusiasts are more frequently targeted.	‘University student’, ‘crypto newcomer’, ‘Steam users’	Target Profile
AI Makes Phishing Harder to Detect	Participants agreed that AI-generated phishing is harder to detect due to fluency and polish.	‘Looks legitimate’, ‘no grammar mistakes’, ‘difficult to tell’	Validate all the dimensions
Trust Exploitation via Account Compromise	Phishing attempts that exploit compromised friend accounts to gain user trust, especially on platforms like Steam.	‘Message from friend’s account’, ‘Steam wallet giveaway’, ‘No personal details used’, ‘No reply after response’	Phishing Vector, Platform Feature Exploited, Target Profile

3.5 Case Study Selection

This research describes six real AI-driven phishing cases from 2025 along with surveys and interviews. The suggested taxonomy is practically validated by examining these cases. It also demonstrates how social media and generative AI make phishing campaigns more effective. Each case was evaluated in terms of five core taxonomy dimensions, such as Generative Model Modality, Target Profile, Phishing Vector, Social Platform Feature Exploited, and Delivery & Automation Pattern. This method validates the taxonomy's performance in practical situations by demonstrating its ability to manage several new phishing attacks.

3.6 Taxonomy Development Process

Different steps are followed to design the taxonomy for AI-generated phishing attacks. First, around twenty-five research papers are reviewed to understand the existing phishing types and taxonomies. The research paper helps to identify the gaps and scope for the research.

Next, some real-world phishing cases are collected from social media and security reports that ensure the proposed taxonomy covered new attack methods. Lastly, security expert interviews and surveys are conducted to evaluate the taxonomy. Then, the modification of the taxonomy is done based on their feedback to improve accuracy and usability. The taxonomy development process is shown in Figure 20.

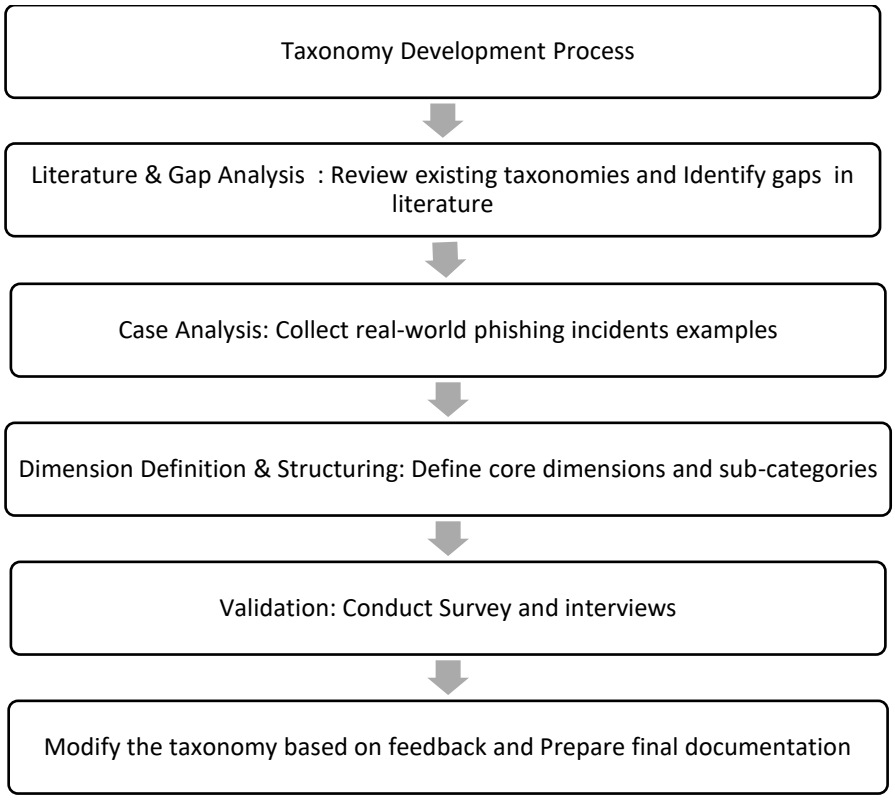

Figure 20 The taxonomy development process of AI-based phishing attacks on social media

3.7 Defense Framework Development Process

This section describes a step-by-step methodology and the logic behind the design of the proposed defense framework against AI-driven phishing attacks on social media. The proposed framework combines real-world threat analysis, a clear breakdown of the system, and AI-powered methods to detect attacks. The first step was to study the different AI tools like LLM, image generators, and automated chat systems. This step was important to understand how phishing attacks are evolving with the help of AI tools. It is easy to create more realistic phishing content using AI. These contents are delivered across different platforms like Facebook, Twitter, Instagram, and LinkedIn. To counter these threats, the defense system must be flexible. The next step was identifying the problem and the gap in the current defense framework. This step is done by analysing the existing research on AI detection frameworks.

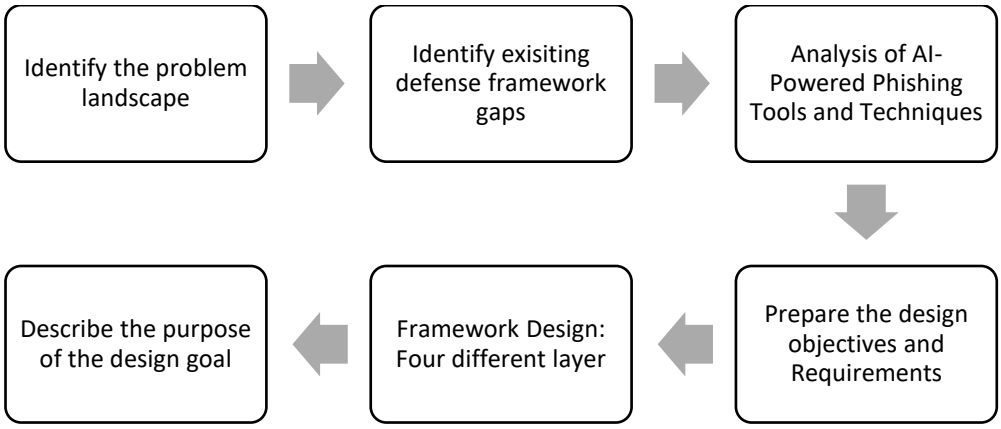

Figure 21 The development process of the defense framework

Then, the research objectives and requirements section are designed. Data availability, threat complexity, platform regulations, and privacy limitations have been evaluated to determine requirements. The defense system was organized into four interconnected layers, each layer built systematically using a logical progression from problem identification to solution implementation. The proposed defense framework is more than a technical solution. It represented a logically structured response to a new type of dynamic cyber threats. Each layer was designed to manage a distinct challenge within the phishing lifecycle. Together, they create a robust and adaptive defense mechanism against AI-powered attacks on social media platforms. Figure 21 shows the defense framework design process.

4 Taxonomy design for AI-based Phishing Attacks on social media

Chapter 4 consists of two sub-sections. In the first sub-section, a taxonomy design for AI-powered phishing campaigns on social media platforms is provided. The structured framework helps to understand and classify the AI-generated techniques. This section introduces the overall taxonomy structure that is designed to detect the evolving nature of phishing attacks driven by AI. The taxonomy has five core dimensions, such as generative model modality, phishing vector, social platform feature exploited, target profile, and delivery and automation pattern. Each dimension is explained in detail using a table to illustrate its categories and characteristics in the following sections. The next sub-section of this chapter is about the validation process of the taxonomy. The validation process is conducted in three different ways such as recent case studies, survey data, and expert interviews. Six recent cases are analyze to validate the taxonomy. Each survey and interview questions are designed to assess which dimension of the taxonomy are validated, along with the specific purpose of the validation.

4.1 Proposed Taxonomy Design

A taxonomy is a hierarchical classification method that organizes concepts into clear categories and subcategories [28]. In cyber security, taxonomies help to categorize threats, incidents, and vulnerabilities. On social media platforms, the taxonomy gives a structured understanding of AI-powered phishing attacks by systematically categorizing the components and techniques used in different campaigns. This classification helps to expose attack vectors such as deepfakes, LLMs, and voice synthesis that have emerged alongside generative AI [31]. By clarifying the distinctive features of these threats, the taxonomy supports the design of automated detection systems, risk assessments, and user-training initiatives. It provides a comparative analysis of traditional and AI-driven phishing attacks. Finally, this framework equips social media platforms and policymakers with a contextualized, platform-aware perspective to inform targeted mitigation strategies. Figure 22 shows the proposed taxonomy for AI-powered phishing Campaigns on social media.

The AI-powered phishing taxonomy framework illustrates how phishing attacks are evolving with the use of AI on social media. AI tools, such as LLMs and deepfakes, are making phishing messages more realistic and difficult to detect. These days, bots that pose as actual users frequently carry out these attacks via social media tools like direct messaging, comments, and advertisements. The information is carried out by bots that act like real users.

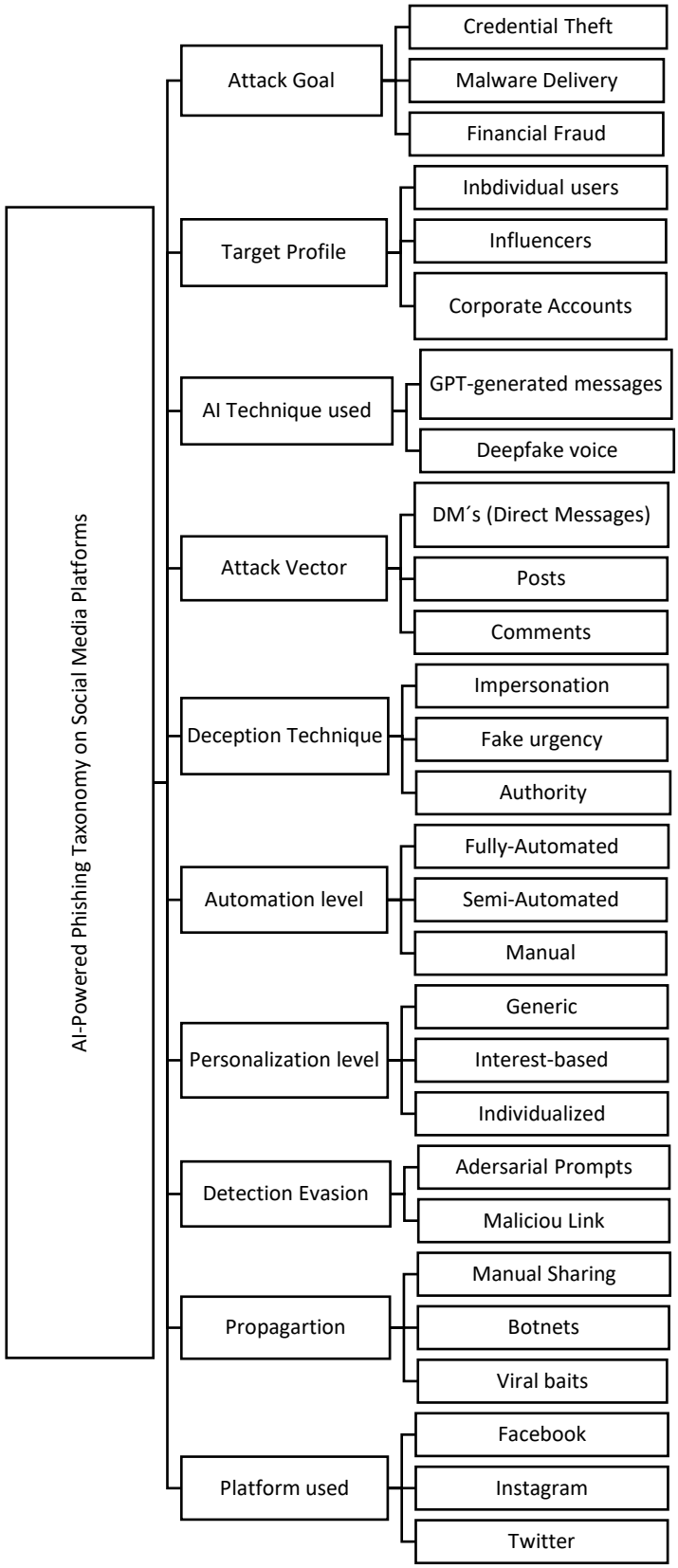

Figure 22 The Proposed taxonomy of AI-powered phishing attacks on social media

Unlike old scams that were generic, AI-driven phishing is highly deceptive. These attacks imitate real people and trigger emotions like urgency or trust. A major shift is the personalization of messages using personal data stolen from users' profiles, making them seem more legitimate. On the other hand, with the help of automation, AI can now handle all the parts of the phishing process, helping attackers scale their efforts with less human involvement. As the phishing attack on social media is increasing at a high rate, this taxonomy will help experts to understand, detect, and defend against these new types of threats. For an in-depth analysis of an AI-driven phishing attack on social media, the mentioned ten dimensions are necessary. For a practical analysis of any attacks, five core dimensions are needed among all the dimensions. The five core dimensions are Generative Model Modality, Target Profile, Phishing Vector, Social Platform Feature Exploited, and Delivery & Automation Pattern. Using these five dimensions, researchers can analyze any phishing attacks. In the next section, five core dimensions are discussed with the necessary table.

4.1.1 Dimension 1: Generative Model Modality

Generative models are transforming AI by enabling common applications, such as image generators (DALL-E and Stable Diffusion) and chatbots (ChatGPT and LaMDA) [29]. By identifying patterns in data, these models generate new content. There are two primary categories such as implicit models and explicit models. Implicit models become creative and flexible because of their gradual improvement through experimentation. More control and predictability are achieved by explicit models, which respond to accurate mathematical principles derived from the data. AI may produce practical and realistic results with the help of both categories. Table 5 illustrates the three main generative modalities: text, image, and multimodal. The table also includes features and an explanation of each generative modality.

Table 5 Features and Explanation of Different Generative Modalities

Category	Features	Explanation
Text	Purely text–based outputs from LLMs like ChatGPT.	Direct Messages (DMs) and posts are common attack vectors, as LLMs are highly effective at generating modified and tailored content.
Image	AI-synthesized still images to mimic official branding.	Branded "alert" images or QR codes are increasingly being used to mislead users by replicating actual prompts.
Multimodal	AI-generated text and pictures (or video) combined for more complex deception.	Combining linguistic and visual deception to increase engagement and trust, posing a growing threat to targets with high value.

4.1.2 Dimension 2: Phishing Vector

Phishing vectors are the channels for attackers to execute phishing attacks, depending on the target they exploit [30]. They determine how phishing content is delivered effectively to target victims. These vectors are essential because they define how phishing reaches and targets victims effectively. Social media is highlighted in the research as a major phishing vector because users prefer to trust their online contacts and provide personal information easily. Attackers utilize false identities, fake profiles, and misleading messages with malicious links on social media sites like Facebook, LinkedIn, Pinterest, Tumblr, and Twitter to conduct targeted phishing attacks. Social media is a very powerful medium for phishing attacks because of its large user base. The relationships between the users are based on loyalty. Table 6 presents the three main vectors used in generative AI-driven phishing campaigns. These vectors mislead people into sharing their personal information. The Category column lists each vector, such as Link-based, Stealing Credential, and Malware Delivery. The feature column explains how each vector works. The importance of differentiating between these approaches is emphasized in the Explanation column. These vectors represent the fundamental methods that attackers use, such as payload delivery, direct data collection, and malicious link clicks.

Table 6 Different Phishing Vectors

Category	Features	Explanation
Link-based	Users clicking on a malicious URL that leads to credential-stealing pages.	A Facebook post: “Verify your account here” hyperlink leading to a spoofed login form.
Stealing Credentials	Uses embedded forms or chat prompts on the platform itself to steal login details.	These attacks are getting popular on sites that allow embedded apps or online forms.
Malware Delivery	Distributes malware via attachments or links (e.g., trojans, ransomware)	Less common but highly effective for ransomware or remote access trojans via seemingly harmless files

4.1.3 Dimension 3: Social Platform Feature Exploited

Social media platforms like Facebook, Instagram, and TikTok collect personal information from users without clear permission from the users [5]. The attackers use this data to make money by using it in ads and business deals. The social media users don’t realize that their data has been compromised. Users don’t realize it because of the hard privacy rules and regulations.

These platforms also use algorithms to keep people online for a long time by showing exciting content. Sometimes the content can spread false information and create online bubbles. To address this, more robust privacy laws and transparent information from platforms are required. Increasing knowledge about online privacy can prevent attacks.

Table 7 Different media for sending malicious content

Category	Features	Explanation
Direct Messages (DMs)	One-to-one or small group private messages.	In high-trust contexts, personalized messages tend to achieve higher click-through rates
Public Posts (Feeds/Timeline)	Broadcast messages to all followers or public streams.	Attackers can use popular hashtags or keywords to reach more people
Sponsored Ads	Paid advertisements are inserted into users' feeds	Bypasses friend/follow restrictions; often less reviewed by automated filters
Group/Community Posts	Posts within any community or private groups	Attackers rely on group trust and limited oversight to spread harmful content more easily

Table 7 shows four main social media features that attackers use in AI-powered phishing campaigns. The feature column explains the way of using each category by attackers. For example, DMs for private messages, public posts for mass exposure, ads for targeted attacks, and groups for reaching specific communities.

4.1.4 Dimension 4: Target Profile

Attackers target high-profile executives, brands, and public figures by creating deepfake profiles on social media platforms. They spread fake news by compromising trustworthy accounts. They also use malicious links and social engineering to steal sensitive data and launch targeted attacks. The goal of narrative attacks against large companies is to affect their brand reputation. These threats require verifying profiles and using strong security measures to protect oneself from attacks. Taking steps such as multi-factor authentication, limiting the sharing of personal information, and having proper knowledge about phishing and scams can help prevent users from falling victim to attacks. Table 8 shows the three main target groups of AI-powered phishing attacks: regular social media users, company accounts, and influencers. The features and explanation are included in the table. Regular users don't take any steps to protect their information. As a result, they become frequent targets of attacks. Company accounts hold

valuable information and cause significant damage if hacked by attackers. Influencers have many followers, so if their accounts are hacked, harmful content can spread quickly.

Table 8 Different Target profiles by the Attackers

Category	Features	Explanation
Individual Users	General end-users with no special privileges, e.g., regular users	The largest group of victims: attackers always target them because they often lack awareness and protection
Corporate Accounts	Employees, executives, or official brand profiles	These targets are valuable because attackers can use them to steal money or sensitive business information
Influencers	High-reach or high-engagement personal brands, such as celebrities, leaders	This affects the company's reputation as people lose trust when the brand is linked to scams or security issues

4.1.5 Dimension 5: Delivery & Automation Pattern

Attackers deliver phishing messages in three ways: fully automated (AI handles everything and sends numerous generic messages), semi-automated (AI drafts messages that humans customize before sending), and manual (humans craft each message with some assistance from AI). Table 9 shows the categories of delivery and automation patterns. Fully automated campaigns use AI to generate, schedule, and send phishing messages without any human help. This allows them to target many users quickly and efficiently. Fully automated patterns enable mass targeting with minimal cost but often lack personalization.

Table 9 Delivery and Automation Pattern of Phishing Attacks

Category	Features	Explanation
Fully Automated	AI handles everything from creating content to sending it, without any human assistance.	Highly efficient for large-scale phishing with low cost
Semi-Automated	AI drafts are reviewed or customized by humans before being sent.	Combines personalized messages with a wide reach: frequently utilized in targeted attacks.
Manual	Human-written messages with little AI assistance	The scam employs sophisticated and persuasive techniques to steal money or information from prominent individuals.

Human judgment and AI's innovative ability are combined in semi-automated activities. An operator selects, modifies, and approves the templates or drafts created by large-language or image models before they are distributed, balancing efficiency with customized social engineering. While messages are crafted, AI supports manual delivery patterns. They require expert review and are challenging to discover. Defenders can select the best course of action by being aware of these trends. Rate limits and behavior anomaly detection are the best defenses against automated attacks. Content analysis and expert evaluation are often necessary for semi-automated or manual ones.

4.2 Taxonomy Validation

This section discusses the methods used to validate the proposed taxonomy of AI-driven phishing attacks on social media. The first approach involves analyzing six real-world scams from 2025 that illustrate how each incident aligns with the taxonomy's five dimensions. This analysis highlights the practical applications of taxonomy dimensions. The second method is a user-centered survey designed to assess how well users' experiences align with the core dimension of the taxonomy. By combining these two methods, the taxonomy's relevance, clarity, and practical value are effectively demonstrated for security professionals.

4.2.1 Validation Through Case Studies

Six real-world scams [32] are investigated below to validate the five core dimensions of the proposed taxonomy. Every scam has a different combination of platform manipulation, target profiles, attack delivery, and AI techniques. These cases confirm that the taxonomy can categorize current and emerging threats in a mutually exclusive, comprehensive, and practical manner.

- **Case 1: AI-powered Romance Scams (Pig Butchering)**

In a "pig butchering" scam, the attackers build a relationship slowly and carefully [33]. They use ChatGPT to create emotional conversations and face-swapping apps for fake video chats. They act like supportive and affectionate people until they conduct the conversation toward investment opportunities. These scams build trust over time and lead victims to fake investment platforms. The five core dimensions have analyzed of this attack as follows:

Table 10 Dimension Analysis of AI-powered Romance Scams

Dimension	Classification
Generative Model Modality	Hybrid (LLM Text + Deepfake Video)
Target Profile	General End-User (Emotionally vulnerable)
Phishing Vector	Direct Message / Private Chat
Social Platform Feature	Chat interface, profile video, story reply
Delivery & Automation Pattern	Semi-Automated

This case validates the multiple dimensions, such as a hybrid modality, emotional targeting, and feature exploitation. LLMs and deepfake technology are combined in the hybrid approach. The case highlights the need to recognize emotionally vulnerable individuals. The use of direct messaging and edited video content highlights the significance of phishing techniques and the misuse of social media platform functionalities. Since these interactions happen slowly and combine both human and AI input, it’s important to understand which parts are automated and which come from a real person.

- **Case 2: Deepfake Scams (Hong Kong $25M Incident)**

In this scam [34], the attackers used deepfake speech and video technology during a video conference to pretend as a company's chief financial officer and other employees. The chief officer was convinced that the conference was authentic by the real staff's images and voices. As a result, he approved a transfer of $25 million. The convincing deepfakes eventually neutralized his initial doubts about a message asking for a transaction. The dimensions analysis of this scam is shown in Table 11.

Table 11 Dimension Analysis of Deepfake Scams

Dimension	Classification
Generative Model Modality	Deepfake Audio/Video
Target Profile	Corporate Account (Finance Individuals)
Phishing Vector	Direct Message / Video Call
Social Platform Feature	Embedded video call or livestream
Delivery & Automation Pattern	Semi-Automated

This high-profile case demonstrates that the attacks can happen during corporate video calls and validates the applicability of deepfake audio/video modalities. It explains why real-time communication should be included under the Phishing Vector. To confirm the taxonomy's scalability beyond general consumers, it also supports extending the Target Profile to include corporate or enterprise-level targets.

- **Case 3: AI-powered Social Media Bots**

Bots use LLMs to post misleading comments, DMs, and fake interactions across social media platforms [36]. They mimic human behaviour and amplify scam content using AI to avoid detection. The AI-based social media bots are analyzed in terms of five core dimensions, as shown in Table 12.

Table 12 Dimension Analysis of Social Media Bots

Dimension	Classification
Generative Model Modality	LLM-Generated Text
Target Profile	General End-User
Phishing Vector	Public Post / Timeline Feed / Group Post
Social Platform Feature	Comments, replies, bot accounts, and reposts
Delivery & Automation Pattern	Fully Automated

This case validates the LLM-generated text modality and bot-based delivery as fully automated attacks. Public posts and comments are used by the bots to emphasize how important it is to separate them as social media features. Additionally, it provides a more comprehensive understanding of Target Profile, including both mass-targeted audiences and general users.

- **Case 4: AI-generated Phishing Emails & Ads**

Generative AI creates phishing messages with perfect grammar and tone, making the content indistinguishable from authentic information [35]. Social media advertisements and emails are used to spread the malicious content. This scam is examined for classifying five dimensions as shown in Table 13. This case supports the use of LLMs to create phishing content and validates the inclusion of Sponsored Ads and Public Posts as attack vectors. The automation pattern validates the importance of adding the category as Fully Automated Delivery. Additionally, it

illustrates how Phishing Vectors can use multi-platform channels and not just traditional inboxes.

Table 13 Dimension Analysis of E-mail and Ads Phishing

Dimension	Classification
Generative Model Modality	LLM-Generated Text
Target Profile	General End-User / Corporate
Phishing Vector	Public Post / Sponsored Ad
Social Platform Feature	Bio links, ad redirect buttons
Delivery & Automation Pattern	Fully Automated

- **Case 5: AI-powered Conversational Phishing (Chatbots)**

AI chatbots are increasingly used in different phishing attacks and take over the conversations once a user responds [36]. These bots use LLMs to maintain a natural tone, mimic a sense of urgency, and build false trust before delivering malicious links. The dimension analysis of AI-powered Conversational Phishing is shown in Table 14. This case explains the increasing use of AI chatbots for real-time manipulation in phishing attacks. It emphasizes the value of the LLM text modality and the necessity of documenting in-app chatbot interactions under social media Features. AI bots can conduct conversations without human assistance. This also provides strong evidence for classifying the attack under Fully Automated Delivery. This case also demonstrates the application of dynamic social engineering techniques by focusing on common end users.

Table 14 Dimension Analysis of AI Chatbots

Dimension	Classification
Generative Model Modality	LLM-Generated Text
Target Profile	General End-User
Phishing Vector	Direct Message
Social Platform Feature	In-app chatbot / DM
Delivery & Automation Pattern	Fully Automated

- **Case 6: AI-driven Investment Scams (Pump-and-Dump via Astroturfing)**

In this incident [37], the attackers used AI to create fake websites and social media profiles. These websites and profiles advertise fake investment possibilities, particularly in stocks and cryptocurrency. The astroturfing technique is used to manipulate public opinion and stock prices by spreading false support. AI may also simulate real-time trading, which makes fraud more difficult to detect by making it appear authentic. The AI-Driven Investment Scams dimension analysis is given in Table 15. This case validates community-targeted scams under target profiles and LLM text modality via fake reviews and posts. The use of social forums, polls, and group posts reinforces group-based phishing vectors and shows how platform features like comment threads and forums can be exploited. The taxonomy's coverage of scalable misinformation methods is validated by the fully automated bot-driven distribution.

Table 15 Dimension Analysis of Pump-and-Dump via Astroturfing

Dimension	Classification
Generative Model Modality	LLM-Generated Text (Astroturfing Content)
Target Profile	Specific Community (Crypto/finance groups)
Phishing Vector	Group Post / Forum Post / Public Comment
Social Platform Feature	Crypto forums, polls, comment threads
Delivery & Automation Pattern	Fully Automated

4.2.2 Validation Through Survey Data

To validate the taxonomy, this research conducted a survey using ten questions related to social media platforms. The survey collects data about users' real-world experiences, perceptions, and behaviors. The survey validates the taxonomy by demonstrating how these data align with the five proposed dimensions. The responses show that the taxonomy accurately captures how AI-powered phishing is perceived and encountered across social media. Each question was designed to capture a variable that aligns with at least one taxonomy dimension. Table 16 describes the validation of each dimension.

Table 16 The validation of the Proposed Taxonomy through Survey Questions

Survey Question	Dimension	How It Validates
1. What is your age group?	Target Profile	This question helps to identify how different demographic groups (especially age-based) experience phishing attacks.
2. What is your primary background?	Target Profile	This question assesses users' technical knowledge for classifying victims as non-technical, technical, or professional.
3. How many hours per day do you spend on social media?	Phishing Vector/Social Platform Features	It shows that people who use the platform more frequently are more likely to face certain security threats.
4. Which three platforms do you use most often?	Phishing Vector/Social Platform Features	This identifies which platforms are most relevant for taxonomy coverage. This confirms the need for platform-specific features in the proposed model.
5. Have you heard of generative AI tools?	Generative Model Modality	This helps to check if users are familiar with AI Tools to see if these categories make sense to include.
6. Do you believe AI tools can create convincing phishing messages?	Generative Model Modality/Delivery & Automation Pattern	This question helps to confirm whether users identify AI-generated scams as realistic. It validates both the presence and believability of automated content.
7. Have you ever encountered a suspicious message that you think was AI-generated?	Phishing Vector/Delivery and Automation Pattern	This validates if real-world attacks match the proposed taxonomy and confirms the use of specific methods and automation.
8. If yes, in which form did that suspicious content appear?	Phishing Vector/Social Platform Features Exploited	This question directly validates the relevance of delivery channels (e.g., DM, ad, story).
9. How confident are you in identifying whether a message is AI-generated versus human-written?	Target Profile/Generative Model Modality	This question informs how different users perceive modality by validating the core dimension. It identifies the need to adopt training and detection subcategories.
10. Suppose you receive a message from someone you "know" (influencer or friend) that has a minor change in their name and includes a link. How likely are you to click that link?	Target Profile/Delivery and Automation Pattern	This question helps to capture behavioral risk by user type and validates the target profile. It also shows how automation can increase the success rate of phishing attacks.

4.2.3 Validation Through Expert Interviews

To ensure a comprehensive validation of the proposed taxonomy, this research adopts a mixed-methods approach. In this section, the description of quantitative validation is provided by expert interviews.

Table 17 Validation of the proposed taxonomy by conducting expert interviews

Interview Question	Validated Taxonomy Dimensions	Validation Purpose
Can you describe any phishing attempt you've experienced or heard about recently on social media?	All Dimensions	This question captures complete real-world scenarios to test the full proposed taxonomy.
Did the content seem human-written or AI-generated? What made you think so?	Generative Model Modality	This helps to test user awareness of AI-generated content and supports modality classification.
Through what channel or feature did the phishing message reach you (e.g., DM, comment, ad, video)?	Phishing Vector, Social Platform Feature Exploited	This question identifies the platform-specific attack vector and the exploited feature.
What part of the message or interaction seemed most suspicious to you?	Delivery & Automation Pattern	The signs of automation or social engineering attacks can be assessed by this question.
Who do you think the attackers were targeting, and why?	Target Profile	The question reveals target demographics and supports user profiling accuracy.
Did the phishing message seem like it was made just for you? If yes, how?	Target Profile, Delivery & Automation Pattern	Measures message customization and help to identify automation patterns.
Was there any indication that the message was automated or part of a bot-like interaction?	Delivery & Automation Pattern	This question aims to confirm whether participants noticed signs of automation, such as repetitive patterns or a lack of personalized context.
Which social media features (e.g., stories, polls, live streams) do you think are easiest to exploit?	Social Platform Feature Exploited	This question investigates participants' awareness of how specific social media features may be exploited. This supports the taxonomy's focus on platform-specific vulnerabilities.
Do you believe AI-generated phishing is harder to detect than traditional phishing? Why or why not?	Generative Model Modality, Target Profile	This question evaluates participants' understanding of AI-generated phishing.
If you were to categorize the phishing attacks, what words or criteria would you use (e.g., delivery type, target, intent)?	All Dimensions	This question describes how participants categorize phishing attacks and whether their responses align with the proposed taxonomy's dimensions.

Ten semi-structured interview questions were prepared to ask all participants. These questions were designed to understand the user experiences with phishing attacks on social media and to analyse how those experiences align with the five dimensions of the proposed taxonomy. Table 17 shows how every question validates the proposed taxonomy dimensions.

5 Defense Framework Architecture

This chapter describes a conceptual design for a multi-layered defense framework that detects, investigates, and mitigates AI-powered phishing campaigns. The chapter begins with the design objectives and requirements of the defense framework, followed by a high-level architectural design. Additionally, this research describes the main components and their functions in the following section. The section concludes with notes on prototyping considerations and integration with existing platform processes.

5.1 AI-powered Phishing Tools and Techniques

AI has made phishing attacks more sophisticated. AI-powered phishing generates realistic and personalized messages based on the user´s information. AI-powered Phishing makes them more difficult to identify than traditional scams that deliver the same message to many victims. These days, cybercriminals can create deepfakes: fake calls or videos that seem real by using AI to mimic voices and pictures. These strategies are employed to mislead users into approving unauthorized transactions or disclosing personal information [34]. It is necessary to understand the techniques and tools used by attackers in various phishing attacks before developing a defense strategy. There are some common phishing tools and strategies powered by AI that are used by attackers on many platforms. Table 18 shows the different AI-based tools and techniques used in phishing attacks.

Table 18 AI-based Tools and Techniques that are used in designing phishing content

Category	Tool/Technique	Description
Email Generators	Chatbots and LLMs such as ChatGPT and Gemini	This tool generates realistic, error-free, and highly personalized phishing emails.
	Email Spoofing (Deep Learning)	In this category, malicious emails are sent to the target user by analysing and mimicking the writing style of reliable contacts. Email spoofing increases users' trust.
Social Engineering and Vishing	Deepfake Voice techniques such as Voicify AI, ElevenLabs	These techniques are used to create synthetic voice messages of real executives/managers for vishing calls.
	Social Media platform scraping, such as Maltego, OpenAI Whisper	Attackers use this tool to collect personal information from public profiles to make phishing messages more trustworthy.
Image-Based Attacks	Deepfake Image Generators, such as Stable Diffusion, Midjourney, DeepFaceLab	These tools create realistic photos of trusted individuals or companies with their logos to use in fake conversations.

	CAPTCHA Bypass	During automated phishing attacks, attackers use a CAPTCHA solution to get access over website security measures.
Fake Website Cloning	Automated Website Cloners	These tools are used to replicate real login pages to steal information from Banks or email services.
	Keystroke and Behavior Analysis	Analyze typing patterns or mouse movements to bypass two-factor authentication.
Spear Phishing Automation	Automation Bots, such as Selenium, Puppeteer	These tools automatically send thousands of personalized phishing emails.
	Malware Generator, such as WormGPT, FraudGPT	These are specialized “underground” LLMs trained to generate malicious code, malware templates, and targeted phishing content.

5.2 Design Objectives and Requirements

The AI-powered phishing detection and prevention system is an advanced cyber security method that is designed to identify and mitigate phishing attacks on different social media platforms. This system works by collecting information from sources like emails, URLs, websites, and network traffic. Social media is a common place for attackers because of its vast amount of information. Attackers are changing their methods and tricks for every new attack. So, the defense framework needs to be kept updated. The design process of the defense framework for an AI-driven phishing attack on social media platforms is discussed in this chapter. The main objective of the defense framework is to identify, neutralize, and mitigate sophisticated phishing attacks by using different AI techniques. The AI techniques are given below:

- Machine learning (ML)
- Deep learning (DL)
- Natural language processing (NLP).

The framework aims to provide a strong, flexible, and transparent defense against advanced phishing attacks. The primary goal is to reach complete multimodal coverage, which will allow the system to identify phishing content in text, photos, audio, and video on social media platforms. In this recent era, attackers are using AI-generated content in phishing attacks. So,

depending on only one detection technique would create significant gaps. The framework also addresses the challenges posed by deepfake technology, voice phishing (vishing), automated social engineering, and highly personalized phishing campaigns. There are some essential requirements to design an effective defense against AI-powered phishing attacks. These requirements will help to detect the threat efficiently and protect the user from advanced phishing scams. The requirements are given below:

- Real-time detection of AI-generated phishing content on different social media platforms.
- Robust integration capabilities with various platforms such as Twitter, Facebook, Instagram, LinkedIn, and external intelligence databases like PhishTank and VirusTotal. PhishTank is a database of phishing websites. On the other hand, VirusTotal collects data about files and URLs to detect viruses, malware, and other threats.
- Scalable response and mitigation strategies capable of swift intervention.
- Continuous learning mechanisms to adapt effectively to evolving phishing methodologies and attack patterns.

5.3 Framework Design

This research proposed a multi-layered defense framework designed to prevent AI-based phishing attacks. This model aims to prevent, mitigate, and detect AI-generated phishing attacks on social media. The multi-layered defense framework is shown in Figure 23.

The architecture of the defense framework comprises four different layers designed to mitigate social media phishing attacks. The layers are:

1. Integration Layer
2. Detection Layer
3. Response & Mitigation Layer
4. Feedback & Learning Loop

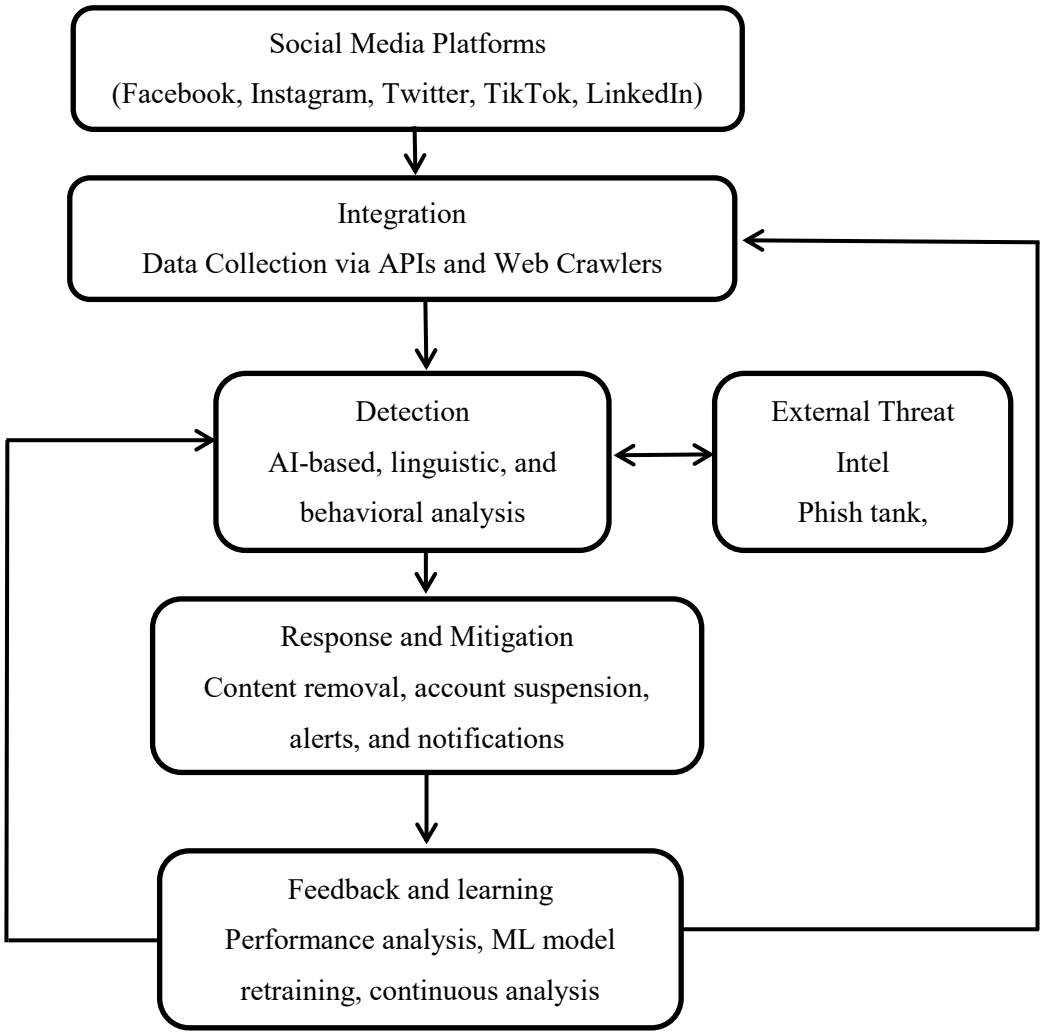

Figure 23 Multi-layered defense framework architecture

Each layer is connected to show how data moves and actions are taken. The first step is gathering data from websites such as LinkedIn, Instagram, Twitter, and Facebook. All the data is gathered in the Integration layer via APIs. The information is being improved with threat analysis from resources such as VirusTotal and PhishTank. Then, AI is used to analyze the content in the detection layer. These models use visual, behavioral, and linguistic analysis to detect deepfake content, impersonation, and phishing messages. In the next step, the Response & Mitigation layer is activated. It automatically takes some actions, such as content removal, account suspension, and notifying users who may have been affected. Finally, the Feedback & Learning Loop turns off the system. It evaluates the system's performance and gathers feedback from the user. Then, it uses new data to retrain models. This also ensures that the defense evolves continuously in response to emerging threats. The following flow chart shows the process of detecting AI-generated content.

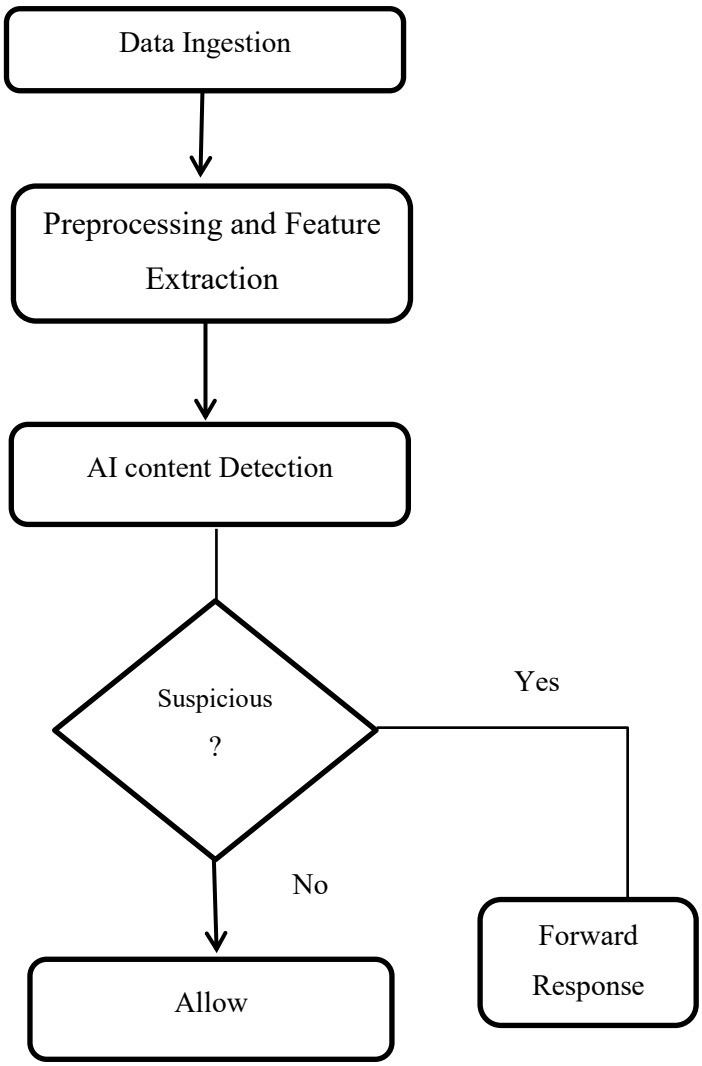

Figure 24 The detection process of AI content as a part of a defense framework

5.3.1 Integration Layer

The system is started by collecting various types of content, like posts, direct messages, ads, and group chats from social media platforms. The user's content and information are monitored for phishing threats. It connects to these platforms through official APIs to get data securely and reliably. It also gathers additional threat information from trusted sources, such as PhishTank and VirusTotal. All incoming data is cleaned up and placed into a single, standard format to work easily. The system is designed to handle both public and private information safely by using encryption to protect the data. It also includes tools to avoid getting blocked by platforms due to excessive or rapid data collection. It uses message queues (like Kafka) to keep the data in the right order for later analysis.

5.3.2 Detection Layer

The use of AI specifically Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP), has significantly enhanced phishing detection [45]. These methods help to recognize the signs of malicious activity and learn patterns from data, hence increasing the accuracy of phishing threat detection.

- **Machine Learning (ML) Approaches**: Machine learning models are used to detect any phishing attacks because of their ability to learn from labelled datasets and generalize to unknown threats. There are some commonly used algorithms for phishing detection, like decision trees, Support Vector Machines (SVMs), K-Nearest Neighbours (KNN), and random forests. Decision trees are useful for feature-based detection such as URL length [38]. SVMs perform well with clear margin separation [39]. Finally, KNNs are simple and non-parametric [40]. These algorithms can effectively identify and detect malicious activities based on data patterns. These models use lexical patterns, WHOIS information, email metadata, URL structures, and domain age to categorize emails, webpages, and messages. A key advantage of ML is its adaptability to evolving attack vectors, including zero-day threats. However, issues like data imbalance can decrease effectiveness, where there are many fewer phishing samples than authentic ones. To mitigate this problem, methods like SMOTE (Synthetic Minority Over-sampling Technique), under sampling, and oversampling are utilized. Performance optimization still depends on efficient feature engineering and selection.

- **Deep Learning (DL) Techniques:** Deep learning models provide better performance by automatically learning complex and hierarchical features from raw data. This model also eliminates the need for manual feature engineering. There are two popular DL architectures, such as Convolutional Neural Networks and Recurrent Neural Networks [41]. Convolutional Neural Networks (CNNs) are used to examine phishing websites and messages for structural or visual patterns. On the other hand, Email and message content can effectively capture sequential connections via recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) networks.

- **Natural Language Processing (NLP) Applications:** NLP help to process, understand, interpret, and generate human language by using software [42]. These applications are widely used in phishing detection. These methods are used in analyzing the text on webpages and emails. The goal is to identify the characteristics that distinguish phishing

contents from normal texts. NLP can analyze messages or emails to detect suspicious words, mismatched domain names, and spelling or grammatical anomalies [43]. There are following three techniques that are included in the NLP Application:

1. Text classification identifies whether a message is authentic or phishing.
2. Emotional manipulation can be identified with the help of sentiment analysis.
3. Advanced models such as Bidirectional Encode Representations from Transformers (BERT) and other transformer-based architectures are used in semantic and contextual analysis.

NLP helps to comprehend the language's intention in the systems. In addition, social engineering techniques and psychological manipulation are frequently used in phishing attacks.

5.3.3 Response & Mitigation Layer

The Response and Mitigation Layer is designed to prevent AI-driven phishing attacks by combining some automated methods like risk scoring, user protection, and human oversight. When phishing attacks are detected, the system automatically blocks or remove malicious content and suspends phishing accounts, and blacklists dangerous URLs via platform APIs. AI-generated risk scores serve as the basis for threat classification. High-risk items are escalated immediately to security teams, while medium-risk items are queued for human analyst review. Low-risk events are monitored for continuous learning and model improvement. The system displays in-app alerts to protect users from any phishing risks. It recommends password changes and implements multi-factor authentication. These responses help to minimize risks and secure user accounts after an incident. It sends targeted notifications and training materials to more vulnerable users, such as executives or finance staff. The framework ensures that all responses are logged and auditable. This system also enables security analysts to modify decisions. This approach uses both automation and human judgment to quickly stop clear threats. It also helps handle complex cases more accurately.

5.3.4 Feedback & Learning Loop

The feedback and learning loop help to improve phishing detection by using system data and user feedback. AI models are regularly updated with new threats to stay effective and accurate. Detection algorithms are improved using reinforcement learning and behavior analysis. This

helps the system get more accurate over time. The malicious content can be scored depending on the risk score as follows:

- High-risk (score > 0.9): Immediately send alerts to security teams via API.
- Medium-risk (0.7–0.9): Queued for human analyst review.

When phishing is confirmed, the system can eliminate malicious content, disable accounts, and stop scam ads. Users will receive notifications and account security advice through app alerts or SMS. All activities are recorded so that analysts can examine and modify automatic decisions. The system can quickly correct errors, prevent excessive notifications, and follow the legal requirements.

6 Results and Discussion

This chapter is divided into three parts. The survey results are analysed in the first section using pie and bar charts. The pie and bar charts help to understand the scenario of the users' experience and awareness properly. In the following two sections, the summary of the experts' interviews is provided with thematic analysis. The thematic analysis result from the interviews is described using Braun and Clarke's six-phase framework [9].

6.1 Survey Findings

The survey gathered a total of 37 responses from social media users. The survey provides an overview of users' awareness, experience, and detection confidence of AI-powered phishing attacks. The result from each question is described below:

Around 72% of participants were between 18 and 34 years old, and the rest of the participants were between 35 to 44 years old. Only one participant under the age of 18 responded. This result indicates that mainly young adults are highly active on social media, and phishing awareness and confidence vary by age. This result also indicates that attackers may design their strategies based on the target's age demographic. The pie chart shown in Figure 25 presents the Demographic results by different Age groups.

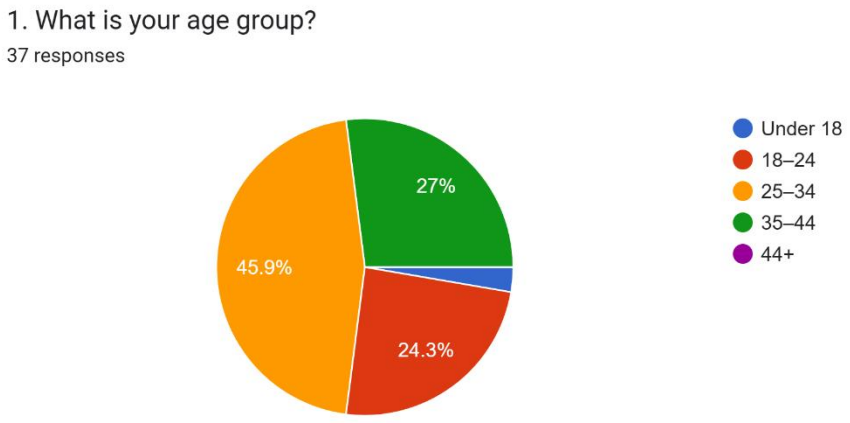

Figure 25 Distribution of respondents by age group

All the participants had different technical backgrounds. 37.8% have some technical knowledge, while 29.7% are identified as non-technical. Only 32.4% had a cyber security or IT background. Technical expertise affects the user's ability to detect phishing content. The survey

results show clear differences in detection confidence by skill level, which highlight the need to categorize user types in the proposed taxonomy. Figure 26 shows the participants' Technical Expertise Levels.

Figure 26 Respondents' technical expertise levels

Around 76% of participants spend more than three hours per day on different social media platforms, followed by almost 22% who responded only spend 1-2 hours. Participants who are spending a lot of time on social media platforms are more likely to be targeted by attackers. Figure 27 shows the percentage of time spent on social media platforms daily.

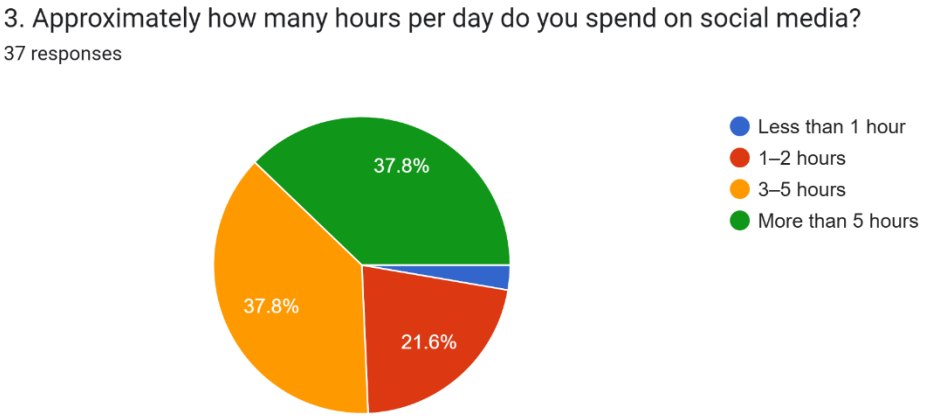

Figure 27 Amount of time spent on social media platforms daily

The survey response to this question found that Facebook and Instagram are the most widely used platforms among the participants. Almost all the participants are using these two

platforms. Out of 37 respondents, 30 people are using Facebook and Instagram, followed by 24 people using YouTube, 15 people using LinkedIn, and 9 people using Twitter. Other platforms such as Twitter, TikTok, and Reddit were selected by the less participants. Facebook and Instagram are not only popular but also High-Risk Platforms for Phishing Attacks, as shown in Figure 28. The result from the survey confirms that any defense framework or taxonomy of any AI-based phishing must include platform-specific feature exploitation.

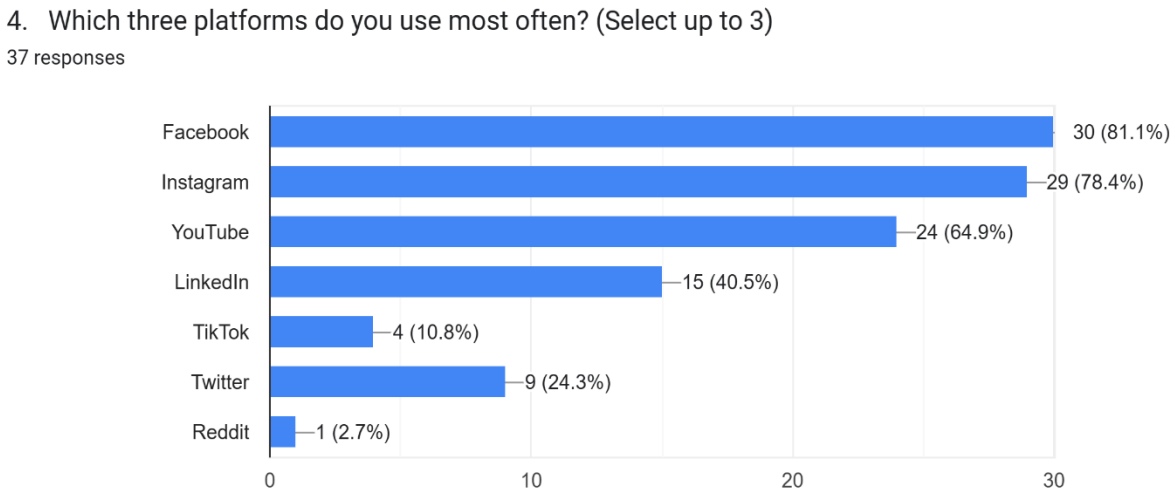

Figure 28 Most frequently used social media platforms

The survey results show that 94.6% of participants were familiar with generative AI tools such as ChatGPT and DALL·E. Only 6% responded “No” or were unsure. This result confirms that people know about AI-generated content, as shown in Figure 29. As the different tools are easy to use, all the respondents know about the tools.

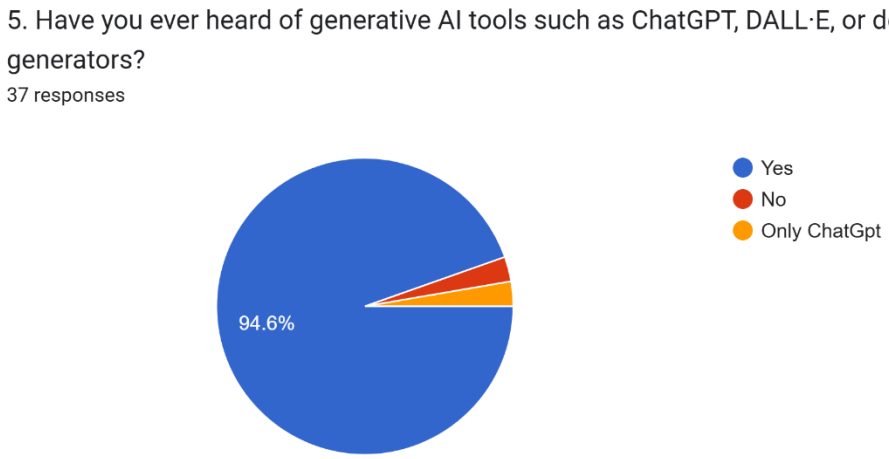

Figure 29 Awareness of generative AI tools among respondents

In Figure 30, 91.9% of respondents answered “Yes,” and the remaining respondents answered “Maybe” or “No”. This result confirms that social media users believe that AI tools can be used to generate malicious content. This result also shows that social media users are becoming aware of how AI technology is used in real-world attacks. The pie charts show the percentage of how many people who believe that AI tools can be used to create Phishing content.

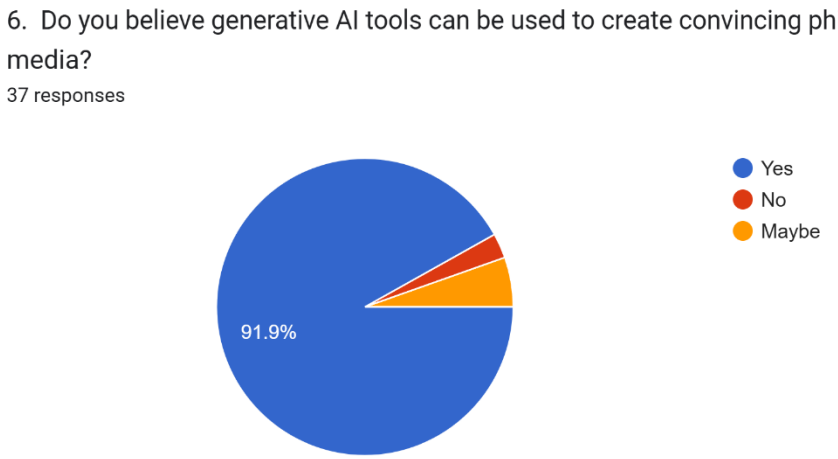

Figure 30 Users’ views on generative AI’s content for phishing

86.5% of participants responded “Yes” to this question, as shown in Figure 31. The remaining participants answered “No” or “Not Sure.” This result indicates real-world confirmation that users are actively encountering a malicious message or post on social media. The response

result validates the Generative Model Modality and Phishing Vector dimensions of the proposed taxonomy by showing that both content generation method and delivery channel are meaningful, observable, and relevant.

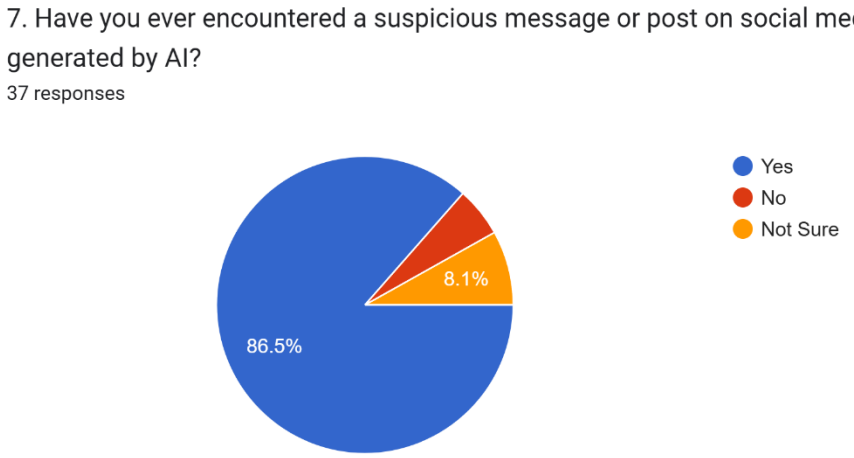

Figure 31 Frequency of facing malicious content

Respondents who indicated that they had encountered suspicious content were asked to identify the form of malicious content in which it appeared. The results showed the following distribution in the pie chart. The results showed the following distribution, as shown in Figure 32. 56.8% through a Fake profile or sponsored ad, 16.2% through Direct message, 10.8% through public messages or suspicious links. The remaining 16.2% is through video content and other content. This question provides direct evidence supporting the Phishing Vector dimension in the proposed taxonomy, which classifies how phishing content reaches the users.

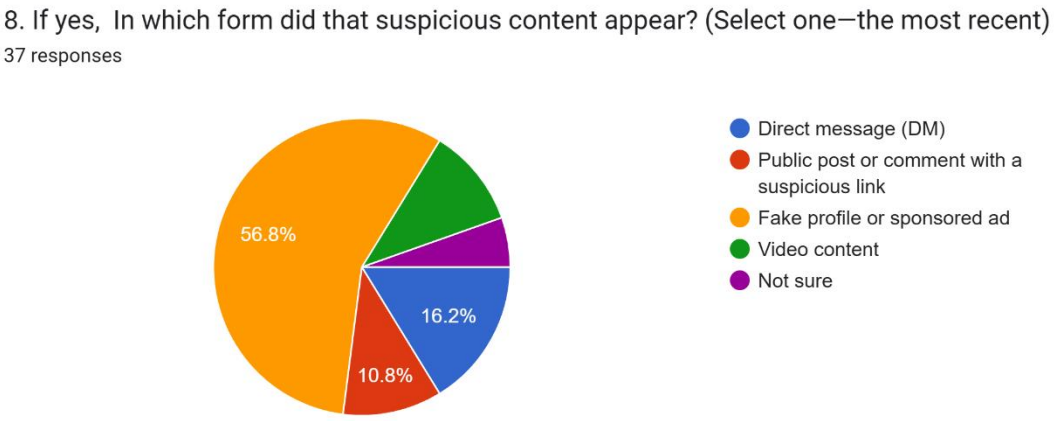

Figure 32 Most common forms of suspected ai-driven phishing content

As we can see in Figure 33, the participants rated their confidence level in identifying phishing attacks on social media from 1 (no confidence) to 10 (complete confidence). The breakdown was 13.5% low confidence (1–4), 35.1% moderate or neutral (5–6), and 51.3% high confidence (7–10). Among them, only 8.1% of participants think they have the highest confidence of 9 or 10. The result from this question strongly supports the inclusion of the Target Profile dimension in the proposed taxonomy. This also helps to classify users based on characteristics like digital literacy, experience level, and vulnerability.

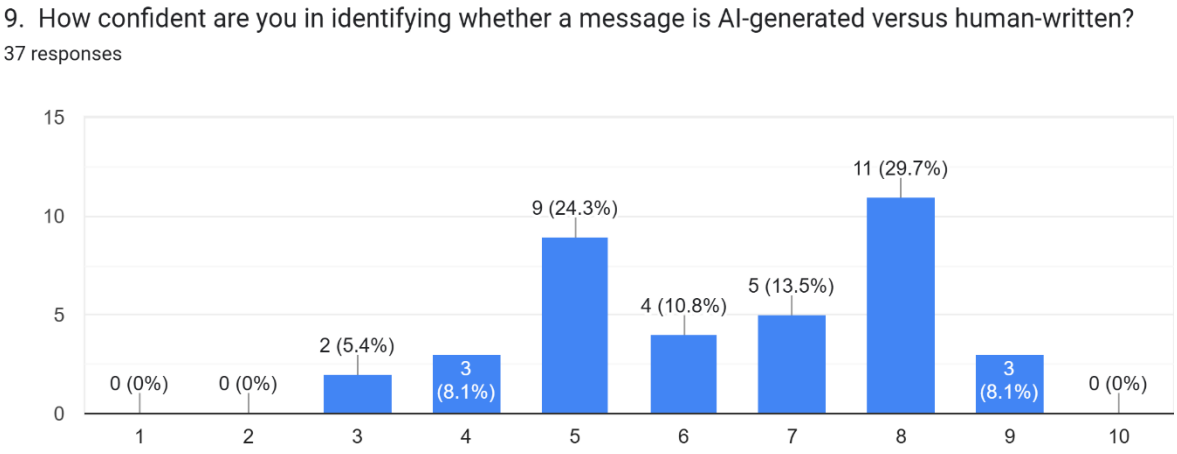

Figure 33 Confidence levels in distinguishing AI-generated messages

In this question, participants were asked how likely they would be to click on a suspicious message from someone pretending to be familiar. Figure 34 shows that the responses revealed that 56% were unlikely or very unlikely to click, 22% were neutral, and 22% indicated they were likely or very likely to click. Almost 8% (Likely + Very Likely) would still consider clicking such a link. The results from Q10 confirm the real-world effectiveness of automated attacks and validate both the Delivery & Automation Pattern and Target Profile dimensions of the proposed taxonomy.

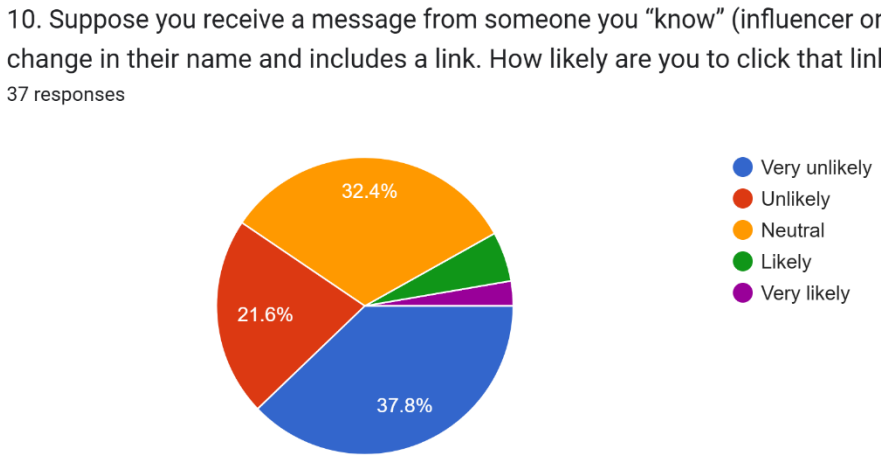

Figure 34 The possibility of clicking suspicious messages from a person pretending to be familiar

The survey results show that most participants (ages 18–34) are familiar with generative AI techniques. The participants believe that AI tools can be used to detect convincing phishing attacks. While 75% of participants have experienced malicious content or a phishing attack on social media. Non-technical users experienced the most issues due to their lack of confidence in identifying AI-generated phishing. The common attack vectors reported by the respondents include fake profiles, DMs, and sponsored ads. Users mostly experience the attacks on Facebook and Instagram. Moreover, behavioral responses revealed that over 40% of users were neutral or even likely to click on a suspicious message, especially when it appeared to come from a familiar-looking source. These findings validate all five core dimensions of the proposed taxonomy, especially Generative Model Modality, Phishing Vector, and Target Profile, by confirming their relevance to real-world user experiences.

6.2 Interview Insights

The interview summary of all five participants is given below:

6.2.1 Participant 1 (Cyber Security Specialist)

The participant recently experienced a phishing scam on Instagram. Fake profiles were created by the attacker to pretend to be Meta's support team. These accounts were sending messages regarding the profiles' copyright violations. The user's profile would be disabled unless they clicked on a link. A similar scam was observed on Facebook that involved a fake verification method. The content of the messages was like AI-generated or based on a template. The grammar was clean and formal but lacked personal touches, such as addressing the recipient by name. It appeared very generic. The phishing attack was delivered through direct messages on Instagram. A friend had forwarded the message to ask if it was legitimate. The attacker had also left a comment under his friend's post and told them to check their inbox immediately. The account had some followers and a recently added profile picture, which was more suspicious. According to the participant, the attackers were targeting university students, especially those under academic pressure or deadline stress. They are making them more likely to fall for impersonated academic or institutional messages. The overall incident was financial fraud, and the malicious content was sent through direct messaging.

6.2.2 Participant 2 (Engineer and Instagram Content Creator)

The participant never experienced a phishing attack, but he saw multiple phishing contents on social media platforms. He described fake cryptocurrency giveaways as phishing on platforms like Twitter, Facebook, and Instagram, using fake accounts. According to his statement, he thought the messages were human-written, but then he identified the content as AI-generated. Because the message was flawless and persuasive. Phishing came via comments, DMs, sponsored ads, and fake profiles. There were some suspicious signs, including too-good-to-be-true offers, urgency, and different usernames. The main target profiles were crypto enthusiasts, newcomers, and followers of public figures. Some attacks were personalized using replies or public data. Signs of bots included rapid posting, generic replies, and inconsistent profiles. He thinks that the easiest features to exploit are DMs, comments, ads, fake profiles, live streams, and stories. He also believes AI phishing is harder to detect due to better language, personalization, scale, and deepfakes. Finally, he categorized the attack as social media delivery, targeting crypto users, using fake profiles, and urgency for financial fraud.

6.2.3 Participant 3 (A General User)

Participant 3 has not personally experienced a phishing attack on social media, but some of their relatives have. These relatives received phishing emails or messages pretending to be about upcoming deliveries. The attackers claimed there was an issue with the delivery address and asked the recipients to re-enter their card details. They were asking for a small payment, which ultimately led to financial fraud. When asked whether the content seemed human-written or AI-generated, the participant said that most phishing messages appeared to be human-made because he thinks they contain spelling mistakes. The phishing attacks reached their targets through various channels, including emails, direct messages, Instagram ads, and similar digital platforms. The participant found the URLs, attachments, or the structure of the webpages themselves to be the most suspicious aspects of these interactions. They believed the attackers were motivated by financial gain. Their main aim was to steal bank or card details. There was no indication that any of the phishing attempts were personalized. However, he did observe that many Instagram ads and URLs seemed automated and suggested some level of bot activity. According to him, many Instagram ads and URLs are automated, and stories and feeds are easiest to exploit. He believes AI-generated phishing is harder to detect because the content looks legit.

6.2.4 Participant 4 (Cyber security Student)

This participant said that he saw a fake iPhone giveaway on Instagram asking users to click and enter their personal information. The message felt AI-generated because it was too polished and generic. It appeared as a comment, and it was tagged by some random users. The fake URL and urgency (“claim within 30 minutes”) were clear red flags. He believed it was designed to target younger users or those who were seeking free gadgets. It was not personalized, but many users were tagged. This was a sign of automation, like repeated comments and an empty profile. He also added that stories and polls with swipe-up links are easily exploited. He believes that AI phishing is harder to detect due to natural language and better grammar.

6.2.5 Participant 5 (Facebook Content Creator)

The participant experienced a phishing attack directly through Steam. A message came from someone on their friend list. The message mentioned that she had won a $50 Steam Wallet code via a giveaway. They also send a link to claim the giveaway. The message seemed to be credible because it came from a trusted friend’s account. However, the language felt suspicious because

it was grammatically perfect but robotic. There was a lack of any personal greeting or reference, which made it feel either AI-generated or copy-pasted. The phishing content was delivered through Steam's direct messaging system. That was suspicious for her because the friend had never messaged about giveaways or anything money-related before. The most suspicious element was the link because it was a fake domain resembling Steam. The participant believed the attackers were targeting content creators and gamers. The attackers were targeting younger or less experienced users who might be tempted by offers of free wallet credit. Such scams can spread quickly once an account is compromised because the same message can be sent to other people on the friend list. This phishing attack was categorized as a credential theft or account takeover attempt by using social engineering through a compromised account.

7 Conclusion

The conclusion part is divided into three sections. The first section describes the key contributions of the research. In the next section, the proposal for a community resource is given to keep continuous engagement with the proposed taxonomy and defense framework. Finally, some future scope of this research is added.

7.1 Summary of the Key Contributions

The key contributions of this research are summarized as follows:

1. **Development of a comprehensive Taxonomy:** The research introduces five core dimensions of the taxonomy of AI-generated phishing attacks on social media platforms. When the existing models only focused on email or static threats, this taxonomy includes dimensions such as generative modality, phishing vector, social platform feature exploited, target profile, and delivery/automation pattern to classify AI-generated content.

2. **Design of a Multi-Layered Defense Framework:** A four-layer defense framework was developed to prevent, detect, and respond to AI-powered phishing on social media. It includes data integration, detection, automated response actions, and a feedback loop for continuous learning. This structure ensures adaptive protection against AI-based phishing techniques.

3. **Validation through Case Studies, Survey, and Interviews:** The proposed taxonomy was validated using real-world AI-powered phishing case studies that were analyzed to validate all five taxonomy dimensions. Furthermore, a survey of social media users and expert interviews was conducted to provide qualitative and quantitative insights. The description of how the survey and interview questions validate the proposed taxonomy is also provided.

4. **Behavioral Insights into User Vulnerability:** This research highlights the key behavioral trends, such as low confidence in detecting AI-generated phishing and oversharing on social media, which contribute to the success of these attacks. These findings directly mentioned the need for the defense framework and the prioritization of awareness-based methods.

5. **Proposal of a Community-Based Resource Platform:** This research provides a Community-Based Resource Platform design for users, educators, and researchers. They

can explore real phishing examples and conduct self-assessments about AI-based phishing. This helps the research to keep updated, useful, flexible, and relevant over time.

7.2 Proposal for a Community Resource

To make this research more beneficial, a future initiative can be developed based on the proposed taxonomy and defense framework. The method can be taken as a form of an online phishing risk assessment tool. Users, educators, and organizations can use this to categorize the real examples of AI-generated phishing, test users' awareness, or access prevention guidelines. This research gives a validated structure for understanding and mitigating phishing threats on social media platforms. However, it also requires ongoing adaptation and user engagement beyond the research. The proposal for a community resource can fill the gap between academic research and public awareness. A fixed taxonomy or defense model will become outdated as AI-generated phishing attacks are becoming more sophisticated. So, by proposing a collaborative platform, the proposed taxonomy and framework stay updated through community updates, encouraging input from other experts and platforms. The following figure 35 illustrates a proposal for a community resource. The proposed taxonomy provides a comprehensive way to analyze any cyber threats and cyber security strategies for EMS. The community research consists of the following three layers:

- Foundational Layer
- Platform Core Layer
- End User Layer

The main components of the foundation layer are Validated Taxonomy and Multi-layered Defense Framework. This layer represents the core output from this research. It serves as the foundation for the entire community resource. The platform core layer consists of four key modules: The first module is a library of categorized examples of AI-generated phishing based on real-world case studies and taxonomy dimensions. The second module is self-assessment and prevention tools that help users evaluate their vulnerability and improve awareness. A live dashboard is added to the model to track emerging phishing techniques. Finally, a collaborative update framework is used to enable researchers and educators to contribute to the improvements of the taxonomy over time. It allows cyber security experts, educators, and developers to update the taxonomy by submitting new attack examples and refining defense strategies. Regular users,

educators, and cyber security researchers are the main contributors to the resource. To keep the research applicable, flexible, and relevant, it promotes individual learning, continuous observation, and group contributions.

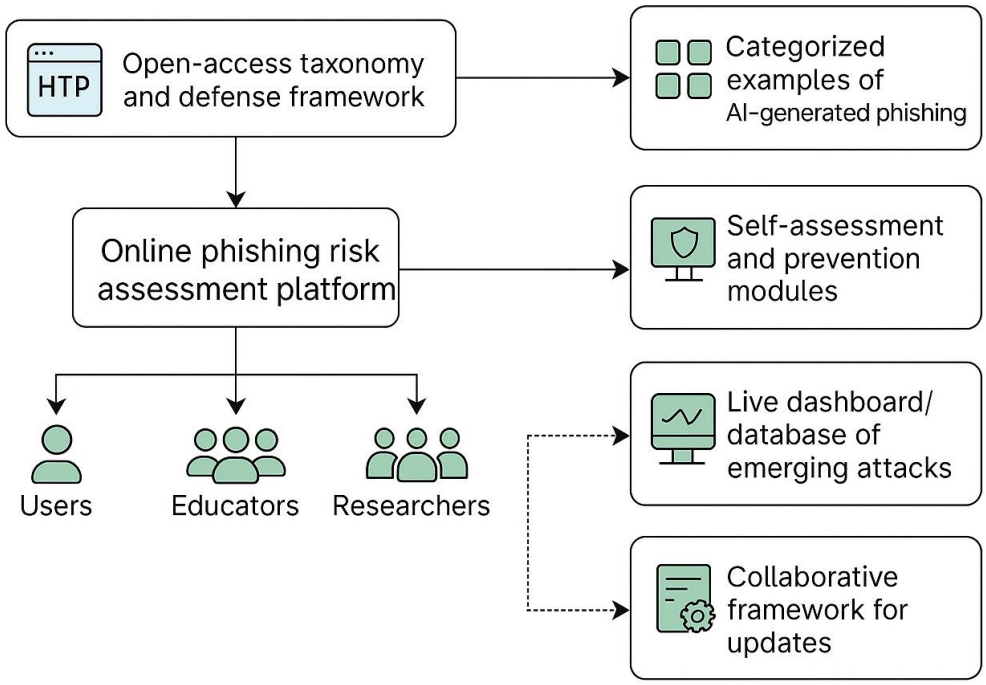

Figure 35 Proposal for a community resource

The live dashboard and collaborative framework are kept separate to show their distinct roles in the community resource design. The dashboard describes the real-time user awareness, and the collaborative framework examines the taxonomy evolution. This ensures that the platform remains both informative and adaptable over time. In addition, if the dashboard fails or changes, it won't affect the editorial logic of the taxonomy update workflow.

As AI-powered phishing attacks on social media are becoming more advanced, this research highlights the urgent need for long-term community-based solutions that will fix the academic models. The proposed community resource is based on the proposed phishing taxonomy and defense framework that offers a practical and flexible approach to cyber security. This initiative converts complex phishing threats into simple tools and teaches people to stay safe. It creates a stronger, team-based defense against future AI-powered attacks by getting help from experts.

7.3 Future Research

This research includes a taxonomy design, a multi-layered defense framework, survey/interview validation and analysis, recent case studies, and the proposal of a community resource of AI-generated Phishing attacks on social media. There are still some scopes for future work. The future recommendations include as following:

- The future taxonomies can be improved by adopting user behavior dimensions such as engagement patterns, click tendencies, and decision fatigue.
- The research will become more effective if modular components can be added with sub-classifications like platform features and culture.
- Platform-level implementation and Collaboration can be done to mitigate phishing attacks on social media. Future research should work with social media platforms to implement and test the defense framework in real environments.
- A live detection system can be implemented on social media by using AI tools to detect phishing content instantly.
- Future projects should build the proposed community resource into a live, user-accessible platform that offers awareness tools, phishing simulations, and interactive defense training.

References

[1] C. Hewage, L. Nawaf, I. A. Khan, and Z. Alkhalil, "Phishing attacks: A recent comprehensive study and a new anatomy," Frontiers in Computer Science, vol. 3, art. no. 563060, 2021. [Online]. Available: https://doi.org/10.3389/fcomp.2021.563060

[2] Valimail, "Complete Guide to Phishing: Techniques & Mitigations". [Online]. Available: https://www.valimail.com/resources/guides/guide-to-phishing/

[3] B. Dean, "Social Media Usage & Growth Statistics," Feb 10, 2025. [Online]. Available: https://backlinko.com/social-media-users

[4] E. Fletcher, "Social media a gold mine for scammers in 2021," Federal Trade Commission, Jan. 25, 2022. [Online]. Available: https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/01/social-media-gold-mine-scammers-202

[5] S. S. Harihar and M. Potdar, "A comprehensive analysis of phishing challenges and AI solutions adopted by IT organizations," J. Neonatal Surg., vol. 14, no. 22s, pp. 17–25, 2025. [Online]. Available: https://www.jneonatalsurg.com/index.php/jns/article/view/5427

[6] E. Mouncey and S. Ciobotaru, "Phishing scams on social media: An evaluation of cyber awareness education on impact and effectiveness," J. Econ. Criminol., vol. 7, no. 1, art. no. 100125, 2025. [Online]. Available: https://doi.org/10.1016/j.jeconc.2025.100125

[7] H. Naseer, "Understanding and exploring core social media features," Sprout Social, Feb. 6, 2023. [Online]. Available: https://sproutsocial.com/insights/social-media-features/

[8] S. J. Dixon, "Instagram: Distribution of global audiences 2025, by age and gender," Statista, Jun. 2, 2025. [Online]. Available: https://www.statista.com/statistics/248769/age-distribution-of-worldwide-instagram-users/

[9] S. K. Ahmed et al., "Using thematic analysis in qualitative research," J. Med. Surg. Public Health, vol. 6, art. no. 100198, 2025. [Online]. Available: https://doi.org/10.1016/j.glmedi.2025.100198

[10] V. Bhavsar, A. Kadlak, and S. Sharma, "Study on phishing attacks," Int. J. Comput. Appl., vol. 182, no. 33, pp. 27–29, 2018. [Online]. Available: https://doi.org/10.5120/ijca2018918286

[11] Mailgun, "The golden age of scammers: AI-powered phishing." [Online]. Available: https://www.mailgun.com/blog/email/ai-phishing/

[12] Hoxhunt, "Phishing Trends Report (Updated for 2025)." [Online]. Available: https://hoxhunt.com/guide/phishing-trends-report

[13] B. B. Gupta, A. Tewari, A. K. Jain, and D. Agrawal, "Fighting against phishing attacks: State of the art and future challenges," Neural Comput. Appl., vol. 28, no. 12, pp. 3629–3654, 2017. [Online]. Available: https://doi.org/10.1007/s00521-016-2275-y

[14] G. Rabitti et al., "A taxonomy of cyber risk taxonomies," Risk Anal., vol. 45, no. 2, pp. 376–386, 2025. [Online]. Available: https://doi.org/10.1111/risa.16629

[15] T. T. Nguyen et al., "Deep learning for deepfakes creation and detection: A survey," arXiv preprint, arXiv:1909.11573, Sep. 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1909.11573

[16] S. Albladi and G. R. S. Weir, "User characteristics that influence judgment of social engineering attacks in social networks," Hum. -Centric Comput. Inf. Sci., vol. 8, no. 1, art. 28, 2018. [Online]. Available: https://doi.org/10.1186/s13673-018-0128-7

[17] G. Loukas, D. Gan, and T. Vuong, "A taxonomy of cyber-attacks and defence mechanisms for emergency management networks," in Proc. IEEE PerCom Workshops, San Diego, CA, USA, Mar. 2013, pp. 534–539. doi: 10.1109/PerComW.2013.6529554

[18] M. Rizwan, M. S. Sarfraz, and K. Främling, "Cyber security taxonomies: A comparative study including AI-enabled attacks and defenses," arXiv preprint, arXiv:2401.01374, 2024. [Online]. Available: https://arxiv.org/abs/2401.01374

[19] R. R. Nuiaa and S. Manickam, "A critical review: A new taxonomy for phishing attacks based on phishing techniques used," Wasit J. Pure Sci., vol. 2, no. 2, pp. 251–269, 2023. [Online]. Available: https://doi.org/10.31185/wjps.143

[20] H. Aldawood and G. Skinner, "An advanced taxonomy for social engineering attacks," Int. J. Comput. Appl., vol. 177, no. 30, pp. 1–11, 2020. [Online]. Available: https://doi.org/10.5120/ijca2020919744

[21] J. Rastenis et al., "E-mail-based phishing attack taxonomy," Appl. Sci., vol. 10, no. 7, art. no. 2363, Mar. 2020. [Online]. Available: https://doi.org/10.3390/app10072363

[22] K. Ivaturi and L. Janczewski, "A taxonomy for social engineering attacks," in Proc. KMIS & CONF-IRM Int. Conf., Seoul, South Korea, Jun. 2011, pp. 324–334.

[23] P. Tulkarm, "A survey of social engineering attacks: Detection and prevention tools," J. Theor. Appl. Inf. Technol., vol. 99, no. 18, pp. 1–12, 2021.

[24] N. A. Odeh, D. Eleyan, and A. Eleyan, "A survey of social engineering attacks: Detection and prevention tools," J. Theor. Appl. Inf. Technol., vol. 99, no. 18, p. 4375, Sep. 30, 2021.

[25] S. A. Ogundairo and A. Broklyn, "An intelligent model for AI-powered phishing attack detection using natural language techniques," J. Cybersecur. Intell. Cybercrime, vol. 7, no. 1, pp. 54–71, 2024.

[26] S. Asiri et al., "PhishingRTDS: A real-time detection system for phishing attacks using a Deep Learning model," Comput. Secur., vol. 141, p. 103843, Jun. 2024. [Online]. Available: https://doi.org/10.1016/j.cose.2024.103843

[27] J. Chung, J.-Z. Koay, and Y. B. Leau, "A review on social media phishing: Factors and countermeasures," in Adv. Cyber Secur., M. Anbar et al., Eds., Commun. Comput. Inf. Sci., vol. 1347, pp. 657–673. Singapore: Springer, 2021. [Online]. Available: https://doi.org/10.1007/978-981-33-6835-4_43

[28] European Cybersecurity Atlas, "Cyber security taxonomy." [Online]. Available: https://cybersecurity-atlas.ec.europa.eu/cybersecurity-taxonomy

[29] A. M. Bhat, "Taxonomy of Generative Model," Feb. 22, 2024. [Online]. Available: https://medium.com/@avinashmbhat182/taxonomy-of-generative-model-4c9be2a09748

[30] R. Alabdan, "Phishing attacks survey: Types, vectors, and technical approaches," Future Internet, vol. 12, no. 10, art. no. 168, 2020. [Online]. Available: https://doi.org/10.3390/fi12100168

[31] The SEO Platform, "Are social media users exploited? 6 ways we can change all that." [Online]. Available: https://theseoplatform.co.uk/blog/are-social-media-users-exploited-6-ways-we-can-change-all-that/

[32] CanIPhish, "The 6 most popular AI scams in 2025," Jan. 10, 2025. [Online]. Available: https://caniphish.com/blog/ai-scams

[33] DFPI, "Pig butchering – how to spot and report the scam." [Online]. Available: https://dfpi.ca.gov/news/insights/pig-butchering-how-to-spot-and-report-the-scam/

[34] CNN, "Finance worker pays out $25 million after video call with deepfake's chief financial officer." [Online]. Available: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

[35] TotalDefense, "AI emails are getting scary good: Here's how to spot the fakes!" 2025. [Online]. Available: https://www.totaldefense.com/security-blog/ai-emails-are-getting-scary-good-heres-how-to-spot-the-fakes/

[36] The Nation, "10 subtle ways scammers use AI to deceive victims." [Online]. Available: https://thenationonlineng.net/10-subtle-ways-scammers-use-ai-to-deceive-victims/

[37] AARP, "How AI is making ‘pump and dump’ investment scams easy for criminals." [Online]. Available: https://www.aarp.org/podcasts/the-perfect-scam/info-2025/pump-and-dump-scam.html

[38] D. I. Mienye and N. R. Jere, "A survey of decision trees: Concepts, algorithms, and applications," IEEE Access, vol. PP, no. 99, pp. 1–1, Jan. 2024, doi: 10.1109/ACCESS.2024.3416838.

[39] S. Ding, B.-J. Qi, and H.-Y. Tan, "An overview on theory and algorithm of support vector machines," J. Univ. Electron. Sci. Technol. China, vol. 40, no. 1, pp. 2–10, Jan. 2011, doi: 10.3969/j.issn.1001-0548.2011.01.001.

[40] GeeksforGeeks, "K-Nearest Neighbours (KNN) Algorithm." [Online]. Available: https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/. [Accessed: Jul. 22, 2025].

[41] D. ul Haq, M. H. Faheem, and I. Ahmad, "Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks," Appl. Sci., vol. 14, no. 22, art. no. 10086, Nov. 2024. doi: 10.3390/app142210086.

[42] R. Jonker et al., "Using natural language processing for phishing detection," in Optimization, Learning Algorithms and Applications, Commun. Comput. Inf. Sci., vol. 1234, pp. 540–552, Jan. 2021. doi: 10.1007/978-3-030-91885-9_40.

[43] D. L. Yse, "Your guide to natural language processing (NLP)," Towards Data Science. [Online]. Available: https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e19. [Accessed: Dec. 20, 2020].

[44] D. Sánchez-García, T. S. F. Gilabert, and J. A. Calvo-Manzano, "Countermeasures and their taxonomies for risk treatment in cybersecurity: A systematic mapping review," Comput. Secur., vol. 128, art. no. 103170, May 2023. doi: 10.1016/j.cose.2023.103170.

[45] M. Manjula, "Cyber security threats and countermeasures using machine and deep learning approaches: A survey," J. Comput. Sci., vol. 19, no. 1, pp. 20–56, Jan. 2023. doi: 10.3844/jcssp.2023.20.56.

[46] E. Burns, “Combating WormGPT: What You Need to Know,” Abnormal Security, Jul. 2025. [Online]. Available: https://abnormal.ai/blog/combating-wormgpt