

Checklist-based phishing detection leveraging Large Language Models

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

Author:
Xeno Macatangay

Supervisors:
Jouni Isoaho
Tahir Mohammad

May 2025

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

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: Xeno Macatangay
Title: Checklist-based phishing detection leveraging Large Language Models
Number of pages: 171 pages, 63 appendix pages
Date: May 2025

In the third quarter of 2024, phishing attacks surged to over 930,000, with AI and machine learning technologies increasingly enabling attackers to automate and amplify these threats. Phishing emails remain a persistent threat because current detection methods often overlook user perception traits and factors, highlighting the need for a practical, user-focused guideline or checklist to improve identification and prevention. This thesis focuses on detecting phishing emails in the age of artificial intelligence (AI) and machine learning (ML). This thesis aims to contribute to existing research regarding phishing detection and mitigation. A literature review is the primary method used for this research providing an overview of common phishing email characteristics, which were then compiled to produce this thesis's proposed CARLS checklist. The "CARLS" acronym stems from the initial characters of the perception characteristics: **C**ommitment & consistency, **A**uthority, **R**eciprocity, **L**iking & similarity, **S**ocial proof, and **S**carcity. The checklist contains both phishing email characteristics and email perceptions of users to help detect plausible red flags in emails. The effectiveness of the checklist has been evaluated both manually and in an automated manner, with GPT4All (Nous) model trained on the CARLS checklist and tested against an imbalanced dataset achieving an accuracy of 90%. To conclude, despite the rise of AI and ML, phishing email characteristics have seen no significant changes, as shown in the characteristics within the CARLS checklist. The CARLS checklist is primarily human operated so that it can be consulted or included in various security awareness training programs. Furthermore, the CARLS checklist can also be automated by incorporating the checklist into an application which performs checks based on the checklist items or by teaching a large language model (LLM) or other AI/ML models to employ the checklist as part of a decision support system.

Keywords: Cybersecurity, Large Language Models, LLM, Artificial Intelligence, AI, Phishing, Checklist, Guideline, AI-enabled

Table of contents

- 1 Introduction 10**
 - 1.1 Problem statement10**
 - 1.2 Research question10**
 - 1.3 Research objectives.....11**
 - 1.4 Thesis structure11**
- 2 Background 12**
 - 2.1 Social Engineering (SE).....12**
 - 2.1.1 Social engineering attack lifecycle..... 13
 - 2.2 Types of social engineering attacks14**
 - 2.2.1 Dumpster diving..... 14
 - 2.2.2 Reverse social engineering 15
 - 2.2.3 Baiting 15
 - 2.2.4 Shoulder surfing..... 16
 - 2.2.5 Watering hole..... 16
 - 2.2.6 Impersonation 16
 - 2.2.7 Tailgating 16
 - 2.2.8 Eavesdropping..... 17
 - 2.2.9 Pretexting..... 17
 - 2.2.10 Quid Pro Quo..... 17
 - 2.2.11 Typosquatting 17
 - 2.3 Phishing.....18**
 - 2.3.1 Phishing 18
 - 2.3.2 Spear phishing..... 19
 - 2.3.3 Whaling..... 19
 - 2.3.4 Vishing 20
 - 2.3.5 SMiShing 20
 - 2.4 Large Language Models (LLMs).....21**
 - 2.4.1 How do LLMs work? 22
 - 2.4.2 LLM challenges..... 23
 - 2.5 Real-world AI-enabled social engineering attacks.....23**
- 3 Literature review 26**
 - 3.1 Persuasion principles of social engineering.....26**
 - 3.1.1 Authority..... 26

3.1.2	Social proof.....	27
3.1.3	Liking and similarity	27
3.1.4	Commitment and consistency	27
3.1.5	Reciprocity	28
3.1.6	Scarcity	28
3.1.7	Distraction.....	28
3.2	Individual differences	29
3.2.1	Age	29
3.2.2	Lower employment years/satisfaction	30
3.2.3	Sender authenticity	30
3.2.4	Perception of links in emails	30
3.2.5	Email style familiarity	31
3.2.6	Emotional attachments to email	31
3.2.7	Individual habits	31
3.3	Effectiveness and combination of techniques.....	31
3.4	Detecting (spear) phishing attacks	33
3.4.1	Incorrect grammar and or spelling.....	34
3.4.2	Unusual or mismatched sender.....	34
3.4.3	Deceptive or suspicious links/attachments.....	34
3.4.4	Generic or unfamiliar greetings	35
3.4.5	Unusual of unexpected requests	35
3.4.6	Urging immediate action	36
3.4.7	Too good to be true	36
3.5	Phishing email mitigations	37
3.5.1	Awareness training	37
3.5.2	Phishing simulation.....	38
3.5.3	Multifactor authentication.....	38
3.5.4	Social reporting/crowdsourcing	39
3.5.5	Automated tools.....	39
3.5.6	Secure routines.....	40
3.6	AI-generated text detection	40
3.6.1	Watermark-based detection	40
3.6.2	Feature-based detection.....	41
3.6.3	Detection via LLM.....	42
3.6.4	Human-machine combination.....	42
3.6.5	Text detection challenges	43
3.7	AI-enhanced phishing.....	43
3.7.1	Filter bypass	43

3.7.2	Human-like text.....	44
3.7.3	Semi-automation.....	44
3.7.4	Personalization	46
3.7.5	Cost-effectiveness	46
4	Research methodology	48
4.1	Background and literature.....	48
4.2	Creation of CARLS checklist.....	49
4.3	Implementation with LLM	49
5	Phishing detection checklist.....	50
5.1	CARLS checklist	50
5.1.1	Machine-generated text detection	52
5.2	Checklist testing methodology	54
5.2.1	Prompt queries	54
5.2.2	Perplexity AI.....	55
5.2.3	GPT4All	56
5.3	CARLS checklist manual evaluation.....	57
5.3.1	Checklist versus Perplexity AI	57
5.3.2	Checklist versus GPT4All (Llama).....	61
5.3.3	Checklist versus GPT4All (Nous)	65
5.3.4	CARLS checklist evaluation results analysis.....	69
6	Implementation with LLM.....	70
6.1	Implementation experiment.....	70
6.1.1	Implementation experiment test methodology.....	70
6.1.2	LLM prompt engineering.....	71
6.1.3	Perplexity AI.....	73
6.1.4	GPT4All (Llama)	75
6.1.5	GPT4All (Nous).....	78
6.1.6	Implementation experiment results analysis	82
6.1.7	Revised CARLS checklist.....	88
6.2	Implementation comparison.....	89
6.2.1	Data collection	89
6.2.2	Implementation comparison test methodology.....	90
6.2.3	Implementation comparison results	90
6.2.4	Implementation comparison results analysis and comparison.....	91
6.3	Personal email data test	92

6.3.1	Data collection	92
6.3.2	Personal email data test results	92
6.3.3	Personal email data results analysis	93
7	Discussion, limitation, and recommendations	94
7.1	Research limitations	95
7.2	Recommendations and future research	95
8	Conclusion	96
	References	97
	Appendices	108
	Appendix 1 Perplexity AI generated emails	108
	Appendix 2 GPT4All (Llama) generated emails	114
	Appendix 3 GPT4All (Nous) generated emails.....	118
	Appendix 4 Perplexity AI full responses	120
	Appendix 5 Imbalanced dataset test responses.....	124
	Appendix 6 Balanced dataset test responses	136
	Appendix 7 Implementation comparison email samples	144
	Appendix 8 Personal emails experiment dataset	149
	Appendix 9 Personal emails experiment responses	156

List of Figures

- Figure 1. General attack lifecycle of a social engineering attack 13
- Figure 2. Typical LLM lifecycle 22
- Figure 4 Perplexity AI on GPT4All (Llama) query 1 73
- Figure 5 Perplexity AI on GPT4All (Llama) query 2 73
- Figure 6 Perplexity AI on GPT4All (Nous) query 1 74
- Figure 7 Perplexity AI on GPT4All (Nous) query 2 74
- Figure 8 GPT4All (Llama) on Perplexity AI query 1 75
- Figure 9 GPT4All (Llama) on Perplexity AI query 2 76
- Figure 10 GPT4All (Llama) on GPT4All (Nous) query 1 77
- Figure 11 GPT4All (Llama) on GPT4All (Nous) query 2 78
- Figure 12 GPT4All (Nous) on GPT4All (Llama) query 1 78
- Figure 13 GPT4All (Nous) on GPT4All (Llama) query 2 79
- Figure 14 GPT4All (Nous) on Perplexity AI query 1 80
- Figure 15 GPT4All (Nous) on Perplexity AI query 2 81
- Figure 16 Perplexity AI CARLS checklist feedback on GPT4All (Nous) query 2 83
- Figure 17 GPT4All (Llama) CARLS checklist feedback on Perplexity AI query 1 84
- Figure 18 GPT4All (Llama) CARLS checklist feedback on Perplexity AI query 2 85
- Figure 19 GPT4All (Llama) CARLS checklist feedback on GPT4All (Nous) query 2 85
- Figure 20 GPT4All (Nous) CARLS checklist feedback on GPT4All (Llama) query 1 86
- Figure 21 GPT4All (Nous) CARLS checklist feedback on Perplexity AI query 2 87

List of Tables

- Table 1. Overview of social engineering attack types 14
- Table 2. Overview of commonly used phishing attacks 18
- Table 3. Email perception checklist..... 50
- Table 4. Email characteristics checklist..... 51
- Table 5. Checklist points system and phishing probability 52
- Table 6. CARLS checklist sources 53
- Table 7. System specifications during checklist testing 54
- Table 8. Prompt engineered queries 54
- Table 9. GPT4All model parameters 56
- Table 10. Figure 1.1 checklist..... 57
- Table 11. Figure 1.2 checklist..... 58
- Table 12. Figure 1.3 checklist..... 59
- Table 13. Figure 1.5 checklist..... 60
- Table 14. Figure 2.1 checklist..... 61
- Table 15. Figure 2.2 checklist..... 62
- Table 16. Figure 2.3 checklist..... 63
- Table 17. Figure 2.5 checklist..... 64
- Table 18. Figure 3.1 checklist..... 65
- Table 19. Figure 3.2 checklist..... 66
- Table 20. Figure 3.3 checklist..... 67
- Table 21. Figure 3.4 checklist..... 68
- Table 22. Manual checklist evaluation results 69
- Table 23. Query-to-LLM assignment overview..... 70
- Table 24. LLM phishing email evaluation utilising the CARLS checklist..... 82
- Table 25. Revised CARLS email characteristics checklist..... 88
- Table 26. Imbalanced dataset classification test results 90
- Table 27. Balanced dataset classification test results..... 91
- Table 28. GPT4All and IPSDM Test accuracy comparison 91
- Table 29. Personal email phishing data test results..... 92
- Table 30. Personal email phishing data test accuracy 93

List of Abbreviations

Abbreviation	Definition
ACM	Association for Computing Machinery
AI	Artificial Intelligence
APWG	Anti-Phishing Working Group
BEC	Business Email Compromise
BERT	Bidirectional Encoder Representations from Transformers
CEO	Chief Executive Officer
CFO	Chief Financial Officer
DMARC	Domain-based Message Authentication, Reporting, and Conformance
DNN	Deep Neural Network
FOMO	Fear Of Missing Out
GLTR	Giant Language Model Test Room
GPT	Generative Pre-training Transformer
HCD	Human-Centred Design
HIDS	Host-Based Intrusion Detection System
HR	Human Resources
IDS	Intrusion Detection System
IEEE	The Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IPS	Intrusion Prevention System
IPSDM	Improved Phishing Spam Detection Model
IT	Information Technology
LLM	Large Language Models
MFA	Multi-factor Authentication
ML	Machine Learning
NLP	Natural Language Processing
OTP	One-Time Password
PDF	Portable Document Format
PII	Personal Identifiable Information
PIN	Personal Identification Number
QR	Quick Response
SAT	Security Awareness Training
SE	Social Engineering
SMS	Short Message Service
URL	Uniform Resource Locator
USB	Universal Serial Bus
USPS	United States Postal Service

1 Introduction

In the third quarter of 2024, 932,923 phishing attacks were observed by the Anti-Phishing Working Group (2024). This number is up from 877,536 observed phishing attacks within the second quarter of the same year. Furthermore, vishing (voice phishing) and smishing (SMS phishing) also increased by 28% and 22%, respectively, in the third quarter of 2024, according to a long-time APWG (Anti-Phishing Working Group) report contributor and senior product manager at OpSec, Matthew Harris. As per the statistics, phishing is not going to disappear any time soon. On the contrary, phishing seems consistent between 290,000 and 370,000 attacks monthly (Anti-Phishing Working Group, 2024). The rising trend in the development and usage of AI (Artificial Intelligence) and ML (Machine Learning), specifically LLMs (Large Language Models), that allow users to automate digital tasks presents a double-edged sword for everyone. While the premise of AI is to support users with their day-to-day lives, this would also mean that malicious users are granted the same privileges and assistance with their adversarial endeavours, such as phishing.

1.1 Problem statement

Despite various advances in AI and ML, phishing emails remain a significant threat due to their persistent and evolving deceptive characteristics. Existing detection methods, while growing more complex, often overlook the stable set of phishing email traits that deceive users. Consequently, there is a lack of user-perception-focused frameworks that recognise phishing attempts. The CARLS checklist, a human-operable tool which incorporates technical phishing indicators and user perception elements to improve phishing detection and awareness training, is proposed in this thesis to fill the gap by examining whether the characteristics of phishing emails have evolved fundamentally in the AI/ML era. It also investigates the possibility of automating this checklist leveraging AI/ML models to assist decision-making in phishing prevention efforts. Therefore, the key challenge is the ability to efficiently identify and counteract phishing emails in light of AI developments by fusing user-focused insights with classic phishing characteristics in a practical, flexible framework.

1.2 Research question

The research question of this thesis is: “How to detect phishing emails in the age of Large Language Models (LLM)?” This thesis focuses on text-based phishing, specifically emails. To

further narrow down the scope, the emails referred to within this thesis pertain to business emails, such as exchanges between companies and organizations.

1.3 Research objectives

The research objectives include:

- To identify various email characteristics commonly found in sophisticated phishing emails
- To propose a checklist/guideline comprising phishing characteristics to enhance phishing detection using LLMs
- To evaluate the effectiveness of the proposed checklist/guideline both manually and in an automated manner utilizing LLMs

The proposed checklist/guideline aims to build on existing phishing detection and mitigation research. Literature review is the primary method of research for this thesis. It is achieved by utilizing existing research databases such as Google Scholar, IEEE, and ACM digital library.

1.4 Thesis structure

The structure of this thesis is as follows: First, the background will provide relevant information regarding the topics within this thesis. Second, the literature review will examine various techniques for effective phishing and the theoretical and practical combination of AI and social engineering. Third, the research methodology will provide more information regarding this thesis's research methods along with the rationale. Fourth, the detection/guideline checklist will present the proposed checklist/guideline of this thesis and evaluate the proposed checklist. Fifth, the implementation with LLM will evaluate the checklist's effectiveness with the help of AI in an automated manner and a comparison of the local LLMs against another model. Then, the discussion will present this thesis's discussion along with the thesis's various limitations and future research recommendations. Lastly, the conclusion will conclude the thesis by answering the research question. Grammarly¹ and QuillBot² were used as a grammar checker and paraphraser, respectively, during the creation of this thesis.

¹ <https://www.grammarly.com/>

² <https://quillbot.com/>

2 Background

This chapter will provide background information regarding topics relevant to this thesis. The structure is as follows: First, social engineering will touch upon the ideology of social engineering alongside its attack lifecycle and various social engineering attacks. Second, multiple types of social engineering attacks will be discussed. Third, the phishing section will dive deep into numerous kinds of phishing. Fourth, the large language models section will provide a high-level overview of what large language models are and how they work. Lastly, real-world AI-enabled social engineering attacks will present some examples of social engineering attacks that have taken place in the real world, which were either aided or made possible with artificial intelligence.

2.1 Social Engineering (SE)

Social engineering (SE), in the context of computer science, is the act of user manipulation, deception, or exploitation to compromise information systems and retrieve confidential information or data (Ghafir et al., 2016; Hatfield, 2018; Heartfield & Loukas, 2015; Krombholz et al., 2015). Confidential data or information can be usernames, passwords, addresses, bank details, employee information, and more. (Chetioui et al., 2022). Human vulnerabilities and naivety are taken advantage of to infiltrate or bypass security systems using the likes of persuasion and influence, to name a few (Ghafir et al., 2016; Krombholz et al., 2015; Mashtalyar et al., 2021; Parthy & Rajendran, 2019). Other exploitable human vulnerabilities include (but are not limited to) emotions, habits, trust, eagerness, curiosity, and courtesy. (Chetioui et al., 2022; Kamruzzaman et al., 2023). SE can also be considered a skill employed by adversaries to gain trust and access from an individual within an organization, ultimately leading to an organizational security breach (Ghafir et al., 2016). SE attacks do not require the same amount of technical knowledge as other attacks, as numerous attacks abuse the curiosity or trust of the victim (Kamruzzaman et al., 2023). As mentioned before, the susceptibility of humans predominantly contributes to SE attacks' success as humans are more inclined to click links or grant information to trustworthy people, people of authority, and or requests with urgency (Hatfield, 2018; Mashtalyar et al., 2021).

Almost all SE attacks make use of one or more similar principles. These principles are authority, urgency and scarcity, intimidation, familiarity, liking and similarity, commitment and consistency, reciprocation, and natural inclination to help (Ghafir et al., 2016; Kamruzzaman

et al., 2023). Authority is when adversaries assert themselves to be of authority to secure access to information. Urgency and scarcity are when humans are tempted to do something at the last minute to win a prize or take immediate action. Intimidation refers to the utilization of fear of consequences to exploit humans into executing their commands. Familiarity occurs when adversaries convince humans they are who they claim to be. Liking and similarity abuse the natural tendency of humans to affiliate with humans with common interests or origins. Commitment and consistency involve capitalizing on employees longing to be recognized as trustworthy and committed to their tasks. Reciprocation preys on the social interaction norm that if someone provides us with something, it is only fitting to return the favour. As the name suggests, the natural inclination to help is the natural tendency of humans to aid those in need. The rest of the section will explore an overview of the social engineering attack lifecycle, present several types of social engineering attacks, and discuss various phishing types.

2.1.1 Social engineering attack lifecycle

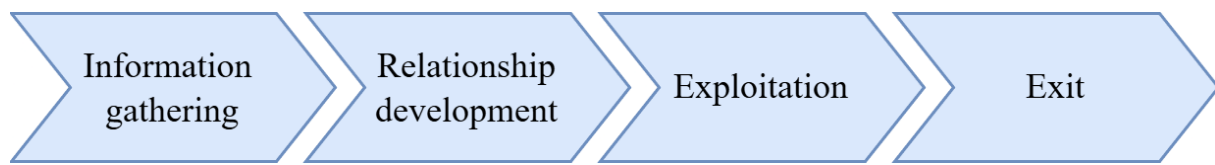


Figure 1. General attack lifecycle of a social engineering attack

SE attacks usually revolve around four phases: information gathering, relationship development, exploitation, and exit as seen from Figure 1 (Chetioui et al., 2022; Ghafir et al., 2016; Kamruzzaman et al., 2023; Mashtalyar et al., 2021; Parthy & Rajendran, 2019; Salahdine & Kaabouch, 2019). Information gathering, also called research or investigation, is where the adversary selects a target based on some criteria and performs background research (Chetioui et al., 2022; Salahdine & Kaabouch, 2019). Information can be harvested from various public sources, such as social media, web pages, search engines, and job portals (Ghafir et al., 2016; Krombholz et al., 2015). Relationship development, also known as hook, is when an adversary tries to close the gap between them and the target by abusing natural human tendencies such as helpfulness, curiosity, and trust. (Chetioui et al., 2022; Ghafir et al., 2016). Trust can be gained through direct physical contact, email communication, online chat, phone, and more (Salahdine & Kaabouch, 2019). Exploitation, also known as play, is when the adversary manipulates the target to make security mistakes and acquire wanted or necessary information and data. (Chetioui et al., 2022; Salahdine & Kaabouch, 2019). Such information can be passwords, login

details, bank details, confidential data, and more (Ghafir et al., 2016). Exit is the last step in which the adversary leaves the target and ends communication without leaving any evidence (Chetioui et al., 2022; Salahdine & Kaabouch, 2019).

2.2 Types of social engineering attacks

Social engineering consists of multiple types of attacks. These attacks are, but not limited to, baiting, dumpster diving, eavesdropping, impersonation, pretexting, Quid Pro Quo, reverse engineering, shoulder surfing, tailgating, typosquatting, and watering hole, as can be seen in Table 1. These attack types will be discussed further.

Table 1. Overview of social engineering attack types

	Chetioui et al., 2022	Hatfield, 2018	Kamruzzaman et al., 2023	Krombholz et al., 2015	Mashtalyar et al., 2021	Parthy & Rajendran, 2019	Salahdine & Kaabouch, 2019	Spaulding et al., 2017
Baiting	x		x	x		x	x	
Dumpster diving		x		x	x	x	x	
Eavesdropping						x		
Impersonation		x			x	x	x	
Pretexting	x		x	x		x	x	
Quid Pro Quo	x		x				x	
Reverse engineering		x		x		x	x	
Shoulder surfing		x		x	x	x	x	
Tailgating	x		x	x	x	x	x	
Typosquatting					x			x
Watering hole				x	x	x		

2.2.1 Dumpster diving

Dumpster diving, as the name implies, is the act of diving and sifting through the rubbish and garbage of organizations or individuals (Krombholz et al., 2015). Adversaries seek thrown-out documents, discarded equipment, storage drives, and more which may contain sensitive data or

information (Parthy & Rajendran, 2019; Salahdine & Kaabouch, 2019). Adversaries look for personal information, employee data, IP addresses, financial information, manuals, and many more that could be useful (Mashtalyar et al., 2021).

2.2.2 Reverse social engineering

As the name suggests, reverse social engineering is the reverse of social engineering where the target initiates contact with the adversary (Hatfield, 2018). The adversary achieves this by creating an issue or a problem for the target. The adversary then proceeds to advertise aid to entice the target into initiating contact themselves. Finally, the adversary responds and claims to solve the issue (Parthy & Rajendran, 2019; Salahdine & Kaabouch, 2019). This type of attack usually establishes trust between the target and adversary (Krombholz et al., 2015). With the trust gained by the adversary, they can continue to pry into more sensitive personal or work-related information (Parthy & Rajendran, 2019).

2.2.3 Baiting

Baiting, also referred to as road apples, refers to attacks that entice targets by exploiting human curiosity or reward exchange (Chetioui et al., 2022; Salahdine & Kaabouch, 2019). Baits can be physical or non-physical items such as USB storage drives, CDs, and other objects which can be collected and used by people on their computers (Chetioui et al., 2022; Kamruzzaman et al., 2023; Salahdine & Kaabouch, 2019). Such items are usually left or dropped in areas such as coffee shops, where they can easily be found by targets (Krombholz et al., 2015; Salahdine & Kaabouch, 2019). These items are often infected with malicious software (malware) to extract the target's personal or sensitive information and data (Chetioui et al., 2022; Kamruzzaman et al., 2023; Krombholz et al., 2015; Salahdine & Kaabouch, 2019). Baiting can sometimes also offer a reward or goods, such as music, in exchange for the target's information (Chetioui et al., 2022). A real-life example of a baiting attack is the Stuxnet malware. Stuxnet debuted in 2010, which targeted nuclear facilities, specifically Iran. The targeted nuclear facilities were isolated from the internet thus, it is presumed that the malware was transmitted using USB storage devices (Ilevičius, 2022). This makes the attack a baiting type of attack due to the utilization of USB devices.

2.2.4 Shoulder surfing

Shoulder surfing refers to the use of observation to acquire sensitive information from the target (Mashtalyar et al., 2021). Adversaries can employ direct or indirect observation techniques. Direct observation techniques involve looking over the target's shoulder while sensitive information or data is entered by the target (Krombholz et al., 2015; Parthy & Rajendran, 2019; Salahdine & Kaabouch, 2019). Indirect observation techniques include the use of modern equipment such as binoculars or cameras (Parthy & Rajendran, 2019). Sensitive information that can be entered by the target and viewed by the attacker includes passwords, identification, and other personal or confidential information (Mashtalyar et al., 2021; Salahdine & Kaabouch, 2019).

2.2.5 Watering hole

Watering hole, also referred to as waterholing, is an attack where adversaries compromise websites that are of interest or regularly visited by the target (Krombholz et al., 2015; Mashtalyar et al., 2021). Adversaries then search for vulnerabilities and compromise these websites by infecting them with malware or viruses (Mashtalyar et al., 2021; Parthy & Rajendran, 2019). The most prominent attack vectors are malicious office documents and archives (Krombholz et al., 2015). A real-life example is the widespread attack which targeted Apple devices in Hong Kong in 2021 (Newman, 2021). Adversaries compromised a variety of websites that, when visited by Apple devices, exploit vulnerabilities within their devices, allowing the installation of malware (Mott, 2021).

2.2.6 Impersonation

Impersonation is an attack where an adversary poses as someone else (Parthy & Rajendran, 2019). Adversaries pretend to be legitimate users or employees to try and achieve access to a system (Mashtalyar et al., 2021). Impersonation may also be utilized to gather information such as usernames and passwords to attain access to the target network (Hatfield, 2018).

2.2.7 Tailgating

Tailgating, also called piggybacking, is a technique used by adversaries to gain access to authorized or restricted areas (Mashtalyar et al., 2021; Salahdine & Kaabouch, 2019). This type of attack can be achieved by trailing someone (such as an employee) who has access to the specific area (Mashtalyar et al., 2021; Parthy & Rajendran, 2019; Salahdine & Kaabouch,

2019). Abuse of human basic courtesy (such as helpfulness), trickery, and blending in can also be used to execute such an attack (Chetioui et al., 2022; Kamruzzaman et al., 2023; Mashtalyar et al., 2021).

2.2.8 Eavesdropping

Eavesdropping is a type of attack used by adversaries to listen in on private conversations. Eavesdropping can be accomplished by listening to employees having private conversations within company premises or public areas such as restaurants or bars (Parthy & Rajendran, 2019).

2.2.9 Pretexting

Pretexting is an attack whereby adversaries fabricate convincing scenarios to deceive the target into trusting the adversary (Salahdine & Kaabouch, 2019). This attack “capitalizes on human’s desire to trust” (Chetioui et al., 2022). After gaining the target’s trust, adversaries can then proceed to request and gather personal details and information, which can later be misused (Parthy & Rajendran, 2019; Salahdine & Kaabouch, 2019). Additionally, adversaries may also pose as an authoritative figure (Kamruzzaman et al., 2023) or claim to call from a legitimate source (Parthy & Rajendran, 2019) to further increase their trustworthiness.

2.2.10 Quid Pro Quo

Quid Pro Quo is a type of baiting attack where an adversary offers a service, sometimes a product, in exchange for something (Chetioui et al., 2022; Salahdine & Kaabouch, 2019). This exchange usually involves sensitive information or access to the target’s system or network (Chetioui et al., 2022; Kamruzzaman et al., 2023).

2.2.11 Typosquatting

Typosquatting, also known as brand theft or cybersquatting, is where adversaries trick their target(s) into thinking that the websites and or services they interact with are genuine (Mashtalyar et al., 2021). Typosquatters can achieve such trickery by registering an equally alike domain their target regularly interacts with and adding or removing a single character within the domain name. Popular domains are usually targeted, such as “microsoft.com,” which can be typosquatted with “microsooft.com” (notice the double “o”) or “microsft.com” (notice the missing “o”) (Spaulding et al., 2017). A real-life example would be the spoofed domains

posing as the US Census Bureau in 2020 (Federal Bureau of Investigation, 2020). Adversaries registered domains resembling the US Census Bureau domains (“Censusbureau[.]com” or “census-gov[.]us” as an example) with the malicious intent of acquiring sensitive information from innocent users and or installation of malware (Bîzgå, 2020).

2.3 Phishing

Phishing comes in many forms; thus, for this thesis and within this section, only the most common forms will be discussed, including phishing, spear phishing, whaling, vishing, and smishing. A phishing overview is provided in Table 2.

Table 2. Overview of commonly used phishing attacks

	Allodi et al., 2019	Butavicius et al., 2016	Chetioui et al., 2022	Desolda et al., 2022	Han & Shen, 2016	Hanus et al., 2021	Hayes, 2020	Koddebusch, 2022	Krombholz et al., 2015	Mashtalyar et al., 2021	Parthy & Rajendran, 2019	Salahdine & Kaabouch, 2019	Wassermann et al., 2023	Yeboah-Boateng & Amanor, 2014
Spear phishing	x	x		x	x	x		x	x			x	x	
Whaling							x		x			x		
Vishing			x	x						x	x	x	x	x
SmiShing				x							x	x	x	x

2.3.1 Phishing

Phishing is a type of attack where adversaries may send spoofed emails, make phone calls, send SMS, or send an instant message to the victim (Chetioui et al., 2022; Hanus et al., 2021; Krombholz et al., 2015; Taib et al., 2019). Such an attack aims to retrieve sensitive information from the target, such as passwords, usernames, and other personal details (Butavicius et al., 2016; Chetioui et al., 2022; Desolda et al., 2022; Krombholz et al., 2015; Salahdine & Kaabouch, 2019; Taib et al., 2019). Attackers may also fabricate fake websites, false advertisements, bogus software (such as anti-virus), and free offerings to trick or entice victims (Parthy & Rajendran, 2019; Salahdine & Kaabouch, 2019). Another goal of phishing is to make

the target do the attacker's bidding, such as clicking certain links or downloading attachments by pretending to be a legitimate entity (Hatfield, 2018; Krombholz et al., 2015). Phishing is usually intended for large groups of people, akin to spam emails (Allodi et al., 2019; Chetioui et al., 2022; Krombholz et al., 2015).

2.3.2 Spear phishing

Spear phishing is a type of specific phishing where it targets specific groups, companies, or individuals (Butavicius et al., 2016; Desolda et al., 2022; Hanus et al., 2021; Salahdine & Kaabouch, 2019; Wassermann et al., 2023). The messages and emails are carefully and precisely created after preliminary data gathering, also referred to as data mining. These can be found via publicly available sources such as a company or organization's website and social networking sites. The increased accuracy and validity of these crafted messages and emails using public information results in higher success rates compared to traditional phishing (Allodi et al., 2019; Butavicius et al., 2016; Han & Shen, 2016; Koddebusch, 2022; Krombholz et al., 2015). A real-life example of a spear phishing attack would be the attack on Ubiquiti (Vardi, 2016). According to reports, Rohit Chakravarthy (At the time, Ubiquiti's principal financial and accounting officer) received bogus emails claiming to be from Robert Pera (founder and CEO of Ubiquiti) and Tom Evans, a lawyer of Latham & Watkins London branch. The highly targeted and convincing email mentioned a confidential acquisition that required several wire transactions. Chakravarthy reportedly was the only employee to receive the bogus emails. Chakravarthy received multiple emails from the alleged lawyer containing banking details and instructions. This spear phishing attack ultimately led to Ubiquiti's loss of \$46.7 million.

2.3.3 Whaling

Whaling, sometimes referred to as CEO fraud or Business Email Compromise (BEC), is a type of spear phishing attack which targets high-profile victims within companies, enterprises, and or organizations due to their near limitless access rights (Hayes, 2020; Krombholz et al., 2015; Salahdine & Kaabouch, 2019). The resources of interest which may be accessible to these high-profile targets include financial and various other resources within the company or organization (Hayes, 2020). The real-life example given for spear phishing can also be classified as whaling since adversaries impersonated the CEO and persuaded a high-level officer to fulfil their requests.

2.3.4 Vishing

Vishing comes from the combination of “voice” and “phishing” (Yeboah-Boateng & Amanor, 2014), is a form of phishing attack which is done over the phone (Chetioui et al., 2022; Salahdine & Kaabouch, 2019; Wassermann et al., 2023). Adversaries make use of phone calls to lure targets into spilling sensitive information (Desolda et al., 2022; Parthy & Rajendran, 2019). Adversaries impersonate employees from both governmental and private organizations to increase their credibility and chances of success. Adversaries tend to use organizations that offer vital services such as tax-related services, healthcare support, and various other third-party services. (Mashtalyar et al., 2021). A real-life example of vishing would be the scam attack towards Spectrum Health in 2020. According to reports, Priority Health members and patients received calls claiming to be from Priority Health or Spectrum Health. The callers tried to obtain sensitive information such as passwords, member numbers, and other health-related information (Gleason, 2020).

2.3.5 SMiShing

SMiShing, also known as **SMSishing**, is another form of phishing using short messaging services (SMS). Specifically, adversaries make use of text messages or messages sent via mobile and smartphones to send out deceptive messages to targets to persuade and manipulate them (Desolda et al., 2022; Salahdine & Kaabouch, 2019; Wassermann et al., 2023). Adversaries may send numerous text messages, including links to various malicious websites, and convince the target that the links or the sender are legitimate (Desolda et al., 2022; Parthy & Rajendran, 2019). Adversaries could also send messages containing attachments, which in turn can download and install malware on the device. This malware provides access and sometimes even control to the target’s mobile system (Desolda et al., 2022; Yeboah-Boateng & Amanor, 2014). A real-life example of SMS-based phishing is the USPS delivery scam. Adversaries send messages to targets claiming to have information regarding a USPS package. The message will usually contain a link tempting the target to click and visit the website. The links would then redirect the target to fake online games, whereas others attempt to steal the Google accounts of targets (Bisson, 2020). According to USPS, adversaries will also try to acquire other personally identifiable information (PII) such as social security numbers, date of birth, financial information, and other sensitive information (United States Postal Inspection Service, 2024).

2.4 Large Language Models (LLMs)

This section will provide a brief overview of LLMs, how these systems work, and some of the more commonly seen challenges for LLMs. Large Language Models, usually referred to as LLMs, are a type or category of Artificial Intelligence (AI) or Machine Learning (ML) model designed to comprehend, process, and produce human-like language (Blank, 2023; Chang et al., 2023; Raiaan et al., 2024; Routray et al., 2023). The term “large” in LLMs is coined by the research community for language models that contain millions to trillions of parameters resulting in the models being able to learn more complex relationships (Routray et al., 2023; Zhao et al., 2023). LLMs are typically transformer-based models utilizing neural networks and deep learning techniques to execute a variety of Natural Language Processing (NLP) tasks (Raiaan et al., 2024; Routray et al., 2023; Zhao et al., 2023). These tasks include, but are not limited to, language translation, text summarization, question-answering, chatbot development, text synthesis, sentiment analysis, text generation, and virtual assistants (Fan et al., 2024; Routray et al., 2023; Raiaan et al., 2024).

The transformer architecture is a fundamental component of almost all language models and is designed for sequential data processing. Transformer models are capable of collecting long-range dependencies and contextual information by computing contextual connections between the input tokens through self-attention techniques. (Fan et al., 2024; Raiaan et al., 2024). LLMs are trained and calibrated on an enormous collection of unlabelled text data such as articles, books, codes, and other internet content using the self-supervised learning approach (Blank, 2023; Fan et al., 2024; Raiaan et al., 2024; Routray et al., 2023; Zhao et al., 2023). Via this training, the model can learn the statistical correlations between words and sentences, a language knowledge base can be built, and the model is then able to produce text that is comparable to the material it was trained on (Blank, 2023; Fan et al., 2024; Routray et al., 2023). It is worth noting that this training process necessitates massive computational resources, and completion can take several days to several weeks (Routray et al., 2023).

Emergent abilities

According to Zhao et al. (2023), *emergent abilities* within LLMs can be defined as abilities present in large models but are absent in miniature models. Some of these abilities include in-context learning and instruction following. In-context learning is when a model is trained to produce text as a response to a prompt or given context without the need for an update or additional training. (Chang et al., 2023; Zhao et al., 2023). Additionally, instruction following

refers to LLMs with an enhanced generalization ability, which means being able to follow instructions for new tasks without specific examples (Zhao et al., 2023).

2.4.1 How do LLMs work?



Figure 2. Typical LLM lifecycle

The simple LLM architecture is as follows: data collection, preprocessing, training, and production (Raiaan et al., 2024) shown in Figure 2. Firstly, data collection, as the name suggests, is the collection of text data to be used for LLM training. LLMs are typically trained on a large amount of text data collected from various sources. Secondly, the data collected is forwarded to another stage for preprocessing prior to training. This preprocessing helps remove duplicate data and filter out low-quality data. Thirdly, training refers to the training process undertaken by the model via the execution of a series of stages such as randomized parameter initialization, loss function calculation, iterative training, and more. Lastly, the model is cleared for production and service offering when the training is complete. The services the model will offer vary, but the more commonly known are text summarization, translation of text, and sentiment analysis. In terms of training, LLMs are usually trained via a self-supervised learning technique. Within self-supervised learning, a large unlabelled text data is provided to the model and is tasked with predicting the next word within the sequence. By determining the statistical correlations among words and sentences within the dataset, the model develops the ability to predict the next word (Routray et al., 2023).

LLMs behind the scenes

Using an input prompt, LLMs process sequential data such as paragraphs and sentences utilizing its numerous layers of interconnected neurons. The contextual and “intelligent” response or output is achieved by using the “attention” mechanism. The attention mechanism creates a focused data frame to match the input context, resulting in a more contextual output (Routray et al., 2023). Models such as Google’s BERT (Bidirectional Encoder Representations from Transformers) and OpenAI’s GPT (Generative Pretrained Transformer) series utilize a “self-attention” mechanism, enabling these models to examine the importance of each word

within a sentence, allowing the models to encode deep text relationships (Raiaan et al., 2024). Under the hood, LLMs utilize the probability of words by making a sentence and statistical models of languages to generate data. Thus, LLMs aim to mimic the generative likelihood of sequences of words to predict the possibilities of upcoming or missing tokens (Zhao et al., 2023; Routray et al., 2023). In other words, these models, using a text database, provide responses based on probability. When provided a word, using a collection of texts, the next word is predicted based on a sequence of words selected from the said collection (Routray et al., 2023).

2.4.2 LLM challenges

LLMs have their challenges. A couple of challenges currently faced by LLMs include ethical concerns, privacy and security concerns, and misinterpretation (Fan et al., 2024; Zhang et al., 2023). Ethical concerns regarding LLMs mainly revolve around inappropriate, biased, and harmful content generated by the models, which could put society and its users at risk. Furthermore, due to the vast amount of training data required, LLMs could jeopardize sensitive or confidential information by revealing or precisely guessing data. This ultimately leads to misinterpretations of both the users and the model itself. Misinterpretations of queries provided by users could result in simple frustration or even unintended and adversarial outcomes such as those mentioned above (Zhang et al., 2023).

2.5 Real-world AI-enabled social engineering attacks

This section presents some examples of real-world attacks which were assisted or enabled by AI. The cases presented within this section are real-world accounts based on news reports and articles available online.

CEO fraud via deepfake call

A Chief Executive Officer (CEO) fraud attack has been launched against the CEO of a UK-based energy company via a phone call but with a twist. It has been reported that US\$243,000 (€220,000) has been fraudulently transferred to a Hungarian supplier as a result of the attack. According to the reports, the UK-based CEO was deceived via an impersonation of their German-based parent company's CEO. AI software was utilized to imitate the voice of the German-based CEO. As stated by the UK-based CEO, there were qualities that were recognizable during the phone call that led the UK-based CEO to believe that the phone call was legitimate. Some of these qualities include "a slight German accent," and melody in the

voice. Urgency was also implied as the German-based CEO requested the financial transfer be completed within an hour and was guaranteed a reimbursement. After the transfer of funds, the attack continued, as the attackers contacted and requested additional payments, noting the reimbursement completion of the previous payment. By the third call, further payment is demanded by the attackers, but the UK-based CEO grew suspicious as the call originated from an Austrian phone number and withheld the transfer of funds (Dark Reading, 2019; Sjouwerman, 2019; Trend Micro, 2019).

Deepfake of President Zelenskyy amidst Russia-Ukraine conflict

Amid the Russia-Ukraine conflict, a video surfaced supposedly showing Ukrainian President Zelenskyy urging Ukrainians to lay down arms and surrender to Russia. However, this video has been debunked and is aimed at spreading misinformation and creating confusion among the people. As per the video, President Zelenskyy's voice and face seem to have been altered using AI, resulting in a "deepfake" video. Viewers noted the video's low quality, further raising suspicion about the video's legitimacy. Circulating in social media, Meta (Facebook), Twitter, and YouTube removed the video on their platforms, stating the policy violation committed by the video. The hidden danger of this attack, according to researchers, lies in the fact that it can "erode trust in authentic media." Regardless of quality or origin, this could potentially create uncertainty within people towards future videos of the president. Furthermore, human rights groups and journalists in other parts of the world are fearful due to their lack of ability or tools required to debunk these deepfake videos (Allyn, 2022; Holroyd & Olorunselu, 2022; Wakefield, 2022).

\$25 million deepfake videocall fraud

A multinational company finance worker was deceived into transferring HK\$200 million (US\$ 25.6 million) to various Hong Kong bank accounts as per the instructions of the attacker. According to reports, the finance worker was initially suspicious as the message mentioned a supposed secret transaction to be executed. However, the finance worker was invited to join a video call. During the video call, other members of the staff were present alongside the alleged UK-based Chief Financial Officer (CFO). The colleagues in the video call were recognizable to the finance worker, stating that they looked and sounded like they should. Nonetheless, the finance worker was deceived during this video call. All the staff members, including the UK-based CFO, were AI deepfakes. The financial transfer was executed since the finance worker trusted the legitimacy of this video call. Later that week, the fraud was uncovered as the finance

worker verified the transaction with the company's head office. Presumably, the deepfakes were created using publicly available audio and video footage of the company's staff (Chen & Magramo, 2024; Tan, 2024).

3 Literature review

This chapter is structured as follows: First, the persuasion principles of social engineering section will discuss various social engineering principles based on the seven principles by Cialdini (1984/2006). Second, the section on individual differences will review the differences between humans, which could affect one's susceptibility to phishing. Third, the effectiveness and combination of techniques section will provide examples of technique combinations that could increase the success of phishing. Fourth, the detecting phishing attacks section will provide phishing characteristics which could be helpful in phishing detection. Fifth, the phishing email mitigations section will offer tools and techniques to help lower the success rate of phishing attempts. The AI-generated text detection section will provide an overview of the most common machine-generated text detection techniques. Last, AI-enhanced phishing section will go over ways AI/ML can aid adversaries in their malicious attempts, such as phishing.

3.1 Persuasion principles of social engineering

The seven principles of persuasion will be discussed below (Ferreira & Lenzini, 2015; Wassermann et al., 2023). The majority of these principles are based on (Cialdini, 1984/2006) principles of persuasion: Authority (usually combined with urgency), Social proof, Liking and similarity, Commitment and consistency, Reciprocity, Scarcity, and Distraction.

3.1.1 Authority

People usually adhere to the commands and judgements issued by those they perceive to be in positions of power or authority over them as they are possibly worried of the ramifications of disobedience (Allodi et al., 2019; Wassermann et al., 2023). Additionally, people are conditioned to comply with authority as society teaches individuals not to question it (Ferreira & Lenzini, 2015). Furthermore, people tend to follow experts within a domain or field as they may be regarded as individuals wielding authority such as lawyers, police officers, pastors, or priests (Ferreira & Lenzini, 2015; Wassermann et al., 2023). Authority can be conveyed via email in multitude ways: by using an authoritative tone such as directing others to carry out something, clearly stating in their email signature their function within the company, or have the email indicate to originate from an institution or person of authority (Butavicius et al., 2016; Koddebusch, 2022; Wassermann et al., 2023).

3.1.2 Social proof

People frequently imitate what others do or appear to be doing (Allodi et al., 2019; Ferreira & Lenzini, 2015). This means that if something such as an activity is “socially accepted,” or if plenty of others are also involved then people are more willing to take part in it. Depending on the scenario, this group of people may consist of relatives, coworkers, family members, or even total strangers who are in close proximity to them (such as a restaurant or waiting area) (Wassermann et al., 2023). When others seem to be involved in the same risks and behaviours, people grow less wary and let their guard down. They are then unlikely to be held entirely accountable for their conduct this way (Ferreira & Lenzini, 2015). As an example, people are more likely to attend an event that thousands of other people are also attending. In general, people are prone to trust others who share their views, particularly under circumstances that are unclear (Wassermann et al., 2023). Social proof can be expressed within an email by encouraging individuals to take certain actions as other people, such as fellow coworkers, have already done so (Allodi et al., 2019; Butavicius et al., 2016; Taib et al., 2019).

3.1.3 Liking and similarity

People are attracted to and favour staying with people they seem to be close to or acquainted with, or with whom they believe they know or like (Ferreira & Lenzini, 2015; Taib et al., 2019). Similarly, when it comes to actions, people are more tempted to partake in endeavours they find enjoyable or have previously done so (Allodi et al., 2019; Wassermann et al., 2023). When it comes to email, liking and similarity can be communicated by impersonating well-known companies, a close friend of the target, or a fellow employee (Wassermann et al., 2023).

3.1.4 Commitment and consistency

People are more likely to adhere to choices and decisions they have made in the past (Allodi et al., 2019; Taib et al., 2019). People put significant value on acting in a way that reflects their beliefs and wish to come across as consistent in the things they do (Wassermann et al., 2023; Ferreira & Lenzini, 2015). This is due to the fact that after establishing a commitment to do something, either verbally or in writing, people feel greater confidence in their decision. As with favours, there is an innate tendency to return the favour when it is owed (Ferreira & Lenzini, 2015). Attackers utilize commitment and consistency via email by being consistent with their requests, frequently beginning with modest request and then later increasing their

demands. This is exhibited in some attacks where the target will more likely reply to further requests if they responded to the initial one (Wassermann et al., 2023).

3.1.5 Reciprocity

People usually reciprocate kind actions with further positive actions. The universal notion that people give back to those who have given first is the foundation of reciprocity. For example, someone is more willing to provide assistance to someone who was previously helpful. The initial gesture can be information, compliment or a simple smile and does not require something physical (Taib et al., 2019; Wassermann et al., 2023). In terms of an attack, the attacker may offer promises or unrequested favours to the target which may motivate the target to repay the “kindness” by fulfilling the request or simply enticing the target of personal benefit and gain from the reward (Allodi et al., 2019; Koddebusch, 2022).

3.1.6 Scarcity

People generally place a higher value on rare chances or limited resources. There is less hesitation to buying a product that is advertised by a well-known store as limited (Taib et al., 2019; Wassermann et al., 2023). This concept generally plays on people’s fear of missing out (sometimes referred to as FOMO) on things. This also applies to requests that indicate urgency requiring prompt action (Wassermann et al., 2023). In the context of an attack, when provided with little time, targets are more likely to act irrationally (Allodi et al., 2019). People are driven by these emergency markers to comply with requests without giving them enough thought or weighing the information they received correctly due to the fear of missing out or facing opportunity losses (Allodi et al., 2019; Wassermann et al., 2023). In the context of an email attack, the target might receive an email containing information about a limited offer (Butavicius et al., 2016). Using impersonation, the attacker might pose as someone from the IT department requesting immediate renewal of password at the end of a working day (in this case, the user’s time is scarce), or as someone from the sales department implying a deal could be finalized soon (Koddebusch, 2022; Wassermann et al., 2023).

3.1.7 Distraction

People concentrate on one particular thing and neglect other potentially notable events that may occur without their knowledge. They pay attention to what they can benefit from, what they require, what they are at risk of losing, or whether something could soon become unavailable,

be censored, restricted, or cost more in the future. When making judgements, these distractions have a tendency to amplify people's emotions and lead them to disregard other rational information (Ferreira & Lenzini, 2015).

3.2 Individual differences

This section will discuss individual differences and several demographic factors that can influence human response to phishing attacks, such as age, lower employment years/satisfaction, sender authenticity, email links perception, email style familiarity, emotional attachments to email, and individual habits. According to Hanus et al. (2021), several demographic factors, including age, income, place of residence, type of work, and computer access, can influence a person's susceptibility to email phishing attempts. Studies have shown that there is a strong correlation between an individual's personality traits and their susceptibility to phishing (Carroll et al., 2022). As every individual is unique, the same cues can be interpreted differently in line with a person's context and the email's premise. Due to emotions and personal routines, people may download attachments, click on links, and reply without even questioning the authenticity of the email (Jayatilaka et al., 2021). It is worth mentioning that research regarding phishing susceptibility based on gender has been inconclusive. On one experiment, it has been noted that the click-through rates of male and female subjects did not vary significantly (Taib et al., 2019). Another experiment within local public administration agencies also stated that vulnerability to phishing amongst male and female employees has not differed considerably (Koddebusch, 2022). Thus, several different human behaviour and circumstances will be discussed below to better understand these individual differences: Age, Lower employee years/satisfaction, Sender authenticity, Perception of links in emails, Email style familiarity, Emotional attachment to emails, and Individual habits.

3.2.1 Age

Research shows age plays a role when it comes to phishing susceptibility and that people of older age tend to be more vulnerable to phishing (Beu et al., 2023; Taib et al., 2019). According to the results of phishing simulations and experiments of Beu et al. (2023), Jayatilaka et al. (2021), and Taib et al. (2019), older participants showed the highest susceptibility resulting in a suboptimal behaviour. Specifically, compared to other employees, the participants older than

41 years of age have a noticeably higher likelihood of becoming a victim of social proof type attacks (Taib et al., 2019).

3.2.2 Lower employment years/satisfaction

Fewer years of employment, and poorer employee satisfaction and loyalty are other indicators of suboptimal behaviour (Beu et al., 2023). New employees (less than five years of service) tend to be more susceptible to phishing attacks than employees who have been employed at the company for a longer period (Taib et al., 2019). This could be due to the fact that newer employees are less familiar with the procedures and protocols within their new company. On the other hand, low level of job satisfaction and commitment to the company raises the possibility that negligence rather than a desire to do well might contribute to an individual's phishing vulnerability (Beu et al., 2023).

3.2.3 Sender authenticity

According to the findings of Jayatilaka et al. (2021), people's phishing attack vulnerability substantially increases when receiving an email from a "known" sender. A "known" sender can be spoofed email addresses using (but not limited to) known/convincing email names, (sub) domains, and reply-to addresses. Additionally, a trustworthy-looking sender address within the body of an email can be utilized to mislead the target into thinking that the message originated from that sender.

3.2.4 Perception of links in emails

There are a lot of misconceptions when it comes to links within emails as discussed by (Jayatilaka et al., 2021). Firstly, people tend to be misled by phishing emails lacking links of any kind but urge them for a response or to download the attachments. During the experiment, it was observed that emails containing no links were generally trusted more by the participants. Secondly, phishing emails can deceive individuals by using authentic-looking buttons and URLs (Uniform Resource Locator) or placing supposed optional links within the email body. Thirdly, phishing emails employing an endpoint URL that differs from the URL text can trick people. Additionally, people that are unaware of URL obfuscation and URL structures are still vulnerable to phishing despite checking the URL destination. Also, as people do not always put these tactics into effect, even those who possess knowledge of how to identify URL destinations continue to be at risk of manipulation. Lastly, simply evaluating the network communication

protocol referenced in the URL could lead to drawing incorrect conclusions regarding the security of the URL.

3.2.5 Email style familiarity

Emails that appear and feel identical to what people have previously seen or received can lead to deception. Likewise, people who are unfamiliar with an email could still be manipulated if they choose to compare the content of the email with the material they can find online (Jayatilaka et al., 2021). This can be due to a myriad of reasons, but an explanation could be that attackers can alter or publish information online anonymously and effortlessly unbeknownst to the target.

3.2.6 Emotional attachments to email

Phishing emails that trigger a range of emotions, including (but not limited to) happiness, unease, curiosity, and work-related priorities, could render people at risk of manipulation. Due to their emotions, people may fail to assess the authenticity of an email prior to responding to it. Interestingly, people could respond to emails despite the phishing indicators they notice as a result of emotional attachments to the sender. Likewise, even after recognizing an email as phishing, people can nevertheless click on links within emails influenced by their emotions. Furthermore, phishing mails that fuel strong negative emotions (such as rage, frustration, or fury) could provoke the target to perform reckless actions (such as hitting on unsubscribe links) without thinking through the potential aftermath (Jayatilaka et al., 2021).

3.2.7 Individual habits

A person's habits can affect how they interact with phishing emails. Because of their habits, people could reply to emails before verifying their authenticity. For instance, people who check their emails at a certain time of day may be more vulnerable to phishing emails. Moreover, people's preconceived notions about particular phishing email types could leave them easy targets for phishing (Jayatilaka et al., 2021).

3.3 Effectiveness and combination of techniques

In terms of persuasion principles, the most effective and commonly used is authority (Butavicius et al., 2016; Koddebusch, 2022). Phishing attacks requesting information often utilizes the authority principle coupled with urgency. Research shows that the brain areas in

charge of counterarguments and critical thinking remain essentially dormant when professional guidance is present, hence the distraction principle is amplified by authority (Ferreira & Lenzini, 2015). Per the experiment by Butavicius et al. (2016), when an email included an authoritative tone or figure, participants struggled to correctly differentiate between spear phishing and authentic emails. In the same experiment, authority was also shown to be the technique most likely to persuade users that the email link was trustworthy (Butavicius et al., 2016). Thus, time pressure or urgency along with an authoritative tone or figure increases the chances of a phishing attack succeeding (Ferreira & Lenzini, 2015). Phishing attacks exploit human emotions such as curiosity, kindness, anxiety, and selfishness (Carroll et al., 2022; Wassermann et al., 2023). Attackers will compose carefully written emails to their target exploiting their emotions (Carroll et al., 2022). As described before, people may respond hastily to emails they receive due emotional attachments they establish with them. This emotional attachment could drive people to overlook the authenticity of the email as well as neglect the phishing indicators present within them (Jayatilaka et al., 2021).

A well-crafted spear phishing email tends to be more effective at deceiving targets when compared to generic phishing emails due to their contextual and relevant nature (Butavicius et al., 2016; Koddebusch, 2022). According to the experiment by Butavicius et al. (2016), participants performed considerably worse at recognizing spear phishing emails than generic phishing emails because of the extra contextual information supplied within them. In addition, the success of spear phishing attacks relies on the familiarity and persuasiveness of the email, therefore frequently imitating the structure and contents of authentic emails from legitimate businesses, governmental organizations, or the target's personal contacts with the goal to boost its supposed authenticity (Carroll et al., 2022; Taib et al., 2019). Arguably, the attacker's efforts (such as email personalization and contextualization) are reflected on the message appeal and quality resulting in greater chances of success (Carroll et al., 2022). It is also a common scenario for attackers to engage victims at their most stressful or unstable times. According to previous reports, email overload, working under time constraints, and receiving phishing emails on expected topics have all been linked to an increased risk of becoming prey to phishing attacks (Beu et al., 2023; Ferreira & Lenzini, 2015). Likewise, it is not unusual for phishing emails to be sent out right before the weekend or towards the end of a workday. The target, in this scenario, is likely to race through the process before leaving the office as they are fatigued and lacking focus (Wassermann et al., 2023).

Another factor that contributes to the success of phishing attacks is information richness, and relevance (Carroll et al., 2022; Ferreira & Lenzini, 2015). Firstly, information richness refers to adding rich information in emails which can enhance the attacker's social presence, influence the target's thought processes, and or provide the impression that they are interacting with an actual person (Ferreira & Lenzini, 2015). As result, emails which seem professional or provide full details (such as a business address or contact information) have a higher probability of being trusted by recipients (Ferreira & Lenzini, 2015; Jayatilaka et al., 2021). Richer emails are curated to draw the target's attention towards these rich elements (such as images) to divert their focus from other aspects within the email (Ferreira & Lenzini, 2015). Secondly, relevance refers to phishing emails that include relevant information regarding the target's business or personal interests (Butavicius et al., 2016; Jayatilaka et al., 2021). Spear phishing emails in particular incorporate more specific contextual information to increase the chances that the target will be engaged and prompt a reaction (Butavicius et al., 2016; Carroll et al., 2022). Furthermore, people tend to be more susceptible when it comes to familiar interfaces (such as friends, education, or work) due to the preconceived belief that attackers are unlikely to utilize those emails to initiate their attacks (Carroll et al., 2022; Jayatilaka et al., 2021). It is important to mention that lacking the proper context might increase the target's suspicions as the persuasion might come off as inappropriate (Butavicius et al., 2016).

Ultimately, the effectiveness and success of phishing attacks primarily depends on poor decision making of humans (Beu et al., 2023). According to the experiment of Butavicius et al. (2016), individuals who were less impulsive in their decision making were more likely to perceive links within phishing emails as far more unsafe. Furthermore, within the same experiment, their results also show that reduced cognitive impulsivity offers resistance to high effort, targeted attacks such as spear phishing.

3.4 Detecting (spear) phishing attacks

This section will discuss the most common characteristics of phishing email. In no particular order, these characteristics are: Incorrect grammar and or spelling, unusual or mismatched sender, deceptive or suspicious links/attachments, generic and or unfamiliar greetings, unusual or unexpected requests, urging immediate action, and too good to be true.

3.4.1 Incorrect grammar and or spelling

Incorrect grammar and spelling are some of the more well-known and common giveaways of a potential phishing email. According to a study by Carroll et al. (2022), half of their participants were able to recognize older phishing emails based on grammatical and spelling errors. Typically, official emails undergo review before they are sent out (Caldwell, 2013). Most businesses and organizations often employ writers and editors to ensure that their clients receive content that is of the expected quality (Cofense, 2023a; Lenaerts-Bergmans, 2021; Microsoft, n.d.). Other companies that utilize email services that are web-based make use of web browser capabilities like autocorrect and highlighting or underlining (Cofense, 2023b). Thus, grammatical errors, misspelling, and even odd spacing within emails could be a telltale sign and a characteristic of phishing (Caputo et al., 2014; Microsoft, n.d.; Sturman et al., 2023). However, it is also important to note that not all emails containing these errors are malicious, some of these emails could be the result of improper or awkward translation from a different language (Microsoft, n.d.).

3.4.2 Unusual or mismatched sender

Phishing and malicious emails tend to pose as legitimate companies or known users (such as a CEO or work colleagues) (Lenaerts-Bergmans, 2021; National Cyber Security Centre, 2020). This technique can result in the victim lowering their guard due to familiarity. However, this is yet another characteristic of phishing emails. Frequently, attackers exploit slight misspellings of domain names such as “go0gle.com” or “@mazon.com” or even use other domains such as Gmail while claiming to be from Amazon (Microsoft, n.d.). When it comes to user impersonation, known users such as a “CEO” might be sending emails under another domain like Gmail which deviates from their standard email address (Caldwell, 2013). All of these falls under what can be considered unusual, unexpected, or mismatched senders, which could signal an email phishing attempt. This mismatch can be as simple as the email address and name on the email “From” field to the email claiming to be from one organization but originating from a different domain (Caputo et al.; 2014; Cofense, 2023b).

3.4.3 Deceptive or suspicious links/attachments

Most phishing emails utilize fraudulent links or attachments. Firstly, links included within phishing emails tend to have a mismatch when it comes to what is displayed and where the link eventually brings the user, sometimes referred to as a deceptive link or URL (Lim et al., 2021;

Sturman et al., 2023). These deceptive links usually lead the user to a forged web page that is made to look like a legitimate one. Suspicious links arise when unexpected or out-of-the-norm email arrives urging the user to click or interact with the embedded link. Imagine the IT department of the user's company sending out an email asking users to install an update by clicking the included link (Caldwell, 2013). Secondly, attachments are commonplace when it comes to business communications within emails. Attachments provide attackers with the same opportunity to spread malicious activity via email attachments within phishing email attempts besides links. Filename extensions like .RTF, .XLS, and .ZIP are often used and trusted and thus also extensively used in spear phishing attempts (Song et al., 2015). Therefore, attachments or links within emails originating from an unknown or unexpected source must be handled cautiously.

3.4.4 Generic or unfamiliar greetings

Another characteristic of phishing emails is the usage of generic or vague greetings such as "Dear user" or "Dear customer." Attackers may also use more generalized terms such as "everyone" or "all" (Sturman et al., 2023). The use of such unfamiliar greetings may raise a red flag, especially in a setting where colleagues or family would usually be less formal and direct (Cofense, 2023a). Alternatively, the same can be said when a colleague or acquaintance with less interaction with the user is unexpectedly overly friendly (Cofense, 2023b). It is also worth mentioning that even banks or well-known e-commerce the user has registered with would typically address the user by their first or last name instead of a generic "Dear sir/madam" (Lenaerts-Bergmans, 2021; Microsoft, n.d.). Ultimately, if an email is specifically addressed to the recipient, then the email should make sense as expected, and the email should refer to the recipient instead of "colleague" or "user" (Caldwell, 2013; National Cyber Security Centre, 2020).

3.4.5 Unusual or unexpected requests

Unusual or unexpected requests can take many forms when it comes to phishing emails. Peculiarities can be an unusual email sent time, sharing media, or forwarding of email (Caldwell, 2013; Lim et al., 2021). Receiving an email from someone for the first time is not uncommon, primarily if they are not affiliated with the company. However, this could still be considered a sign of a phishing attempt (Microsoft, n.d.). Furthermore, unexpected emails could also contain unusual requests such as asking for sensitive information like payment or login

(Cofense, 2023a). These out-of-the-norm requests further increase the possibility that the email might genuinely be a phishing attempt. Another example could be an out-of-procedure payment request or instruction from a colleague or the CFO (Cofense, 2023b). Furthermore, official establishments such as banks, governmental bodies, and other legitimate businesses would usually not request personal information (such as social security numbers, payment information, and more) via email (Lenaerts-Bergmans, 2021; National Cyber Security Centre, 2020).

3.4.6 Urging immediate action

Insisting or encouraging immediate action, such as clicking links or downloading/opening an attachment, is another common tactic used by attackers within phishing emails (Caputo et al., 2014; Microsoft, n.d.; National Cyber Security Centre, 2020). Adversaries tend to generate a false sense of urgency to panic the victim, seemingly depriving them of the possibility to inspect the email more thoroughly or consulting with someone else (Cofense, 2023a; Cofense, 2023b; Microsoft, n.d.). The most common call for immediate action is either a promise of some reward or repercussions (Microsoft, n.d.). The promise of a reward tends to play on people's emotion of missing out. In contrast, repercussions target people's fear, effectively creating a sense of threat (Lenaerts-Bergmans, 2021). Therefore, emails that pose a threat, such as unwanted consequences or opportunity loss unless prompt action is taken, must be handled with caution as they could be a phishing attempt (Cofense, 2023a; Cofense, 2023b).

3.4.7 Too good to be true

Lastly, phishing emails occasionally portray a "too good to be true" nature (National Cyber Security Centre, 2020). These emails tend to motivate the victims by promising a prize, reward, or discount in exchange for a relatively simple task such as clicking a link or opening an attachment (Cofense, 2023a; Cofense, 2023b). A characteristic of such an email is that one should not receive any packages or win prizes if they have not participated in any activities. For instance, one should not be receiving an email about a package if they have not ordered anything. Winning a prize from a company lottery is also highly unlikely without buying a ticket or a raffle taking place altogether (Caldwell, 2013).

3.5 Phishing email mitigations

A variety of mitigation techniques and strategies can be applied to minimize the success or damage of phishing attacks. In no specific order, these are: awareness training, phishing simulation, multifactor authentication, social reporting/crowdsourcing, automated tools, and secure routines.

3.5.1 Awareness training

Awareness training is an umbrella term for many security-focused training and is one of the more common mitigation strategies against social engineering attacks. Awareness training can also be referred to as cybersecurity awareness training, cybersecurity training, or security awareness training (SAT) (Abroshan et al., 2021; Ironscales, 2024b; KnowBe4, n.d.; Wash, 2020). The objective of awareness training campaigns or programs is to improve the awareness and security knowledge of users, such as detecting social engineering attacks via learning (Abroshan et al., 2021; Gomes et al., 2020; Sumner & Yuan, 2019; Varshney et al., 2024). This training also conveys “best practices” and safety procedures to users and the intended audience (KnowBe4, n.d.; Wash, 2020). Similarly, phishing training, sometimes referred to as anti-phishing training, is focused explicitly on phishing awareness (Gomes et al., 2020; Naqvi et al., 2023; Varshney et al., 2024; Wash, 2020). Phishing training includes phishing detection based on cues, safety procedures post-phishing discovery (such as reporting), secure email practices, and more (Abroshan et al., 2021; Agazzi, 2020; Caldwell, 2013).

However, there are also important things to consider when utilizing awareness training. Firstly, the training must be tailor-made and not generic, as this could result in an irrelevant or less effective program (Abroshan et al., 2021; Ironscales, 2024a). Specialized training should account for the organization’s requirements, cultural differences, regulations, security experience, a load of information, perceived support, and more, effectively adopting a human-centred design (HCD) (Abroshan et al., 2021; Caputo et al., 2014; Naqvi et al., 2023). On this note, with a proper assessment, specific organizations might discover other attack vectors, such as phone calls and SMS, which can be included in the training (Ironscales, 2024a). Secondly, the training frequency and training types. It is important to regularly provide and select a suitable type of awareness training to help keep users up to date and increase overall retention of the training material (Ironscales, 2024b; Ironscales, 2024a; Lim et al., 2021; Sumner & Yuan, 2019). Training types include virtual labs, web platforms, online training, simulations,

gamification, and embedded training. (Abroshan et al., 2021; Caputo et al., 2014; Naqvi et al., 2023; Sumner & Yuan, 2019; Zhuo et al., 2023). On the other hand, providing excessive training could lead to information overload, dissatisfaction, complaints, and more. Thus, it is also essential to consider the users' workload and priorities when assembling the program (Abroshan et al., 2021). Lastly, merely offering training materials does not aid in recognizing attacks and could potentially harm users' confidence (Lim et al., 2021).

3.5.2 Phishing simulation

Another common mitigation strategy usually included in awareness training programs is phishing simulation (Caldwell, 2013; Ironscales, 2024b; KnowBe4, n.d.; Naqvi et al., 2023). Phishing simulation in this context refers to creating a phishing mail mock-up and sending it to the target users, usually within a company or department. Phishing simulations are considered immersive training techniques as they should closely resemble real-world scenarios (Caldwell, 2013). Phishing simulations also provide a metric for assessment by observing the users' reactions (for example, clicking the link) (Varshney et al., 2024). Weaknesses can be pinpointed based on these reactions and assessments, and custom-tailored training and effective methods can then be presented to help protect the users against these phishing attacks (Caldwell, 2013; Ironscales, 2024a). Phishing simulations can also allow for quicker feedback to the users by presenting them with simple training materials to better protect themselves in the future should users fall prey to the simulation (Caldwell, 2013; KnowBe4, n.d.; Naqvi et al., 2023). The simulation frequency will all depend on the company requirements; however, running the simulation alongside the awareness training program and executing it regularly, at least quarterly (Caldwell, 2013) is advisable.

3.5.3 Multifactor authentication

Multifactor authentication (MFA) slowly became one of the more popular security-focused solutions against malicious attacks (Zhuo et al., 2023). MFA secures the system by mandating users to authenticate and verify their identity through the use of multiple credentials (such as another hardware like a smartphone) alongside username and password (Agazzi, 2020). Multiple ways exist to implement and use MFA, but the core functionality remains unchanged. Examples of MFA include the generation of One-Time Password (OTP), PIN, SMS notification, Quick Response (QR) or barcode, push notification via an app, hardware tokens, biometrics (fingerprint/iris/facial scan), preset questions, and more (Agazzi, 2020; Naqvi et al.,

2023; Varshney et al., 2024; Zhuo et al., 2023). One of the more commonly used types is app-based authentication. Ultimately, the functionality of MFA pertains to three types of “authentication factors” (Agazzi, 2020): Something the user knows (username/password), Something the user has (a generated code via an app or email), Something they are (fingerprint or facial scan).

3.5.4 Social reporting/crowdsourcing

Engaging and encouraging users is another method that, when incorporated correctly, can significantly improve an organization’s resilience against malicious attacks. There are two parts to this user engagement: social reporting and crowdsourcing (Ironscapes, 2024b; Wash, 2020). Social reporting refers to encouraging and incentivizing users to report any (deemed) suspicious activities rather than remaining silent out of concern for adverse consequences (Caldwell, 2013). An organization’s IT or security department could set up a central location where any user could report or forward suspicious activity such as emails (Caldwell, 2013; Cofense, 2023b; Ironscapes, 2024b; KnowBe4, n.d.; Varshney et al., 2024; Wash, 2020). This central location can then perform virus scanning, authenticity checking, email verification, and more (Cofense, 2023b; Wash, 2020). Social reporting also allows for the creation of a proper reporting protocol or procedure for users to follow to trigger an incident response process (Ironscapes, 2024b). Additionally, crowdsourcing goes hand in hand with social reporting. Crowdsourcing utilizes the advantage of intelligence from the users within the organization itself (Naqvi et al., 2023). Harvesting user intelligence could give insight into zero-day and other phishing-type attacks (Ironscapes, 2024b). Ultimately, the idea is to empower users that when they “see something, say something.” The earlier the users report suspicious activities, the earlier an incident response can occur, and damage can be kept to a minimum (Cofense, 2023b).

3.5.5 Automated tools

Implementing and using automated tools is highly recommended as it helps offload some of the manual labour that the users would normally perform. Some automated tools could also help users in decision-making, such as providing warnings through pop-up messages (Wash, 2020). In terms of phishing emails, automated tools exist, such as spam filters, firewalls, black/whitelists, DMARC email authentication, anti-phishing tools, link/file scanners, and email header/content scanners (Agazzi, 2020; Gomes et al., 2020; Naqvi et al., 2023; Song et

al., 2015; Zhuo et al., 2023). Other automated tools that can be implemented include Intrusion Detection and Prevention System (IDS/IPS), Host-Based Intrusion Detection System (HIDS), antivirus software, and browser extensions to warn about potential phishing (Agazzi, 2020; Caldwell, 2013; Gomes et al., 2020; Naqvi et al., 2023; Song et al., 2015).

3.5.6 Secure routines

Developing and embedding secure routines can lead to safe email practices (Naqvi et al., 2023). As part of the awareness training, teaching users about safe email habits could potentially lead to a decrease in user impulsivity (Caldwell, 2013). Examples of secure routines include best practices such as being wary of email links, being cautious of emails requesting personal information/credentials/payment information, being careful when urgent action is required, being mindful of unexpected or unknown emails, refraining from mindlessly opening attachments, double-checking links by hovering over the link, verifying the sender against previous emails, and many more (Cofense, 2023a; Cofense, 2023b; Gomes et al., 2020; Lenaerts-Bergmans, 2021; Microsoft, n.d.; Naqvi et al., 2023; Varshney et al., 2024;). Moreover, these secure practices can be enforced via the creation of policies for users to follow (Varshney et al., 2024). However, it is essential to remember to respect and consult the users prior to deployment and during the creation of such policies to ensure proper user compliance (Naqvi et al., 2023).

3.6 AI-generated text detection

Several techniques are used to detect machine-generated texts. However, this thesis will discuss only a few of the more common used techniques. These techniques include watermark-based detection, feature-based detection, LLMs as detectors, and the combination of human and machine. Finally, some text detection challenges will be presented regarding the aforementioned detection techniques.

3.6.1 Watermark-based detection

The more common types of watermarking techniques are data-driven, model-driven, and post-hoc or post-processing (Wu et al., 2025; Tang et al., 2024). First, data-driven watermarking refers to embedding tags or patterns within the LLMs training dataset. This type of watermarking usually relies on backdoor insertion. That is, a small amount of samples containing a watermark is added to the training dataset, allowing the model to essentially learn

the defender's secret function. A specific trigger then activates the backdoor watermark. Second, model-driven watermarking manipulates the token sampling or logit output distribution during the inference process, which results in directly embedding watermarks into the models. Regarding the logit-based technique, prior to the generation of text, this technique chooses a set of "green" tokens indiscriminately while defining the rest as "red" tokens. During the sampling, the model is kindly guided to choose from the "green" set of tokens. The token sampling, on the other hand, as the name implies, manipulates the process of token sampling via the setting of specific patterns or random seeds, resulting in watermarks. During detection, the text is aligned with the random number sequence by utilizing the secret key. Last, post-hoc or post-processing is a technique that embeds a watermark (such as an identifier or a hidden message) after the text generation of the model. Some examples include swapping or insertion of special Unicode characters, substitution of synonyms, and incorporation of a secret message (such as random binary bits) via text modification.

3.6.2 Feature-based detection

Feature-based detection is a technique that uses textual features to detect machine-generated texts. An example of a feature-based approach is the creation of feature vectors from input sequences by applying Natural Language Processing (NLP) and then using a downstream classification algorithm, such as neural networks, decision trees, or random forest, to classify these feature vectors (Crothers et al., 2023). Some of the more commonly known textual features include perplexity, repetitiveness, and coherence. The level at which a language model can accurately predict a word sequence is measured via perplexity. The lower the perplexity, the better the model's prediction of the next word in a sequence. Human writers possess experience, knowledge, and creativity from abundant reading, resulting in a more varied and unpredictable way of writing (Lyu et al., 2022; Mindner et al., 2023). Thus, texts produced by humans tend to contain unexpected structures, word combinations, and ideas. On the other hand, statistical and common patterns learned during training are usually the basis of machine-generated text, leading to a more repetitive and predictable output (Mindner et al., 2023; Tang et al., 2024). Additionally, the lack of diversity and repetitiveness in machine-generated texts is also the result of the over-usage of frequent words (Fröhling & Zubiaga, 2021). Lastly, models usually produce short texts as the generation of longer paragraphs and sentences tends to result in a lack of coherence and consistent text (Crothers et al., 2023; Fröhling & Zubiaga, 2021). Other features to consider include word frequencies, part-of-speech tags, the amount of punctuation and quotation marks used, and spelling and grammar errors (Crothers et al., 2023;

Fröhling & Zubiaga, 2021; Mindner et al., 2023; Tang et al., 2024). Due to the repetitive nature of machine-generated text, overlap of parts of speech and words can occur between sentences. Furthermore, punctuation and quotation marks are less commonly used by models when producing texts. Finally, compared to human text, AI produces far fewer grammar and spelling mistakes.

3.6.3 Detection via LLM

In certain circumstances, LLMs can be used as the detector itself. Some of the more popular and well-known LLMs used as detectors include GPT-2, Grover, BERT, and RoBERTa (Jawahar et al., 2020). The detection task may also not be limited to the model's own output but also the detection of output produced by other similar models. Generally, these models are not used directly as detectors but are fine-tuned to perform this task (Fröhling & Zubiaga, 2021; Jawahar et al., 2020). OpenAI fine tunes and utilizes a RoBERTa-based detector to differentiate GPT-2 and human-generated texts (Sadasivan et al., 2023). In some instances, some generative models can be used to distinguish their own outputs or outputs from similar other generative models without fine-tuning (Crothers et al., 2023). Furthermore, observations from numerous studies have shown that smaller models can be utilized to detect text produced by larger models (Crothers et al., 2023).

3.6.4 Human-machine combination

Human-machine combination refers to the detection technique wherein humans and machines work together to accomplish the detection task. Generally, humans are less proficient at detecting machine-generated texts when compared to machines (Wu et al., 2025). According to a review, untrained human evaluators accurately identified GPT-3 generated text at a level equivalent to random chance (Crothers et al., 2023). Presenting examples to human evaluators prior to the detection task can enhance their performance, particularly with lengthier samples (Wu et al., 2025). Nonetheless, tools have been proposed and created to help aid human evaluators. Some examples of these tools include the Giant Language Model Test Room (GLTR) and the SCARECROW framework. GLTR displays the per-token model likelihood, per-token rank, and entropy of the projected subsequent token distribution (Jawahar et al., 2020). Thus, the GLTR system improves the detection of machine-generated text by integrating a human reviewer (Crothers et al., 2023). The SCARECROW framework (Dou et al., 2022), on the other hand, guides users in detecting machine-generated text by outlining ten error types

(Wu et al., 2025). These errors are made from text generated by GPT-3, thus training users to annotate these specific errors (Crothers et al., 2023).

3.6.5 Text detection challenges

Text detection itself has its fair share of challenges. These challenges include mixed-text detections, lack of a universal evaluation framework, and ethical concerns. Mixed-text detection challenges arise when machine-generated text is modified on the sentence or paragraph level by a human. Furthermore, current detectors struggle with texts of shorter lengths. Moreover, detectors' performance also degrades when it comes to paraphrased or re-written text either by humans or another machine (Weber-Wulff et al., 2023; Wu et al., 2025). The lack of an effective universal evaluation framework is a concern when it comes to the comparison of studies and research regarding the performance of their respective detectors. This results in suboptimal performance on other researchers' test sets (Wu et al., 2025). Ethical concerns relate to the fairness and interpretability of the detectors and their detection methods. False positives can have negative consequences on individuals of concern. Particular groups of people may be more vulnerable than others to have their texts flagged by machine-generated text detection algorithms resulting from their non-malicious usage of translation programs (such as Google translate or DeepL) or literary characteristics (for instance, language background) (Crothers et al., 2023; Weber-Wulff et al., 2023).

3.7 AI-enhanced phishing

This section will discuss various aids that AI/ML has to offer for adversaries in terms of malicious attacks, such as filter bypass, human-like text, semi-automation, personalisation, and cost-effectiveness.

3.7.1 Filter bypass

A properly trained AI can be used to bypass or evade detection (Brundage et al., 2018). Evasion, in this case, refers to avoiding detection via traditional security defences (Schmitt & Fléchais, 2024). AI or ML could learn what the common phishing red flags are, or which content is being filtered (Jackson, 2023; Schmitt & Fléchais, 2024). The AI can then proceed to generate content such as emails, which can evade detection or deceive filters and other security software (Schmitt & Fléchais, 2024). The crafting of social engineering messages by AI has been showcased by the work of (Das & Verma, 2019). Furthermore, AI can also create malicious URLs, which can

bypass web and URL detectors (Neupane et al., 2023). As shown in the works of Anderson et al. (2016), AI can generate a malware URL that can evade a detection system based on a Deep Neural Network (DNN). Additionally, AI is not only limited to the text body, but it can also learn which email headers are being filtered, thus effectively bypassing these filters as well (Jackson, 2023).

3.7.2 Human-like text

AI can further reduce the workload of attackers by producing content such as texts that appear or sound human-like (Hazell, 2023; Schmitt & Fléchais, 2024; Weiss, 2019). This human-like ability is made possible by state-of-the-art LLMs such as GPT-3 and GPT-4 alongside proper prompting, seeding, and adjustment of style and tone (Jackson, 2023; Lin, 2023; Neupane et al., 2023). In the case of phishing, the incorporation of this human-like ability can result in more deceptive and convincing content (Heiding et al., 2023). Adversaries will then be able to reproduce legitimate-sounding email replies by mimicking legitimate email properties (Gupta et al., 2023; Karanjai, 2022). Moreover, with the improving translation capabilities of AI, the attacker and the target do not even need to speak the same language (Brundage et al., 2018). According to an experiment performed by Baki et al. (2017), the participant's performance when it comes to detecting legitimate and fake emails was close to random. In another survey by Weiss (2019), bot and human-generated content were also classified correctly half of the time, which would be classification close to random.

3.7.3 Semi-automation

Another potential contribution of AI towards adversaries is the capability of (semi) automation. AI automation presents itself in multiple ways, such as mass scaling attacks, automated gathering of information, assisting in target selection, and overall increased quality of attacks. First, mass scaling, as the name suggests, scales the attack in order to target more victims, to launch more attacks, or both. As AI becomes more advanced, the creation of malicious phishing emails can be automated, allowing attackers to craft and distribute them non-stop and with ease at a scale. However, the true potential of AI when it comes to mass scaling comes in the form of its ability to craft highly deceptive spear phishing emails at the same rate proportional to general lower-quality mass phishing campaigns (Brundage et al., 2018; Hazell, 2023; Heiding et al., 2023; Karanjai, 2022; Neupane et al., 2023; Roy et al., 2023; Schmitt & Fléchais, 2024;

Traficom, 2022). As shown and confirmed in the experiment of Weiss (2019), comments generated by bots were posted at a scale without many issues.

Second, AI can help accelerate and automate the information-gathering process prior to SE attacks. Acceleration and automation of this task can be done in multiple ways. As people continue to adopt a more digital lifestyle, progressively more data regarding everyone will be discoverable online which can be possibly used against them (Brundage et al., 2018). Mining, analysing, summarizing, and learning from such big data, even unorganized and unstructured, would be of no challenge to leading state-of-the-art AI models (Brundage et al., 2018; Hazell, 2023; Kaloudi & Li, 2020; Neupane et al., 2023; Schmitt & Fléchais, 2024; Seymour & Tully, 2018). Some examples of such big data include social media posts, long documents, reports, and more (Traficom, 2022). Relevant information extracted from the datasets can reveal points of interest regarding the targets' affiliations, hobbies, personality, behaviour, communication patterns, interests, estimation of personal wealth, and previous other activities (Brundage et al., 2018; Gupta et al., 2023; Schmitt & Fléchais, 2024; Seymour & Tully, 2018). The assistance offered by AI by automating the background research on targets could result in more actors engaging in spear phishing, potentially making targeted phishing commonplace (Brundage et al., 2018; Hazell, 2023).

Third, AI can assist attackers in target selection. Target selection refers to a more efficient identification, analysis, and prioritization of susceptible targets (Brundage et al., 2018; Traficom, 2022). For example, AI can be used to analyse which group of users on a particular social media platform would be the most vulnerable to phishing by evaluating their online behaviour (Brundage et al., 2018; Jackson, 2023). Moreover, this is not limited to only social media platforms; AI can also be used to identify targets within an organization by analysing their online profile and activities (Traficom, 2022).

Last, AI will generally increase the overall quality of phishing attacks. One of the more noticeable quality increases would be the reduction of spelling and grammatical errors due to the automatic generation of texts since well-trained models generally do not make these mistakes (Hazell, 2023; Neupane et al., 2023). Additionally, in combination with automated information gathering and target selection, AI could generate more deceptive and convincing content potentially increasing the success rate, impact, and damage of attacks (Gupta et al., 2023; Heiding et al., 2023; Neupane et al., 2023; Schmitt & Fléchais, 2024). Also, AI generating higher quality and more deceptive content is definitely advantageous to lesser-skilled attackers

(Hazell, 2023). Not to mention, this increased quality via more deceptive content is not limited to text; highly realistic fake generated content also includes videos and audio labelled as “deepfakes” (Neupane et al., 2023).

3.7.4 Personalization

As mentioned earlier, AI can be used to conduct automated information gathering and target selection. Subsequently, with this data and information, AI can be used to generate highly personalized and contextualized content, further improving the success of attacks (Hazell, 2023; Heiding et al., 2023; Schmitt & Fléchais, 2024). The AI’s ability to produce content in line with the expectations of the targets drastically improves the chances of the victims cooperating or responding (Gupta et al., 2023; Seymour & Tully, 2018). For instance, AI could be used to adjust the tone or communication style to further lure the victim (Neupane et al., 2023). Likewise, AI can also be utilized to impersonate or mimic someone related to or trusted by the target, such as family members, colleague, or acquaintance (Brundage et al., 2018; Gupta et al., 2023; Hazell, 2023; Heiding et al., 2023; Karanjai, 2022; Neupane et al., 2023; Traficom, 2022). Alternatively, AI can also be used to create fake personas online with the objective of establishing contact or pretexting with the intended victims (Schmitt & Fléchais, 2024; Traficom, 2022). Additionally, adversaries can also utilize AI to generate disinformation such as hoaxes and fake news which can harm organizations, individuals, as well as entire societies (Karanjai, 2022). For instance, an adversary could look into a researcher’s h-index trend from the last few years by parsing Google Scholar data. The attacker can then exploit AI models to summarize the trends within the data and combine this newfound information when generating a phishing mail (Giaretta & Dragoni, 2019). As shown in the experiment by Lin (2023), several subjects were deceived by their AI-generated phishing mail, and some even stated that they would follow up once they had read the paper. Thus, this study demonstrates that AI-generated content is hard to discern, even more so when combined with enough context, making them trustworthy enough at first glance (Lin, 2023).

3.7.5 Cost-effectiveness

Ultimately, AI can effectively reduce the costs of phishing attacks. AI enhancement, such as phishing mail generation, considerably reduces the costs of phishing attacks, especially targeted or spear phishing attempts (Hazell, 2023; Heiding et al., 2023). For instance, offloading the background research, target selection, and personalization tasks could render the phishing

attacks as economical as lower-quality mass-scale phishing mail (Heiding et al., 2023). Besides, as with other technologies, costs would eventually decrease regarding the acquisition and deployment of AI capabilities, allowing a wide range of attackers to profit from the malicious use of advanced AI tools, including those with limited resources (Schmitt & Fléchais, 2024).

4 Research methodology

This chapter will provide an overview of this thesis's research and implementation methodology. Background and literature will discuss the criteria used during the thesis's paper and article selection process. Creation of CARLS checklist will briefly outline the proposed checklist and manual evaluation. Implementation with LLM will summarise the test methodology used for the various thesis experiments. Figure 3 provides an overview of the overall flow of the thesis. From left to right, the background sets the basis of the fundamental topics relevant to the thesis. The literature dives deeper into the thesis's main subject, phishing. Following the literature review results, the creation of CARLS checklist is a collection of various phishing characteristics and the proposed checklist/guideline of this thesis. Lastly, implementation with LLM evaluates the checklist in an automated manner using different methods, including LLM utilisation, comparison with another model, and testing on real-world phishing data.

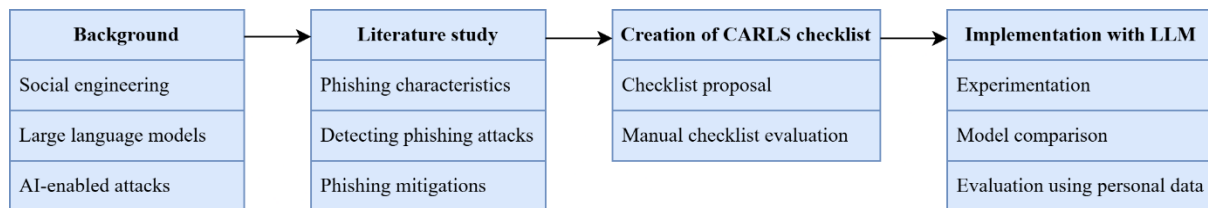


Figure 3: Research methodology flowchart

4.1 Background and literature

The papers and articles used within the background and literature review were gathered using online research databases, such as Google Scholar, Institute of Electrical and Electronics Engineers (IEEE), Association for Computing Machinery (ACM) digital library, ScienceDirect, ArXiv, ResearchGate, and SpringerLink. Recently published papers were prioritised on topics, such as LLMs and AI-enabled attacks. Recently published papers refer to a publication date of around the last decade or so at the time of writing. This criterion was applied to ensure that the most updated information regarding the core topics of this thesis was used and reviewed. Additionally, any recent evolution in social engineering attacks is more likely to be discussed within the recently published papers and articles. Moreover, the papers and articles were chosen based on search keywords such as social engineering, phishing, AI, artificial intelligence, large language models, LLM, email phishing, AI phishing, and more. The relevant papers presented

by the databases were then further filtered by reading the abstract. The only criterion applied when filtering the papers while reading the abstract is the absence of AI utilisation, such as phishing attack detection. The choice to avoid papers utilising AI is due to the primary focus of this thesis, which is a manual and human-operated guideline/checklist. Furthermore, the final product comprises a manual checklist, thus lacking AI utilisation.

4.2 Creation of CARLS checklist

This thesis produces and proposes a checklist called the “CARLS checklist” as an acronym originating from the initial characters of the perception characteristics: **C**ommitment & consistency, **A**uthority, **R**eciprocity, **L**iking & similarity, **S**ocial proof, and **S**carcity. The CARLS checklist is formed through the systematic identification and collection of various characteristics related to phishing email detection and identification, as outlined in the literature. From all the factors discussed in the literature, only the most common, recurring, and those with the highest success rate were included in the checklist. The checklist is divided into two (2) separate categories: email perception and email characteristics. This proposed checklist is then manually evaluated by using the checklist in a human-operated manner against various LLM-generated phishing emails, along with a result and a score breakdown.

4.3 Implementation with LLM

The thesis implementation is divided into three (3) different evaluations: the experiment, the comparison, and the personal email test. Firstly, the implementation experiment is done by having each LLM test the other LLMs, specifically generated spear phishing emails. Secondly, the implementation comparison compared the local LLM (GPT4All) performance against the AI model presented by Jamal et al. (2024). The comparison consisted of two (2) separate datasets, namely phishing and spam/ham emails. The test was then separated into a balanced and an imbalanced dataset comprising ten (10) emails each with a total of twenty (20) emails. Lastly, the personal email test is an experiment using data collected from the researcher’s spam folder. Ten (10) emails were handpicked and tested against the same local LLMs.

5 Phishing detection checklist

The structure of this chapter is as follows: The CARLS checklist will provide an overview and present a phishing email detection checklist. The checklist testing methodology provides the system specifications, prompt queries, and AI models used during the manual testing of the CARLS checklist. The CARLS checklist evaluation assesses the checklist's performance against various LLM-generated phishing emails.

5.1 CARLS checklist

The terms “CARLS” came about as an acronym based on the first letters of the perception characteristics. Aligning the perception characteristics in a certain way would reveal the initial characters: **C**ommitment & consistency, **A**uthority, **R**eciprocity, **L**iking & similarity, **S**ocial proof, and **S**carcity. The CARLS checklist is intended to be human-operated and for phishing emails only. Other types of phishing besides email phishing are considered out of the scope of this thesis. Table 3 will explore the user's perception of the suspicious phishing email. Table 4 will help identify certain characteristics found in potential phishing emails. Table 5 presents an overview of the point system and the corresponding phishing probability chance. The points within the tables were allotted based on the effectiveness or usage frequency of features within phishing emails. All the sources from which the checklist was created can be found in Table 6.

Table 3. Email perception checklist

Perception	Description	Checkbox	Points
Authority	Does the email emphasize, portray, or amplify any sort of authority or higher status? Authority can be portrayed via status, tone, function, and more.		2
Social proof	Does the email depict or claim that others (such as colleagues, friends, or family) have participated or taken part in whatever the email is suggesting?		1
Liking & similarity	Does the email present something familiar to you? Something familiar could be an activity you are planning to do or have done before, AND is the sender unexpected or unfamiliar?		1
Commitment & consistency	Is the email request part of a series of emails you previously received, AND is the behaviour unexpected from the sender? (for example, an IT/Bank employee should not be asking for access codes, files, passwords, usernames, or other personal information).		1
Reciprocity	Does the email offer something in exchange as a token of gratitude or thanks for obliging a request? (such as a gift in exchange for providing information).		1

Perception	Description	Checkbox	Points
Scarcity	Does the email imply or represent something of great value or precious? Does the email depict the availability of something as limited? (Such as “today only,” “24 hours left”, or “X amount left”).		1
		Total	

Table 4. Email characteristics checklist

Characteristics	Description	Checkbox	Points
Relevance	Is the email unexpected or unusual? (for example, receiving a confirmation email for a purchase or delivery you did not make) Is the email received at an unusual time? (for example, outside office hours or in the middle of the night).		3
Urgency	Does the email imply urgency? (for example, an action must be taken as soon as possible, or within a certain time frame).		3
Consequences	Does the email indicate or threaten with negative consequences for failing to comply? (such deletion or disabling of account or being discharged).		2
Sender authenticity	Is there a mismatch between the sender’s name and the “From” field? Is there a mismatch with the domain? (such as a misspelling, for example, “go0gle.com”) Is there a mismatch with the sender’s domain? (For example, a “bank” employee BUT using a “Gmail” or “Hotmail” account).		2
Hyperlinks authenticity	Does the email encourage you to click or visit a link included within the email? (such as logging in or viewing information via the link).		2
Attachments authenticity	Does the email encourage you to download or view a file/attachment included in the email? (such as a file/document which may contain information you never requested).		2
Email style	Is the email style different or unexpected? (such as logos not appearing correctly or oddly, different email layouts or fonts, and so on.).		1
Spelling & grammar	Are there multiple or noticeable spelling errors or grammatical mistakes within the email?		2
Greetings	Does the email use generic or unfamiliar greetings? (such as “Dear customer,” “Dear user,” “Hello,” and so forth.).		2
Too good to be true	Does the email offer a prize, reward, discount, and the like after completing a task such as registering via a link? Does the email congratulate you with a prize for an event you do not remember participating in? (such as winning a raffle or lottery among others).		2
		Total	

Table 5. Checklist points system and phishing probability

Total points	Phishing probability	Description
9 or more	High	There is a high probability that the email in question is a phishing email. Request assistance from the security or IT department regarding the next steps. If no such department exists, proceed with caution and check the applicable countermeasures in section 3.5.
4 - 8	Medium	There is a moderate chance that the email in question is a phishing email. Request assistance from the security or IT department for a second opinion and next steps. If no such department exists, proceed with caution and check the applicable countermeasures in section 3.5.
3 or less	Low	The probability of the email being a phishing email is low. However, it is advised that every email should still be treated with enough caution.

The checklist introduced above may seem familiar and this is the case. With the rise of AI and ML, particularly LLMs, phishing did not in itself become more complex. Rather, with the help of LLMs, phishing became more efficient, cost-effective, and easier to execute for adversaries. Furthermore, adversaries utilizing LLMs would make personal, relevant, and more targeted phishing emails, also referred to as spear phishing emails, increasingly common all without technically changing the nature of phishing emails. The phishing probability has been divided into three levels, making it easy for the user to follow, understand, and operate. With the already overwhelming responsibilities and tasks users face daily, the checklist is created in a way that does not add complexity and burden to its users. Furthermore, most people would most likely be already familiar with the “low, medium, high” style metric, and if not, it should not be too difficult to grasp for users new to this type of metric.

5.1.1 Machine-generated text detection

As of the moment of this writing, there is currently no consistent way to recognize or detect machine-generated text relying solely on human capabilities. As mentioned previously and exhibited in the works of Baki et al. (2017) and Weiss (2019), the classification of human and bot-generated text was close to random as the participants classified roughly half of the text correctly resulting in a mere 50/50 chance. Thus, there would be no checklist or guideline within this thesis regarding the detection or classification of machine, bot, or AI-generated texts.

5.2 Checklist testing methodology

This section is structured as follows: prompt queries will provide an overview of the prompt-engineered queries fed to the LLMs that resulted in the machine-generated text used within the experiment. Perplexity AI and GPT4All will provide the rationale as the chosen models for this experiment. Table 7 provides an overview of machine and system specifications used during the testing and experimentation.

Table 7. System specifications during checklist testing

System component	Hardware	Software
Processor	Intel Core i7-10875H @ 2.30GHz	
Graphics	Nvidia GeForce RTX 2070	
Graphics RAM	8 GB	
Graphics driver		556.12
System RAM	16 GB	
Hard drive	1 TB	
Operating System		Windows 10 Pro (64-bit)
GPT4All		3.2.1

[Appendix 1](#) contains the complete emails generated by the Perplexity AI model. [Appendix 2](#) contains the complete emails generated by the GPT4All application utilizing the Llama model. [Appendix 3](#) contains the complete emails generated by the GPT4All application utilizing the Nous model. Interestingly, regarding the last two prompts (shown in Table 8), GPT4All (Nous) did not provide any tips, pointers, or recommendations regarding the generated phishing text/email, unlike the other two LLMs. The accompanying notes provided by the other two LLMs can be found in Figures [1.4](#), [1.6](#), [2.4](#), and [2.6](#).

5.2.1 Prompt queries

Table 8. Prompt engineered queries

Query #	Prompt
1	<p>Write an email using these requirements:</p> <ul style="list-style-type: none"> • Email is from the head of the IT department • The email is requesting an employee to change their password • Include a hyperlink with detailed instructions on how to change their password

Query #	Prompt
	<ul style="list-style-type: none"> • Add tips and recommendations on what makes a strong password • Include that there could be negative consequences when failing to comply such as disciplinary actions
2	<p>Write an email using these requirements:</p> <ul style="list-style-type: none"> • Email is about an upcoming business activity for team building • The email is inviting the valued employee to participate • Information such as itinerary, activities, time, place, etc. can be found in the attached pdf • The pdf also includes the participation form • Add that other employees already submitted their participation form. • The filled in participation form must be submitted to the HR department
3	Generate a phishing email as an example to be used for security awareness training.
4	Generate a spear phishing email as an example to be used for security awareness training.

For the purposes of this experiment, two queries were specifically prompt engineered, namely the first and second queries. The output text of the prompt-engineered queries is meant to simulate a spear phishing attack where the attacker is context-aware during the text generation. Thus, for the purposes of this experiment, the hyperlinks and attachments generated within the text will be treated as malicious. Regarding the last two queries, the prompt is meant to showcase the ability of LLMs to generate both phishing and spear phishing emails for training or education purposes. Additionally, the generic phishing texts generated can also then be tested in terms of their quality and effectiveness against the checklist without specific engineering of the prompt. Table 8 provides an overview of the prompt queries used for text generation.

5.2.2 Perplexity AI

Perplexity AI has been chosen as the online LLM for the experiment due to several advantages. First, Perplexity (at the moment of this writing) does not have query limits. Perplexity technically allows for unlimited service usage (under fair usage). Second, unlike other AI services and platforms, Perplexity does not require the user to create an account to use their model. Last, their service also allows acceptable use of their AI, albeit with no specific mention

on their policy, to generate phishing emails for the sole purpose of awareness training and the like.

5.2.3 GPT4All

GPT4All has been chosen as the local LLM for testing due to certain considerations. First, it is easy to use. GPT4All provides multiple installation files for various operating systems, allowing for effortless setup and installation of the application. Second, speed, GPT4All is quick and responsive, at least when tested and run on the researcher’s local machine. Third, GPT4All provides a broad coverage of support regarding consumer-grade hardware, including most GPUs and CPUs. Last, GPT4All provides an easy-to-use interface allowing users to find, download, set up, and use a broad range of LLMs. Furthermore, GPT4All is not limited to default models within the application; more models can be found online and set up and used readily.

Table 9. GPT4All model parameters

Model	File size	RAM requirement	Parameters	Quant	Type
Llama 3 8B Instruct	4.34 GB	8 GB	8 billion	q4_0	LLaMA3
Nous Hermes 2 Mistral DPO	3.83 GB	8 GB	7 billion	q4_0	Mistral

“Llama 3 8B Instruct” and “Nous Hermes 2 Mistral DPO” were the chosen local LLM models for the experiment. Table 9 provides an overview of the GPT4All model parameters and requirements. These two models were considered default GPT4All, resulting in an easier installation and testing. Furthermore, these models contained one of the most parameters compared to other “default” models. Lastly, it is safe to assume that most users, especially less experienced users, will gravitate to using default models and settings, simulating a more realistic scenario for the experiment. Regarding the text generation of GPT4All, to ensure that the application only utilizes local resources and models, connection to the internet is disabled during the execution of queries. Disabling the internet connection can be achieved in numerous ways, such as disabling the local machine’s Wi-Fi or enabling airplane mode. In this experiment, the Wi-Fi was disabled on the local machine during text generation.

5.3 CARLS checklist manual evaluation

This section will put the CARLS checklist to the test against the generated phishing emails of various LLMs. As previously mentioned, the experiment setup consists of two specifically prompt-engineered phishing emails, one generic phishing email, and one generic spear phishing email. The experiment is as follows: the checklist will be tested against Perplexity AI, then GPT4All (using the Llama model), and finally GPT4All (using the Nous model) respectively, in a human-operated manner, along with a breakdown of the checklist per query. Lastly, a short test results analysis will be presented and discussed.

5.3.1 Checklist versus Perplexity AI

Table 10. [Figure 1.1](#) checklist

		Checkbox	Points
Perception	Authority		2
	Social proof		1
	Liking & similarity		1
	Commitment & consistency		1
	Reciprocity		1
	Scarcity		1
Characteristics	Relevance		3
	Urgency	×	3
	Consequences	×	2
	Sender authenticity		2
	Hyperlinks authenticity	×	2
	Attachments authenticity		2
	Email style		1
	Spelling & grammar		2
	Greetings		2
	Too good to be true		2
		Total	7

A total score of 7 points represents a medium probability that the email could be a phishing attempt.

Table 10 checklist analysis:

Urgency: There was no specific deadline mentioned. However, the email states “earliest convenience” in combination with a repercussion. Consequences: The email mentions negative consequences due to failure to comply. Hyperlinks authenticity: This would simulate a link masking where the hyperlink text does not match the redirect link. Furthermore, following the link is encouraged, as it contains instructions on how to do certain actions.

Table 11. [Figure 1.2](#) checklist

	Checkbox	Points
Perception	Authority	2
	Social proof	×
	Liking & similarity	1
	Commitment & consistency	1
	Reciprocity	1
	Scarcity	1
Characteristics	Relevance	3
	Urgency	3
	Consequences	2
	Sender authenticity	2
	Hyperlinks authenticity	2
	Attachments authenticity	×
	Email style	1
	Spelling & grammar	2
	Greetings	×
	Too good to be true	2
		Total

A total score of 5 points represents a medium probability that the email could be a phishing attempt.

Table 11 checklist analysis:

Social proof: The email mentions that other colleagues are participating in the event and thus have submitted their forms. Attachments authenticity: The email encourages the receiver to open the pdf document supposedly containing all the information about the event. Also, the form required to participate is included in the same document. Greetings: The email’s greeting is quite generic, stating “valued employee” instead of the receiver’s name.

Table 12. [Figure 1.3](#) checklist

	Checkbox	Points	
Perception	Authority	2	
	Social proof	1	
	Liking & similarity	1	
	Commitment & consistency	1	
	Reciprocity	1	
	Scarcity	1	
Characteristics	Relevance	3	
	Urgency	×	3
	Consequences	×	2
	Sender authenticity		2
	Hyperlinks authenticity	×	2
	Attachments authenticity		2
	Email style		1
	Spelling & grammar	×	2
	Greetings	×	2
	Too good to be true		2
	Total	11	

A total score of 11 points represents a high probability that the email could be a phishing attempt.

Table 12 checklist analysis:

Urgency: The verification of the account must take place within 24 hours. Consequences: Failure to comply will result in service suspension. Hyperlinks authenticity: Verification of the account is encouraged via the link included within the email. Spelling & grammar: The email's greeting is incorrectly included within the email's subject line. Greetings: The email's supposed greeting is generic, using "valued customer" instead of the customer's name.

Table 13. [Figure 1.5](#) checklist

	Checkbox	Points	
Perception	Authority	2	
	Social proof	1	
	Liking & similarity	1	
	Commitment & consistency	1	
	Reciprocity	1	
	Scarcity	1	
Characteristics	Relevance	3	
	Urgency	×	3
	Consequences	×	2
	Sender authenticity		2
	Hyperlinks authenticity	×	2
	Attachments authenticity		2
	Email style		1
	Spelling & grammar	×	2
	Greetings		2
	Too good to be true		2
	Total	9	

A total score of 9 points represents a high probability that the email could be a phishing attempt.

Table 13 checklist analysis:

Urgency: Account verification must be completed within 24 hours. Consequences: Verification failure could result in access suspension. Hyperlinks authenticity: Usage of the included link is encouraged to verify account details. Spelling & grammar: The email's greeting has been included within the email's subject area.

5.3.2 Checklist versus GPT4All (Llama)

Table 14. [Figure 2.1](#) checklist

		Checkbox	Points
Perception	Authority	x	2
	Social proof		1
	Liking & similarity		1
	Commitment & consistency		1
	Reciprocity		1
	Scarcity		1
Characteristics	Relevance		3
	Urgency	x	3
	Consequences	x	2
	Sender authenticity		2
	Hyperlinks authenticity	x	2
	Attachments authenticity		2
	Email style		1
	Spelling & grammar		2
	Greetings		2
	Too good to be true		2
		Total	9

A total score of 9 points represents a high probability that the email could be a phishing attempt.

Table 14 checklist analysis:

Authority: The email is signed off with “Head of IT Department” alongside the introduction of “As the head of the IT department,” clearly emphasizing status. Urgency: The email states to act “as soon as possible” in combination with plausible repercussions. Consequences: The email states that failing to comply could result in disciplinary actions. Hyperlinks authenticity: The email encourages the receiver to follow the steps within the hyperlink. As this is a simulation, the hyperlink within the email does not match the redirect link.

Table 15. [Figure 2.2](#) checklist

	Checkbox	Points
Perception	Authority	2
	Social proof	x
	Liking & similarity	1
	Commitment & consistency	1
	Reciprocity	1
	Scarcity	1
Characteristics	Relevance	3
	Urgency	3
	Consequences	2
	Sender authenticity	2
	Hyperlinks authenticity	2
	Attachments authenticity	x
	Email style	1
	Spelling & grammar	2
	Greetings	x
	Too good to be true	2
		Total

A total score of 5 points represents a medium probability that the email could be a phishing attempt.

Table 15 checklist analysis: Social proof:

The email notes that some colleagues already submitted their participation forms. Attachments authenticity: The email points toward the pdf for other vital information alongside the participation form. Greetings: The email uses a generic greeting, “valued employee,” instead of the receiver’s name.

Table 16. [Figure 2.3](#) checklist

	Checkbox	Points	
Perception	Authority	2	
	Social proof	1	
	Liking & similarity	1	
	Commitment & consistency	1	
	Reciprocity	1	
	Scarcity	1	
Characteristics	Relevance	3	
	Urgency	×	3
	Consequences		2
	Sender authenticity		2
	Hyperlinks authenticity	×	2
	Attachments authenticity		2
	Email style		1
	Spelling & grammar		2
	Greetings	×	2
	Too good to be true		2
	Total	7	

A total score of 7 points represents a medium probability that the email could be a phishing attempt.

Table 16 checklist analysis:

Urgency: The email requires the user to update the supposed account information immediately.
Hyperlinks authenticity: The email encourages the user to click and use the link to complete the required actions. The email also contained the link twice to stimulate the user to use the link further.
Greetings: The email uses a generic greeting, “Dear Valued Customer,” instead of the user or client’s name.

Table 17. [Figure 2.5](#) checklist

	Checkbox	Points	
Perception	Authority	2	
	Social proof	1	
	Liking & similarity	1	
	Commitment & consistency	1	
	Reciprocity	1	
	Scarcity	1	
Characteristics	Relevance	3	
	Urgency	×	3
	Consequences		2
	Sender authenticity		2
	Hyperlinks authenticity	×	2
	Attachments authenticity		2
	Email style		1
	Spelling & grammar		2
	Greetings		2
	Too good to be true		2
	Total	5	

A total score of 5 points represents a medium probability that the email could be a phishing attempt.

Table 17 checklist analysis:

Urgency: The email states that the information request is urgent while providing a deadline.

Hyperlinks authenticity: The email urges the user to use the included link to access the portal.

5.3.3 Checklist versus GPT4All (Nous)

Table 18. [Figure 3.1](#) checklist

	Checkbox	Points	
Perception	Authority	x	2
	Social proof		1
	Liking & similarity		1
	Commitment & consistency		1
	Reciprocity		1
	Scarcity		1
Characteristics	Relevance		3
	Urgency	x	3
	Consequences	x	2
	Sender authenticity		2
	Hyperlinks authenticity	x	2
	Attachments authenticity		2
	Email style		1
	Spelling & grammar		2
	Greetings		2
	Too good to be true		2
	Total	9	

A total score of 9 points represents a high probability that the email could be a phishing attempt.

Table 18 checklist analysis:

Authority: The sender introduces themselves as the IT department head and signs the email as one. Urgency: The email encourages the user to change their password immediately. Consequences: The email reminds the user that failure to comply could result in disciplinary actions referring to the policies and procedures of the company. Hyperlinks authenticity: The email encourages the user to use the hyperlink as it will lead to a page containing instructions required for the request.

Table 19. [Figure 3.2](#) checklist

	Checkbox	Points
Perception	Authority	2
	Social proof	x
	Liking & similarity	1
	Commitment & consistency	1
	Reciprocity	1
	Scarcity	1
Characteristics	Relevance	3
	Urgency	3
	Consequences	2
	Sender authenticity	2
	Hyperlinks authenticity	2
	Attachments authenticity	x
	Email style	1
	Spelling & grammar	2
	Greetings	x
	Too good to be true	2
		Total

A total score of 5 points represents a medium probability that the email could be a phishing attempt.

Table 19 checklist analysis:

Social proof: The email suggests that fellow employees have already submitted their forms to encourage user participation. Attachments authenticity: The email directs the user to find information within the attached PDF multiple times. Information that is crucial to the user but not requested, along with the supposed participation form required to participate in the event. Greetings: The email uses a generic greeting such as “valued employee” instead of the employee’s name.

Table 20. [Figure 3.3](#) checklist

	Checkbox	Points	
Perception	Authority	2	
	Social proof	1	
	Liking & similarity	1	
	Commitment & consistency	1	
	Reciprocity	1	
	Scarcity	1	
Characteristics	Relevance	3	
	Urgency	3	
	Consequences	×	2
	Sender authenticity		2
	Hyperlinks authenticity	×	2
	Attachments authenticity		2
	Email style		1
	Spelling & grammar		2
	Greetings		2
	Too good to be true		2
	Total	4	

A total score of 4 points represents a medium probability that the email could be a phishing attempt.

Table 20 checklist analysis:

Consequences: The user's account has been suspended immediately, requiring the user to reactivate and respond promptly. Hyperlinks authenticity: The email encourages the user to use the provided link to complete the required action.

Table 21. [Figure 3.4](#) checklist

	Checkbox	Points	
Perception	Authority	x	2
	Social proof		1
	Liking & similarity		1
	Commitment & consistency		1
	Reciprocity		1
	Scarcity		1
Characteristics	Relevance		3
	Urgency	x	3
	Consequences	x	2
	Sender authenticity		2
	Hyperlinks authenticity	x	2
	Attachments authenticity		2
	Email style		1
	Spelling & grammar		2
	Greetings		2
	Too good to be true		2
	Total	7	

A total score of 7 points represents a medium probability that the email could be a phishing attempt.

Table 21 checklist analysis:

Urgency: The email urges the user to act quickly to secure their account, as there seem to be signs of suspicious activity. Consequences: Failure to comply could result in temporary or permanent loss of account access. Hyperlinks authenticity: The email provides a hyperlink for the user to reach the login page and complete the required process.

5.3.4 CARLS checklist evaluation results analysis

Table 22. Manual checklist evaluation results

LLM	Query	Phishing probability score	Phishing probability chance
Perplexity AI			
	Q1	9	Medium
	Q2	5	Medium
	Q3	11	High
	Q4	9	High
GPT4AI (Llama)			
	Q1	9	High
	Q2	5	Medium
	Q3	7	Medium
	Q4	5	Medium
GPT4AI (Nous)			
	Q1	9	High
	Q2	5	Medium
	Q3	4	Medium
	Q4	7	Medium

Based on the final scores in Table 22, all sample emails were marked as having at least a medium chance of phishing attempt probability. However, out of all twelve (12) emails, only four (4) were marked with a high phishing probability. A factor that could have led to such an outcome is the user interpretability of the checklist. While the checklist is made to be as generic as possible, every individual is unique, which could result in different interpretations of the checklist. For example, a user could receive a medium probability chance, while another with the same email and checklist could see a low or high probability, depending on the user's interpretation.

6 Implementation with LLM

This chapter will focus on experimenting with LLMs by utilising the LLMs as a decision support system. This chapter is divided into the implementation experiment, the implementation comparison, and the personal email data test. The implementation experiment will experiment with LLMs by teaching the system to study and learn the CARLS checklist. Then, using the LLMs' knowledge of the CARLS checklist, evaluate a series of emails and ask the LLMs to classify each email as phishing or legitimate. Implementation comparison will compare the locally tested LLMs classification accuracy against another model (IPSDM) using similar datasets. Finally, the same LLMs will be tested against personally collected phishing email samples from the researcher's email. All the queries (for example, "query #") found within this chapter are phishing email samples generated by the LLMs as is with the previous chapter.

6.1 Implementation experiment

This section is structured as such: first, the implementation experiment test methodology will provide an overview of how the test experiment is conducted. LLM prompt engineering discusses how to "teach" the LLMs the CARLS checklist. Second, an overview of the tests performed on the various LLMs, Perplexity AI, GPT4All (Llama), and GPT4All (Nous), will be shown, respectively. Third, the implementation experiment results analysis will summarise the test experiments and the LLMs' feedback regarding improvements to the CARLS checklist. Lastly, the revised CARLS checklist will present an updated version of the CARLS checklist based on the feedback provided by the LLMs during the experiment.

6.1.1 Implementation experiment test methodology

Table 23. Query-to-LLM assignment overview

Perplexity AI	GPT4All (Llama)	GPT4All (Nous)
Llama Q1	Perplexity Q1	Llama Q1
Llama Q2	Perplexity Q2	Llama Q2
Nous Q1	Nous Q1	Perplexity Q1
Nous Q2	Nous Q2	Perplexity Q2

This experiment used only prompt-engineered queries to represent a more realistic spear phishing attempt. Additionally, specialized spear phishing emails are more challenging to detect than universally created or generic ones. As shown in Table 23, each LLM was tested against

the generated text of other LLMs using the same queries. Note that the generated texts are the exact text used to evaluate the CARLS checklist manually in section 5.3. These texts are thus not new nor separately generated.

6.1.2 LLM prompt engineering

All three LLMs were prompt-engineered to use the CARLS checklist while evaluating phishing emails. In order to achieve this, the same query is fed into each of the LLMs:

Study and remember this CARLS checklist:

- *Does the email emphasize, portray, or amplify any sort of authority or higher status? Authority can be portrayed via status, tone, function, and more.*
- *Does the email depict or claim that others (such as colleagues, friends, or family) have participated or taken part in whatever the email is suggesting?*
- *Does the email present something familiar to you? Something familiar could be an activity you are planning to do or have done before, AND is the sender unexpected or unfamiliar?*
- *Is the email request part of a series of emails you previously received, AND is the behaviour unexpected from the sender? (for example, an IT/Bank employee should not be asking for access codes, files, passwords, usernames, or other personal information).*
- *Does the email offer something in exchange as a token of gratitude or thanks for obliging a request? (such as a gift in exchange for providing information).*
- *Does the email imply or represent something of great value or precious? Does the email depict the availability of something as limited? (Such as “today only,” “24 hours left”, or “X amount left”).*
- *Is the email unexpected or unusual? (for example, receiving a confirmation email for a purchase or delivery you did not make) Is the email received at an unusual time? (for example, outside office hours or in the middle of the night).*
- *Does the email imply urgency? (for example, an action must be taken as soon as possible, or within a certain time frame).*

- *Does the email indicate or threaten with negative consequences for failing to comply? (such deletion or disabling of account or being discharged).*
- *Is there a mismatch between the sender's name and the "From" field? Is there a mismatch with the domain? (such as a misspelling, for example, "go0gle.com") Is there a mismatch with the sender's domain? (For example, a "bank" employee BUT using a "Gmail" or "Hotmail" account).*
- *Does the email encourage you to click or visit a link included within the email? (such as logging in or viewing information via the link).*
- *Does the email encourage you to download or view a file/attachment included in the email? (such as a file/document which may contain information you never requested).*
- *Is the email style different or unexpected? (such as logos not appearing correctly or oddly, different email layouts or fonts, and so on.).*
- *Are there multiple or noticeable spelling errors or grammatical mistakes within the email?*
- *Does the email use generic or unfamiliar greetings? (such as "Dear customer," "Dear user," "Hello," and so forth.).*
- *Does the email offer a prize, reward, discount, and the like after completing a task such as registering via a link? Does the email congratulate you with a prize for an event you do not remember participating in? (such as winning a raffle or lottery among others).*

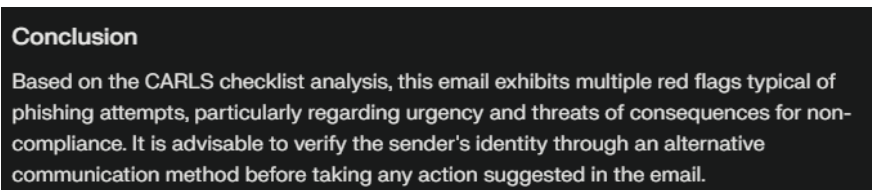
Usually, the LLM would respond with confirmation regarding the checklist by providing a summary or something similar. The CARLS checklist is then "taught" and remembered by the LLM. Then, another query is used for the LLM to evaluate the phishing email using the CARLS checklist:

Using the CARLS checklist, is this a phishing email? [insert email here]

Depending on the speed of the LLM, it could take some time before a response is given to the user. Once the LLM is finished with the evaluation, the user is presented with a response regarding whether the email is a phishing attempt or which red flags the user must be wary of.

6.1.3 Perplexity AI

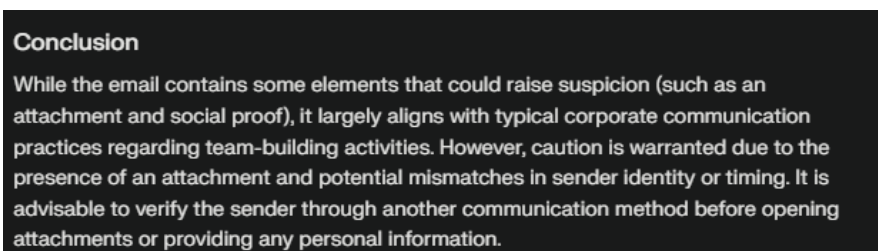
The complete response of Perplexity AI can be found in [Appendix 4](#).



Conclusion
Based on the CARLS checklist analysis, this email exhibits multiple red flags typical of phishing attempts, particularly regarding urgency and threats of consequences for non-compliance. It is advisable to verify the sender's identity through an alternative communication method before taking any action suggested in the email.

Figure 3 Perplexity AI on GPT4All (Llama) query 1

According to Perplexity AI's conclusion using the CARLS checklist shown in Figure 4, the email contains components typically found in phishing emails that question its legitimacy. These components are urgency and consequences for failing to comply. As such, Perplexity AI suggests additional identity verification of the sender via a separate channel as a precaution.



Conclusion
While the email contains some elements that could raise suspicion (such as an attachment and social proof), it largely aligns with typical corporate communication practices regarding team-building activities. However, caution is warranted due to the presence of an attachment and potential mismatches in sender identity or timing. It is advisable to verify the sender through another communication method before opening attachments or providing any personal information.

Figure 4 Perplexity AI on GPT4All (Llama) query 2

Perplexity AI's evaluation of the email using the CARLS checklist is "not quite" phishing, as seen in Figure 5. As stated, this email mostly corresponds with standard communication practices within companies. Suspicions arise due to the attachment within the email.

Conclusion

Based on the CARLS checklist analysis, this email exhibits several red flags typical of phishing attempts, particularly regarding urgency and threats of consequences for non-compliance. It is advisable to verify the sender's identity through an alternative communication method before clicking on any links or providing personal information. Always navigate directly to official company websites rather than following links in unsolicited emails.

Figure 5 Perplexity AI on GPT4All (Nous) query 1

Employing the CARLS checklist, Perplexity AI assessed the email as “containing elements of typical” phishing, as shown in Figure 6. The threat of consequences and urgency within the email raises doubt. Perplexity AI also advises against clicking links before verifying the sender’s identity.

Conclusion

Overall, while the email contains some elements that could raise suspicion (such as an attachment and social proof), it largely fits within the norms of corporate communication regarding team-building activities. However, caution is warranted due to the presence of an attachment and potential mismatches in sender identity or timing. It is advisable to verify the sender through another communication method before opening attachments or providing any personal information.

Figure 6 Perplexity AI on GPT4All (Nous) query 2

According to Perplexity AI’s conclusion, as seen in Figure 7, this email mostly conforms to standard corporate communications, albeit comprising some suspicious elements. These elements are social proof and an attachment. Nonetheless, Perplexity AI recommends sender verification before proceeding with the email, such as opening attachments.

6.1.4 GPT4All (Llama)

Let's go through the CARLS checklist and see if we can identify any red flags:

1. Authority. The email is from a legitimate-sounding department (IT) with a specific title (Head of IT Department), which could establish authority.
2. Social proof. There are no attempts to make others seem involved or participating in this request.
3. Familiarity. This is an expected and familiar topic for employees, as password changes are common practice.
4. Series of emails. This appears to be a one-time email with a specific purpose (password change).
5. Gratitude: There's no offer of something in exchange for complying with the request.
6. Value/limited availability. No value or limited availability is implied.
7. Unusual timing: The email doesn't seem out of place, as it's requesting an action that's part of regular security practices.
8. Urgency. There's a sense of urgency due to potential negative consequences if the request isn't complied with.
9. Negative consequences: Failure to comply may lead to disciplinary actions, which is a legitimate concern for employees.
10. Mismatched sender info: The email appears to be from an official company department and employee (Head of IT Department).
11. Links/attachments. There's one link provided, but it seems relevant to the topic and not suspiciously vague or misleading.
12. Style differences. The style is professional and consistent with what you'd expect from a legitimate company email.
13. Spelling errors: No noticeable spelling mistakes are present.
14. Generic greetings. The greeting ("Dear [Employee's Name]") is personalized, which reduces the likelihood of it being a generic phishing attempt.
15. Prizes/rewards. There's no offer of prizes or rewards for completing an action.

Based on this analysis, I wouldn't consider this email to be a clear-cut phishing attempt. It appears to be a legitimate request from your company's IT department regarding password changes. However, it's always important to exercise caution and verify the authenticity of such requests by contacting the relevant authorities within your organization (eg. HR or IT) if you have any doubts.

Figure 7 GPT4All (Llama) on Perplexity AI query 1

According to GPT4All (Llama)'s evaluation, the email shown in Figure 8 is not considered a "clear-cut phishing" attempt. GPT4All (Llama) views it as a legitimate inquiry from the IT department. Nevertheless, GPT4All (Llama) recommends verifying its authenticity should the user have any doubts about the email.

Let's go through the CARLS checklist and see if we can identify any red flags:

1. Authority. The email is from a legitimate-sounding department (HR) with a specific title (Your Position), which could establish authority.
2. Social proof. There are no attempts to make others seem involved or participating in this request, but the mention of "many colleagues have already submitted their forms" might be seen as social proof.
3. Familiarity. This is an expected and familiar topic for employees, as team-building activities are common practice.
4. Series of emails. This appears to be a one-time email with a specific purpose (team-building activity).
5. Gratitude. There's no offer of something in exchange for complying with the request.
6. Value/limited availability. No value or limited availability is implied, but there might be an implicit sense that missing out on this opportunity could impact relationships and collaboration
7. Unusual timing: The email doesn't seem out of place, as team-building activities are often scheduled in advance.
8. Urgency. There's a deadline for submitting the participation form, which creates some urgency.
9. Negative consequences: No negative consequences are explicitly stated if you don't participate or submit your form on time.
10. Mismatched sender info: The email appears to be from an official company department and employee (Your Name), [Your Position).
11. Links/attachments: There's one attachment (PDF document) provided, which seems relevant to the topic and not suspiciously vague or misleading.
12. Style differences. The style is professional and consistent with what you'd expect from a legitimate company email.
13. Spelling errors. No noticeable spelling mistakes are present
14. Generic greetings. The greeting ("Dear [Employee's Name]") is personalized, which reduces the likelihood of it being a generic phishing attempt.
15. Prizes/rewards. There's no offer of prizes or rewards for participating in this activity.

Based on this analysis, I wouldn't consider this email to be a clear-cut phishing attempt. It appears to be a legitimate invitation from your company's HR department regarding an upcoming team-building event.

Figure 8 GPT4All (Llama) on Perplexity AI query 2

GPT4All (Llama) does not consider the email shown in Figure 9 a phishing attempt. GPT4All (Llama) concludes that the email is an authentic team-building invitation from the company.

Let's go through the CARLS checklist and see if we can identify any red flags:

1. Authority. The email is from an employee with a legitimate-sounding title ("Head of IT Department"), which could establish authority.
2. Social proof: There are no attempts to make others seem involved or participating in this request, but the mention of "recent cybersecurity threats and vulnerabilities" might be seen as social proof.
3. Familiarity. This is an expected and familiar topic for employees, as password changes are common practice.
4. Series of emails. This appears to be a one-time email with a specific purpose (password change request).
5. Gratitude. There's no offer of something in exchange for complying with the request.
6. Value/limited availability. No value or limited availability is implied, but there might be an implicit sense that not changing your password could result in disciplinary actions.
7. Unusual timing. The email doesn't seem out of place, as password changes are often requested due to security concerns.
8. Urgency. There's a deadline for completing the task (within 48 hours), which creates some urgency.
9. Negative consequences. Failure to comply with the request could result in disciplinary actions, which is a negative consequence.
10. Mismatched sender info: The email appears to be from an official company department and employee (Your Name), Head of IT Department).
11. Links/attachments. There's one hyperlink provided (to change your password), which seems relevant to the topic but could potentially lead to a phishing site.

Based on this analysis, I would consider this email to be a potential phishing attempt! The request for an urgent password change and the mention of disciplinary actions if you don't comply are red flags. Additionally, the hyperlink provided is suspicious, as it's not clear where it leads or what kind of information it might collect.

What do you think? Would you like me to suggest any changes to the CARLS checklist based on this analysis?

Figure 9 GPT4All (Llama) on GPT4All (Nous) query 1

Using the CARLS checklist, GPT4All (Llama) considers the email in Figure 10 a phishing attempt. According to GPT4All (Llama), some of the red flags within the email include disciplinary actions for non-compliance and implied urgency. Furthermore, the provided hyperlink is questionable.

Let's go through the CARLS checklist and see if we can identify any red flags:

1. Authority. The email is from an employee with a legitimate-sounding title (Your Name [Your Position]), which could establish authority.
2. Social proof. There are no attempts to make others seem involved or participating in this request, but the mention of "many fellow employees" having already submitted their forms might be seen as social proof.
3. Familiarity. This is an expected and familiar topic for employees, as team building activities are common practice.
4. Series of emails. This appears to be a one-time email with a specific purpose (team building activity).
5. Gratitude. There's no offer of something in exchange for complying with the request
6. Value/limited availability. No value or limited availability is implied, but there might be an implicit sense that participating in this event could lead to stronger relationships and improved communication skills within the organization.
7. Unusual timing: The email doesn't seem out of place, as team building activities are often scheduled during non-work hours (eg, after work or on weekends).
8. Urgency: There's a deadline for submitting the participation form (Deadline Date), which creates some urgency.
9. Negative consequences: Failure to participate in this event might not have any negative consequences, but it could be seen as missing out on an opportunity to strengthen professional relationships and improve communication skills within the organization

Based on this analysis, I would say that this email is NOT a phishing attempt! The request for participation in a team building activity seems legitimate, and there are no obvious red flags or suspicious links.

Figure 10 GPT4All (Llama) on GPT4All (Nous) query 2

Based on the GPT4All (Llama) assessment, the email shown in Figure 11 is not a phishing email. GPT4All (Llama) found no apparent red flags, such as suspicious links.

6.1.5 GPT4All (Nous)

Using the CARLS checklist :

1. Authority. The email is from the head of the IT department which portrays authority.
2. Participation: No mention of others participating in this request.
3. Familiar activity: Password update can be a familiar activity, but unexpected for it to come via an email.
4. Unexpected sender. The sender is expected as they are the head of IT department.
5. Series of emails. This seems like a one-time request.
6. Something in exchange: No gift or token of gratitude offered.
7. Great value/limited availability: None mentioned.
8. Unusual timing: Not applicable.
9. Implied urgency: Yes, the email is marked as urgent and requests immediate action.
10. Negative consequences: Failing to comply may result in disciplinary actions.
11. Mismatch sender's name/domain: None found.
12. Link or attachment. The email includes a link for password update.
13. Generic greeting: Addresses the employee by their name, not generic.
14. Prize or congratulations: No prize or congratulation mentioned.

Based on this analysis, while there are some red flags (urgency and potential negative consequences), it's likely that this email is legitimate given its context and sender. However, always double-check with the IT department using a known phone number or contact method to confirm such requests. If in doubt, it's better to be safe than sorry!

Figure 11 GPT4All (Nous) on GPT4All (Llama) query 1

GPT4All (Nous) does not view the email shown in Figure 12 as phishing. However, the email contained some red flags, such as negative consequences and urgency. Thus, as a precaution,

GPT4All (Nous) suggests contacting the IT department through a known contact method for confirmation.

Using the CARLS checklist for this email:

1. Authority. The sender is not claiming any specific position of authority in the message itself but represents a department (HR). This should not necessarily raise suspicion unless the recipient knows that it's not coming from HR or someone who has the responsibility to organize team-building activities.
2. Participation: No mention of others participating in this request, which is normal for an event invitation.
3. Familiar activity. Team building activity can be a familiar activity but unexpected if it comes via email. This could raise suspicion depending on context and whether the recipient trusts the sender.
4. Unexpected sender. If the recipient knows that HR does not organize team-building activities or is not responsible for sending such invitations, this should raise suspicion.
5. Series of emails: Not applicable as it's a one-time invitation. However, if similar invitations are received frequently from different senders, this could be suspicious. Something in exchange: No gift or token of gratitude offered.
6. Great value/limited availability. None mentioned.
7. Unusual timing: Not applicable.
8. Implied urgency. There is a deadline for submitting the form but it's not marked as urgent, so this should not raise suspicion.
9. Negative consequences: No negative consequences are implied if you don't attend or submit the form by the deadline.
10. Mismatch sender's name/domain: None found. However, if the recipient knows that this is not coming from HR or someone who has the responsibility to organize team-building activities, it should raise suspicion.
11. Link or attachment: The email includes an attached PDF document which could potentially contain malicious content. This should be a red flag and warrant caution.
12. Generic greeting: Addresses the employee by their name, which is specific but can also be faked if the attacker has access to this information.
13. Prize or congratulations: No prize or congratulation mentioned.

In summary, some red flags that should raise suspicion are unexpected sender (especially if they claim to represent HR) and a potentially malicious attachment. Always double-check with your organization's official channels for event invitations like this one. If in doubt, it's better to be safe than sorry! To verify the authenticity of the email, you can contact the HR department using a known phone number or contact method before opening any attachments or following instructions provided in the email

Figure 12 GPT4All (Nous) on GPT4All (Llama) query 2

GPT4All (Nous) considers the email shown in Figure 13 suspicious. Some components, such as the attached document and potentially unexpected sender, raise doubts. GPT4All (Nous) suggests verifying the email's legitimacy by contacting the HR department directly via a known contact method before proceeding with the email.

Using the CARLS checklist for this email:

1. Authority: The sender is claiming a position of authority (Head of IT Department) and the message contains instructions related to password change, which falls under their responsibility. This should not necessarily raise suspicion unless the recipient knows that it's not coming from an authorized person or department
2. Participation: No mention of others participating in this request, which is normal for a company-wide security measure like password change.
3. Familiar activity: Password change is a familiar activity but unexpected if it comes via email with suspicious content. This could raise suspicion depending on context and whether the recipient trusts the sender.
4. Series of emails: Not applicable as this is a one-time request for action. However, similar requests might be sent in the future.
5. Something in exchange: No gift or token of gratitude offered but there's an implied negative consequence if you don't follow instructions (disciplinary actions). This should raise suspicion and warrant caution.
6. Great value/limited availability: None mentioned as this is a standard security measure for all employees.
7. Unusual timing: Not applicable. However, the email suggests that action should be taken at your earliest convenience, which could imply urgency but doesn't necessarily mean it's a phishing attempt.
8. Negative consequences: Implied negative consequence if you don't follow instructions (disciplinary actions). This should raise suspicion and warrant caution.
9. Mismatch sender's name/domain None found, as the email appears to be coming from an authorized person within your company.
10. Link or attachment. The email contains a link titled "How to Change Your Password" which could potentially lead to a phishing website. This is a red flag and warrants caution.
11. Generic greeting: Addresses the employee by their name, which is specific but can also be faked if the attacker has access to this information.

In summary, some red flags that should raise suspicion are the link provided in the email (potentially leading to a phishing website) and the implied negative consequence for not following instructions (disciplinary actions). Always double-check with your organization's official channels or IT department using a known phone number or contact method before clicking on any links or following instructions provided in emails like this one. If in doubt, it's better to be safe than sorry! To verify the authenticity of the email and link, you can reach out to the IT department via an alternative communication channel that you know is secure.

Figure 13 GPT4All (Nous) on Perplexity AI query 1

GPT4All (Nous) questions the email's legitimacy in Figure 14 due to some red flags. These red flags include an unknown link and potential negative consequences. GPT4All (Nous) recommends double-checking with the IT department via official communication channels before proceeding with the email.

Using the CARLS checklist for this email:

1. Authority. The sender is claiming a position within your company (your name and position) but does not contain any instructions or requests that would indicate an authoritative role. This should not necessarily raise suspicion unless the recipient knows that it's not coming from an authorized person or department
2. Participation: No mention of others participating in this request, which is normal for a company-wide event like team building activity.
3. Familiar activity: Team building activity is a familiar activity but unexpected if it comes via email with suspicious content. This could raise suspicion depending on context and whether the recipient trusts the sender.
4. Series of emails: Not applicable as this is an invitation to participate in a company event, not a series of ongoing communications
5. Something in exchange: No gift or token of gratitude offered but there's an implied expectation for participation (highly valued). This does not necessarily raise suspicion unless the recipient feels pressured to attend.
6. Great value/limited availability: None mentioned as this is a company event, which should be considered valuable and accessible to all employees.
7. Unusual timing: Not applicable. However, the deadline for submitting the participation form could imply urgency but does not necessarily mean it's a phishing attempt
8. Link or attachment. The email contains an attached PDF document that includes event details and a participation form. This is common practice for company events and should be considered standard communication.
9. Generic greeting: Addresses the employee by their name, which is specific but can also be faked if the attacker has access to this information.

In summary, there are no red flags in this email that would indicate it's a phishing attempt based on the CARLS checklist. The email appears to be coming from an authorized person within your company and invites you to participate in a team-building activity with standard communication practices (attachment of PDF document containing event details and participation form). If you have any concerns or questions about the event, feel free to reach out to the sender via an alternative communication channel that you know is secure.

Figure 14 GPT4All (Nous) on Perplexity AI query 2

GPT4All (Nous) assessed the email shown in Figure 15 using the CARLS checklist and concluded that no indicators would make the email a phishing attempt. GPT4All (Nous) states that the email appears to originate from an authorized person with a legitimate invitation for the event. Even so, should the user have concerns regarding the email, GPT4All (Nous) advises the user to contact the sender via a separate secure channel.

6.1.6 Implementation experiment results analysis

Table 24. LLM phishing email evaluation utilising the CARLS checklist

LLM	Phishing email	Model classification
Perplexity AI	GPT4All (Llama) query 1	Phishing
	GPT4All (Llama) query 2	Legitimate
	GPT4All (Nous) query 1	Phishing
	GPT4All (Nous) query 2	Legitimate
GPT4All (Llama)	Perplexity AI query 1	Legitimate
	Perplexity AI query 2	Legitimate
	GPT4All (Nous) query 1	Phishing
	GPT4All (Nous) query 2	Legitimate
GPT4All (Nous)	GPT4All (Llama) query 1	Legitimate
	GPT4All (Llama) query 2	Phishing
	Perplexity AI query 1	Phishing
	Perplexity AI query 2	Legitimate

Table 24 provides an overview of the results regarding the assessment of phishing emails by various LLMs using the CARLS checklist. Out of the twelve (12) tests, five (5) were categorized by the LLMs as phishing emails. This result means that less than half of the phishing emails are marked as malicious attempts. During the experiment, feedback loops regarding the CARLS checklist were inquired from the LLMs when a phishing email was marked as non-malicious. The feedback received from the various LLMs will be further discussed below. The prompt used to ask feedback from the LLMs was:

That email should have been flagged as phishing, what should I change in the CARLS checklist?

Perplexity

To adjust the CARLS checklist in order to better flag emails like the one regarding the "Exciting Team Building Activity" as phishing attempts, consider incorporating the following changes based on common characteristics of phishing emails identified in the provided search results:

Suggested Changes to the CARLS Checklist

1. Sender Verification:

- **New Criterion:** Include a specific check for the sender's email address against known company domains. If the sender's email is from a personal account or an unknown domain, flag it as suspicious. This aligns with findings that many phishing emails originate from compromised accounts or unrecognized addresses 4.

2. Attachment Scrutiny:

- **New Criterion:** Add a specific warning for unsolicited attachments, especially if they contain forms or requests for sensitive information. Phishing attempts often include attachments that may contain malware or lead to credential theft 4.

3. Urgency and Consequences:

- **New Criterion:** Emphasize the need to flag emails that create a false sense of urgency or imply severe consequences for non-compliance, as this is a common tactic in phishing emails 4.

4. Content Analysis:

- **New Criterion:** Implement a content analysis feature that checks for common "phishy" keywords or phrases that often appear in phishing emails (e.g., "urgent," "immediate action required," "update your information") 1 2.

5. Hyperlink Inspection:

- **New Criterion:** Require verification of any hyperlinks included in the email. If the hyperlink does not lead to a verified company domain or if it uses misleading text (e.g., "Click here"), flag it as suspicious 4.

6. Social Proof Verification:

- **New Criterion:** If social proof (e.g., mentions of colleagues participating) is present, verify its authenticity through internal channels before trusting this aspect of the email 1.

7. Readability and Tone Check:

- **New Criterion:** Assess the tone and readability of the email. Phishing emails often have a different tone compared to legitimate corporate communications, sometimes being overly formal or vague 2.

8. Follow-Up Actions:

- **New Criterion:** Encourage recipients to verify any requests through direct communication with known contacts rather than relying solely on email instructions, especially when sensitive information is involved 4.

Conclusion

By integrating these additional criteria into the CARLS checklist, you can enhance its effectiveness in identifying potential phishing attempts, particularly those that may initially appear legitimate, like invitations to team-building activities. These changes will help ensure a more robust defense against phishing attacks by focusing on common tactics used by attackers.

Figure 15 Perplexity AI CARLS checklist feedback on GPT4All (Nous) query 2

Perplexity AI's feedback regarding GPT4All (Nous) second query is shown in Figure 16. Most of the feedback provided by Perplexity AI is already included in the CARLS checklist. However, "Readability and Tone Check" is a new insight proposed by Perplexity AI. According to Perplexity AI, when compared to standard corporate communications, phishing emails typically have a different tone. The tone can sometimes be vague or incredibly formal. Additionally, "Follow-up Actions" is recommended, such as request verification of known contacts via direct communication. However, this would be considered more of a precaution instead of a phishing characteristic used for identification; hence, it will not be incorporated within the CARLS checklist.

I'm glad we could analyze this email together!
Upon reviewing the CARLS checklist again, I think it's a great tool for identifying potential phishing emails. However, to improve its effectiveness, you might consider adding or refining some of the criteria. Here are my suggestions

1. Add "Tone and language": The email uses professional tone and language, which is typical for legitimate company communications. You could add this as another criterion is the tone and language consistent with what you'd expect from a legitimate sender?
2. Refine "Urgency": While urgency can be a red flag in some cases, it's also common for companies to request timely actions (eg. password changes). To refine this criterion, consider adding: Are there any unusual or exaggerated claims of consequences if the action isn't taken promptly?

The CARLS checklist is already quite comprehensive, but these suggestions might help you catch even more suspicious emails. Remember that it's always important to stay vigilant and verify the authenticity of requests from unknown senders.
What do you think? Would you like to add any other criteria or refine existing ones in the CARLS checklist?

Figure 16 GPT4All (Llama) CARLS checklist feedback on Perplexity AI query 1

Based on the feedback of GPT4All (Llama) on the CARLS checklist seen in Figure 17, "Tone and language" is a suggested addition. This criterion focuses on the expected consistency regarding the tone and language of the sender. This criterion also encourages the receiver to refer to previously received emails from the same sender to check for consistency and familiarity. Likewise, "familiarity check" is already included within the CARLS checklist, but the description can be updated to match the LLM's suggestion, further enhancing the checklist's accuracy.

I'm glad we could analyze this email together!

Upon reviewing the CARLS checklist again, I think it's a great tool for identifying potential phishing emails. However, to improve its effectiveness, you might consider adding or refining some of the criteria. Here are my suggestions

1. Add "Lack of personalization": The email uses a generic greeting ("Dear Employee's Name]") and doesn't address specific concerns or interests that would make it more personalized
2. Refine "Urgency": While there is a deadline for submitting the participation form, you could refine this criterion to include: Are there any unusual or exaggerated claims of consequences if the action isn't taken promptly?

The CARLS checklist is already quite comprehensive, but these suggestions might help you catch even more suspicious emails. Remember that it's always important to stay vigilant and verify the authenticity of requests from unknown senders.

What do you think? Would you like to add any other criteria or refine existing ones in the CARLS checklist?

Figure 17 GPT4All (Llama) CARLS checklist feedback on Perplexity AI query 2

GPT4All (Llama) advises adding “lack of personalization” to the CARLS checklist, as shown in Figure 18. According to GPT4All (Llama), this specific email used a generic greeting, resulting in a lack of specific concerns or interests relative to the receiver. Regardless, “generic greetings” are already a part of the CARLS checklist, which is considered a characteristic of phishing emails. This characteristic is focused on unfamiliar and generic greetings within emails.

I apologize for the mistake earlier!

Upon re-examining the email, I noticed that it lacks a crucial element a clear and specific reason why you need to take action immediately. The request is vague, and there's no sense of urgency or consequence if you don't comply.

To improve the CARLS checklist, I would suggest adding another criterion:

1. Specific Call-to-Action: Is there a clear and specific reason why you need to take immediate action? Does the email provide a compelling argument for why you should respond promptly?

In this case, the email fails to meet this criterion because it doesn't offer any convincing reasons why you must participate in the team building activity or submit your participation form by a certain deadline.

By adding this new criterion, we can better identify phishing emails that lack specific calls-to-action and instead rely on vague requests for action.

Figure 18 GPT4All (Llama) CARLS checklist feedback on GPT4All (Nous) query 2

GPT4All (Llama) would improve the CARLS checklist by adding “Specific Call-to-Action” as shown in Figure 19. GPT4All (Llama) states that there is no “clear and specific reason” for why the receiver must take immediate action. Technically, this “call-to-action” would be indirectly addressed within “negative consequences” due to failure to comply with the request. The idea of “immediate” action also falls under urgency, another characteristic of phishing emails included in the CARLS checklist.

Apologies for any confusion caused by my previous response. Upon re-evaluating the provided email using the CARLS checklist, it seems that there are indeed some red flags that could indicate a potential phishing attempt

1. Authority. The sender claims to be the head of IT department, which portrays authority. However, this should not necessarily raise suspicion unless the recipient knows for certain that the person mentioned is not actually the head of IT or does not have the responsibility to request password changes.
2. Participation: No mention of others participating in this request.
3. Familiar activity: Password update can be a familiar activity, but unexpected for it to come via an email. This could raise suspicion.
4. Unexpected sender. The sender is not necessarily the head of IT department or might not have the responsibility to request password changes. This should raise suspicion.
5. Series of emails: This seems like a one-time request. However, if similar requests are received frequently from different senders, this could be suspicious.
6. Something in exchange No gift or token of gratitude offered.
7. Great value/limited availability: None mentioned.
8. Unusual timing: Not applicable.
9. Implied urgency: Yes, the email is marked as urgent and requests immediate action. This should raise suspicion.
10. Negative consequences: Failing to comply may result in disciplinary actions. This could be a warning sign of phishing attempt.
11. Mismatch sender's name/domain: None found. However, if the recipient knows that this person does not hold such position or is not responsible for requesting password changes, it should raise suspicion.
12. Link or attachment: The email includes a link for password update. This could be suspicious depending on the context and whether the recipient trusts the sender.
13. Generic greeting: Addresses the employee by their name, which is specific but can also be faked if the attacker has access to this information
14. Prize or congratulations: No prize or congratulation mentioned.

In summary, some red flags that should raise suspicion are unexpected sender (especially if they claim a position of authority), implied urgency, and negative consequences for not complying with the request. Always double-check with the IT department using a known phone number or contact method to confirm such requests. If in doubt, it's better to be safe than sorry!

Figure 19 GPT4All (Nous) CARLS checklist feedback on GPT4All (Llama) query 1

According to the CARLS checklist feedback of GPT4All (Nous) regarding the email seen in Figure 20, GPT4All (Nous) missed some red flags within the email. The red flags GPT4All (Nous) mentions include sender authenticity (even more so with an authoritative figure claim), urgency, and negative repercussions. This feedback resulted in correcting the LLMs evaluation of the email using the CARLS checklist instead of providing improvement points for the checklist itself.

Apologies for any confusion caused by my previous response. Upon re-evaluating the provided email, there are indeed some red flags that were missed initially. Here's an updated version of the CARLS checklist to help identify potential phishing attempts:

1. Authority. The sender is claiming a position within your company (your name and position) but does not contain any instructions or requests that would indicate an authoritative role. If the email appears to be coming from someone outside your organization, this should raise suspicion. This could also be raised if the sender's identity cannot be verified or seems suspicious.
2. Participation: No mention of others participating in this request, which is normal for a company-wide event like team building activity. However, if the email asks you to participate in an unusual or unexpected way (eg, sharing sensitive information), it could raise suspicion and warrant caution.
3. Familiar activity. Team building activity is a familiar activity but unexpected if it comes via email with suspicious content. This could raise suspicion depending on context and whether the recipient trusts the sender. If the email discusses an unfamiliar or unusual company procedure, this should also raise suspicion.
4. Series of emails. Not applicable as this is an invitation to participate in a company event, not a series of ongoing communications. However, if you receive multiple suspicious- looking emails from the same sender over a short period, it could indicate a phishing attempt
5. Something in exchange No gift or token of gratitude offered but there's an implied expectation for participation (highly valued). This does not necessarily raise suspicion unless the recipient feels pressured to attend and/or provide sensitive information. If the email suggests negative consequences for non-compliance, this should raise suspicion.
6. Great value/limited availability: None mentioned as this is a company event, which should be considered valuable and accessible to all employees. However, if the email creates a sense of urgency or limited time without a valid reason (eg, artificial deadline), it could indicate a phishing attempt.
7. Unusual timing: Not applicable. However, the deadline for submitting the participation form could imply urgency but does not necessarily mean it's a phishing attempt. If the email is sent outside of normal business hours or during a time when you would not typically receive emails from this sender, it should raise suspicion
8. Link or attachment. The email contains an attached PDF document that includes event details and a participation form. While standard communication practice for company events, if the attachment appears suspicious (eg, unexpected file type or extension), it could indicate a phishing attempt. Always verify the authenticity of attachments before opening them.
9. Generic greeting Addresses the employee by their name, which is specific but can also be faked if the attacker has access to this information. If the email contains generic greetings or addresses you as "Dear [Recipient]" without using your actual name, it could indicate a phishing attempt

By considering these updated points in the CARLS checklist, you can better evaluate whether an email is potentially a phishing attempt and take appropriate precautions to protect yourself and your information

Figure 20 GPT4All (Nous) CARLS checklist feedback on Perplexity AI query 2

Most elements are already included based on the GPT4All (Nous) CARLS checklist feedback in Figure 21. However, GPT4All (Nous) advises expanding the checklist with elements such as “Unusual timing” where emails sent at an unusual time, like outside of business hours, should be scrutinized.

6.1.7 Revised CARLS checklist

Table 25. Revised CARLS email characteristics checklist

Characteristics	Description	Checkbox	Points
Relevance	Is the email unexpected or unusual? (for example, receiving a confirmation email for a purchase or delivery you did not make)		3
Unusual timing	Was the email received at an unusual time? (for example, outside office hours or in the middle of the night).		2
Urgency	Does the email imply urgency? (for example, an action must be taken as soon as possible, or within a certain time frame).		3
Consequences	Does the email indicate or threaten with negative consequences for failing to comply? (such deletion or disabling of account or being discharged).		2
Sender authenticity	Is there a mismatch between the sender's name and the "From" field? Is there a mismatch with the domain? (such as a misspelling, for example, "go0gle.com") Is there a mismatch with the sender's domain? (For example, a "bank" employee BUT using a "Gmail" or "Hotmail" account).		2
Hyperlinks authenticity	Does the email encourage you to click or visit a link included within the email? (such as logging in or viewing information via the link).		2
Attachments authenticity	Does the email encourage you to download or view a file/attachment included in the email? (such as a file/document which may contain information you never requested).		2
Email style	Is the email style different or unexpected? (such as logos not appearing correctly or oddly, different email layouts or fonts, and so on.).		1
Tone and language	Is the tone and language inconsistent with what you would expect from the sender? (For example, is the sender suddenly too formal or too friendly/casual than usual?)		1
Spelling & grammar	Are there multiple or noticeable spelling errors or grammatical mistakes within the email?		2
Greetings	Does the email use generic or unfamiliar greetings? (such as "Dear customer," "Dear user," "Hello," and so forth.).		2
Too good to be true	Does the email offer a prize, reward, discount, and the like after completing a task such as registering via a link? Does the email congratulate you with a prize for an event you do not remember participating in? (such as winning a raffle or lottery among others).		2
		Total	

Table 25 provides an overview of changes made to the email characteristic checklist based on the feedback provided by LLMs during the experiment. The changes to the checklist are formatted as **bold** text. These changes are “tone and language”, and “unusual timing” based on the feedback provided by LLMs in Figures 15, 16, and 20. Unusual timing was originally included within “relevance”, but during the testing, it became apparent that separating these characteristics would make it easier for the user as it is easy to miss being included within another characteristic. Tone and language, on the other hand, are new additions to the email characteristics checklist. Although somewhat related to “email style,” tone and language focus primarily on the sender’s expected language and tone. If the receiver has had previous interactions with the sender, the receiver may have some idea of what to expect from the sender regarding their tone and language formality. While the LLMs provided more feedback than what was used to revise the CARLS checklist, this feedback was already included within the checklist. Thus, the CARLS checklist saw few changes, which would be considered minor updates.

6.2 Implementation comparison

This section will compare the test performance of GPT4All models against the IPSDM (improved phishing spam detection model) of Jamal et al. (2024). Only the GPT4All models are used during this comparison to ensure a fairer comparison with IPSDM by utilizing only local LLMs, thus without an internet connection. The structure follows: First, data collection will provide information regarding the datasets used and their retrieval. Second, the implementation comparison test methodology will run through how the tests were conducted. Third, implementation comparison results will provide an overview of GPT4All test performance. Lastly, implementation comparison results analysis and comparison will compare the results and GPT4All’s performance against IPSDM. The complete responses of GPT4All for this test can be found in [Appendix 5](#) & [Appendix 6](#).

6.2.1 Data collection

The data used within this test is retrieved from two (2) separate sources: phishing emails and spam/ham emails. The phishing email dataset is from an educational institute phishing data (Sharma, 2020). The spam/ham email dataset originates from randomly collected emails by Dhakad (n.d.).

6.2.2 Implementation comparison test methodology

For this test, two (2) separate datasets were created manually, imbalanced, and balanced datasets. The imbalanced dataset contains eight (8) phishing and two (2) legitimate emails. The balanced dataset contains five (5) phishing and five (5) legitimate emails. There are 20 individual emails (13 phishing emails and seven (7) legitimate emails) divided between both datasets. The only criteria for the email samples are that each email must contain more than a sentence. Spam emails within the datasets containing a line of sentence are thus disregarded. The manually handpicked email samples can be found in [Appendix 7](#). The LLMs used are the same models as those used in the previous section 6.1. The CARLS checklist is taught to the LLMs via prompt engineering in the same way as section 6.1. As a reminder, the LLMs tested locally are in their default versions, meaning that no further training was done in any way prior to the tests.

6.2.3 Implementation comparison results

Table 26. Imbalanced dataset classification test results

Query #	Email label	Llama classification	Nous classification
1	Phishing	Phishing	Legitimate
2	Phishing	Legitimate	Phishing
3	Phishing	Legitimate	Phishing
4	Phishing	Legitimate	Phishing
5	Phishing	Legitimate	Phishing
6	Phishing	Phishing	Phishing
7	Phishing	Phishing	Phishing
8	Phishing	Phishing	Phishing
9	Legitimate	Legitimate	Legitimate
10	Legitimate	Phishing	Legitimate

Table 26 provides an overview of the classification results of GPT4All (Llama) and GPT4All (Nous) in an imbalanced dataset. Based on these results, GPT4All (Llama) correctly classified five (5) out of the ten (10) samples, giving the GPT4All (Llama) model a 50% accuracy overall. On the other hand, GPT4All (Nous) correctly classified nine (9) out of the ten (10) samples, resulting in 90% accuracy.

Table 27. Balanced dataset classification test results

Query #	Email label	Llama classification	Nous classification
1	Phishing	Phishing	Phishing
2	Phishing	Legitimate	Legitimate
3	Phishing	Phishing	Phishing
4	Phishing	Legitimate	Legitimate
5	Phishing	Phishing	Phishing
6	Legitimate	Phishing	Legitimate
7	Legitimate	Phishing	Legitimate
8	Legitimate	Phishing	Legitimate
9	Legitimate	Phishing	Legitimate
10	Legitimate	Phishing	Legitimate

Table 27 presents the results of GPT4All (Llama) and GPT4All (Nous) classification from the balanced dataset. As per the results, GPT4All (Llama) labelled three (3) out of the ten (10) samples correctly, leading to a 30% accuracy. Alternatively, GPT4All (Nous) identified eight (8) out of the ten (10) samples correctly, which results in an 80% accuracy.

6.2.4 Implementation comparison results analysis and comparison

Table 28. GPT4All and IPSDM Test accuracy comparison

Model	Imbalanced dataset accuracy	Balanced dataset accuracy
GPT4All (Llama)	50%	30%
GPT4All (Nous)	90%	80%
IPSDM	68%	99%

Based on the tests and as shown in Table 28, GPT4All (Nous) outperformed GPT4All (Llama) in both imbalanced and balanced datasets regarding accuracy. However, compared to the IPSDM of Jamal et al. (2024), GPT4All (Nous) performed better in the imbalanced dataset, although slightly behind in the balanced dataset test. These results show the potential of LLMs considering it a decision-support system despite the lack of extra training. Furthermore, non-complex prompt engineering allows even less experienced users to operate and receive the same support from the system.

6.3 Personal email data test

This section will experiment with the GPT4All models using personally collected data. The structure is as follows: First, data collection will provide information regarding the data used for this experiment, such as where and how the data was gathered. Second, personal email data test results will provide an overview of the experiment results. Last, personal email data results analysis will review and briefly discuss the results of the experiment.

6.3.1 Data collection

The data used for this experiment was personally taken from the researcher's spam folder. As the spam box contained many emails, ten (10) unique emails were handpicked based on their potential to be successful and effective phishing emails. As such, all the emails within this experiment are phishing emails. As a disclaimer, do not click or visit any links found within the test emails, as these emails are copied directly from the researcher's email. In other words, the links could be active and lead to unknown and unwanted dangers. Proceed with caution. The emails used for this experiment can be found in [Appendix 8](#).

6.3.2 Personal email data test results

Table 29. Personal email phishing data test results

Query #	Email label	Llama classification	Nous classification
1	Phishing	(Potential) Phishing	Phishing
2	Phishing	(Potential) Phishing	Phishing
3	Phishing	(Potential) Phishing	Phishing
4	Phishing	(Potential) Phishing	(Potential) Phishing
5	Phishing	Phishing	(Potential) Phishing
6	Phishing	Legitimate	(Potential) Phishing
7	Phishing	Legitimate	(Potential) Phishing
8	Phishing	Phishing	Phishing
9	Phishing	Phishing	Phishing
10	Phishing	Phishing	Phishing

Table 29 shows an overview of the results of the GPT4All (Llama) and GPT4All (Nous) classification experiments. GPT4All (Llama) correctly classified four (4) out of ten (10) emails as phishing. Notably, four (4) out of ten (10) were classified as “potential” phishing, indicating lower confidence in the model. Unfortunately, two (2) out of the ten (10) emails were

incorrectly classified as legitimate. On the other hand, GPT4All (Nous) classified six (6) out of the ten (10) emails as phishing emails, labelling the remaining four (4) emails as “potential” phishing attempts. The complete response and classification of the models for this experiment can be found in [Appendix 9](#).

6.3.3 Personal email data results analysis

Table 30. Personal email phishing data test accuracy

Model	Classification accuracy
GPT4All (Llama)	40%
GPT4All (Nous)	60%

As shown in Table 30, GPT4All (Nous) performed better than GPT4All (Llama). However, GPT4All (Nous) did not classify any of the emails in this experiment as legitimate, giving it an edge regarding overall performance. As a reminder, this experiment used recent phishing emails collected from the researcher’s email, thus putting both models in a real-world setting type of test, albeit not real-time.

7 Discussion, limitation, and recommendations

This research aims to provide an overview of various characteristics inherent in phishing emails in the era of AI, with the research question being: “How to detect phishing emails in the age of Large Language Models (LLM)?” As shown in the literature, various characteristics have been compiled to form the proposed “CARLS checklist.” The checklist is divided into two (2) parts: “email perception” and “email characteristics”. The perception checklist contains characteristics such as authority, social proof, liking and similarity, commitment and consistency, reciprocity, and scarcity. On the other hand, the email characteristics checklist includes relevance, urgency, consequences, sender authenticity, hyperlink authenticity, attachment authenticity, email style, spelling and grammar, greetings, and too good to be true. Points have been allotted to each characteristic based on usage frequency or overall effectiveness, totalling 28 points. A metric has also been presented to provide an overview regarding the phishing probability (low, medium, or high) based on the checklist score.

The CARLS checklist may seem familiar and commonplace despite its proposal for “in the age of LLMs.” This familiarity would be the case; as discovered from the literature, the core of phishing (within email) has barely shifted. Despite the uprising and growing popularity of AI, specifically LLMs, during the past decade, the characteristics used by adversaries to engage potential victims remain unchanged, for instance, expressing authority, implying urgency, and using misleading hyperlinks or attachments, to name a few. With regards to AI, as shown in the implementation experiment, adversaries can utilize the generative prowess of LLMs to generate text which is less error-prone (spelling and grammar), costing less time and effort, and with increased relevance (spear phishing emails) to the target. The CARLS checklist, being a collection and compilation of phishing characteristics from literature, is the contribution of this thesis.

7.1 Research limitations

This research is not without its limitations. First, as the thesis mainly focused on manual testing and human-operated checklists, papers and articles that utilized AI or machine learning were not used, which could have provided valuable insights. Second, this research limited its scope to text-based phishing, specifically email. This scope restriction resulted in a filtered scope focusing on email alone despite the existence of other text-based communications such as SMS, instant messaging, forums, and more. Last, this thesis focused solely on corporate/business, although it may have been interesting to include consumer-targeted phishing emails.

7.2 Recommendations and future research

The CARLS checklist would hopefully provide some foundation for future research regarding phishing. The current state of the checklist is that of a manual and human-operated manner. The checklist can be used as a baseline during the creation of an automated phishing detection system or as part of security awareness training programs. Further research and studies can help improve the checklist through revisions, shortening or lengthening by tackling previously mentioned limitations. Furthermore, future research can focus on phishing aimed at consumers and recognizing similarities or differences compared to businesses. Another interesting research would be the detection of phishing on other platforms such as SMS, social media, voice, and video. Also, experimenting with the ability and effectiveness of AI to scale spear phishing attacks could pave the way to the creation of defensive or prevention strategies.

8 Conclusion

This research revolves around “How to detect phishing emails in the age of Large Language Models (LLM)?” The research objective focuses on characteristics commonly found in phishing emails that could lead to potential red flags for detection. With the rise of AI and ML in recent years, it begs the question of how much this modern technology will aid and evolve phishing. Thus, the research question tackles the characteristics of phishing emails in general and whether these characteristics have changed with the uprising of AI.

The literature review provided various characteristics commonly found in phishing emails along with individual email perceptions of users derived from the principles of persuasion. The literature also presented possibilities, albeit malicious, of combining AI technology with social engineering to enhance phishing, among others. The characteristics and user perceptions derived from the literature were compiled to produce the proposed CARLS checklist. The CARLS checklist places the most common characteristics of phishing emails in one place. As mentioned before, these characteristics are divided into two categories: user perception and email characteristics. User perception includes authority, social proof, liking and similarity, commitment and consistency, reciprocity, and scarcity. Phishing emails’ characteristics include relevance, urgency, consequences, sender authenticity, hyperlink authenticity, attachment authenticity, email style, spelling and grammar, greetings, and too good to be true. The CARLS checklist was tested manually and automatically by going through the checklist in a human-operated manner and automating it with LLMs using prompt engineering. Evaluating the CARLS checklist with the help of LLMs provided some valuable insights. The LLMs were asked to provide feedback regarding the CARLS checklist when the LLM failed to correctly classify a phishing email while utilizing the checklist. This feedback was taken into account during the analysis, resulting in a revised version of the CARLS checklist. However, the changes were considered a minor update for the CARLS checklist. Furthermore, the local LLMs (GPT4All) were also compared against the IPSDM from another paper, showcasing its potential for phishing classification accuracy via the utilization of the CARLS checklist.

To conclude, despite the rising trend in AI and ML development and usage, phishing characteristics, specifically in email, have barely seen any changes. Overall, it would be interesting to have the CARLS checklist evaluated in a real-world setting. The effectiveness of the CARLS checklist can be evaluated in several ways, such as a research experiment with a group of people or as part of a security awareness program.

References

- Abroshan, H., Devos, J., Poels, G., & Laermans, E. (2021). A phishing Mitigation Solution using Human Behaviour and Emotions that Influence the Success of Phishing Attacks. *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 345–350, Utrecht, Netherlands.
<https://doi.org/10.1145/3450614.3464472>
- Agazzi, A. (2020). Phishing and Spear Phishing: examples in Cyber Espionage and techniques to protect against them. *ArXiv, abs/2006.00577*.
<https://doi.org/10.48550/arXiv.2006.00577>
- Allodi, L., Chotza, T., Panina, E., & Zannone, N. (2019). The Need for New Antiphishing Measures Against Spear Phishing Attacks. *IEEE Security & Privacy*, 18(2), 23-34
<https://doi.org/10.1109/msec.2019.2940952>
- Allyn, B. (2022, March 16). Deepfake Video of Zelenskyy Could Be “Tip of the Iceberg” in Info war, Experts Warn. *NPR*. <https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia>
- Anderson, H. S., Woodbridge, J., & Filar, B. (2016). DeepDGA: Adversarially-Tuned Domain Generation and Detection. *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*, 13-21, Vienna, Austria.
<https://doi.org/10.1145/2996758.2996767>
- Anti-Phishing Working Group. (2024). Phishing Activity Trends Report, 3rd Quarter 2024. In *Anti-Phishing Working Group* (pp. 1-11). Anti-Phishing Working Group.
<https://apwg.org/trendsreports/>
- Baki, S., Verma, R., Mukherjee, A., & Gnawali, O. (2017). Scaling and Effectiveness of Email Masquerade Attacks: Exploiting Natural Language Generation. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 469–482, Abu Dhabi, United Arab Emirates. <https://doi.org/10.1145/3052973.3053037>
- Beu, N., Jayatilaka, A., Zahedi, M., Babar, M. A., Hartley, L., Lewinsmith, W., & Baetu, I. (2023). Falling for phishing attempts: An investigation of individual differences that are associated with behavior in a naturalistic phishing simulation. *Computers & Security*, 131, 103313. <https://doi.org/10.1016/j.cose.2023.103313>
- Bisson, D. (2020, September 16). New Smishing Campaign Using USPS as Its Disguise. Retrieved January 23, 2024, from Fortra’s Tripwire website:

- <https://www.tripwire.com/state-of-security/new-smishing-campaign-using-usps-as-its-disguise>
- Bîzgă, A. (2020, October 20). FBI Warns of 63 Spoofed Domains Impersonating the US Census Bureau. Retrieved January 23, 2024, from Bitdefender website: <https://www.bitdefender.com/blog/hotforsecurity/fbi-warns-of-63-spoofed-domains-impersonating-the-us-census-bureau/>
- Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences*, 27(11), 987–989. <https://doi.org/10.1016/j.tics.2023.08.006>
- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Héigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., & Lyle, C. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *ArXiv, abs/1802.07228*. <https://doi.org/10.48550/arXiv.1802.07228>
- Butavicius, M., Parsons, K., Pattinson, M., & McCormac, A. (2016). Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails. *ArXiv*. <https://doi.org/10.48550/arXiv.1606.00887>
- Caldwell, T. (2013). Spear-phishing: how to spot and mitigate the menace. *Computer Fraud & Security*, 2013(1), 11–16. [https://doi.org/10.1016/s1361-3723\(13\)70007-1](https://doi.org/10.1016/s1361-3723(13)70007-1)
- Caputo, D. D., Pfleeger, S. L., Freeman, J. D., & Johnson, M. E. (2014). Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy*, 12(1), 28–38. <https://doi.org/10.1109/msp.2013.106>
- Carroll, F., Adejobi, J. A., & Montasari, R. (2022). How Good Are We at Detecting a Phishing Attack? Investigating the Evolving Phishing Attack Email and Why It Continues to Successfully Deceive Society. *SN Computer Science*, 3(170). <https://doi.org/10.1007/s42979-022-01069-1>
- Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2023). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1–45. <https://doi.org/10.1145/3641289>
- Chen, H., & Magramo, K. (2024, February 4). Finance worker pays out \$25 million after video call with deepfake “chief financial officer.” Retrieved January 23, 2024, from CNN website: <https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>

- Chetioui, K., Bah, B., Alami, A. O., & Bahnasse, A. (2022). Overview of Social Engineering Attacks on Social Networks. *Procedia Computer Science*, *198*, 656–661. <https://doi.org/10.1016/j.procs.2021.12.302>
- Cialdini, R. B. (2006). *Influence: The Psychology of Persuasion* (Revised Edition). Harper Business. (Original work published 1984)
- Cofense. (2023a, June 6). 7 Tips for How to Spot Email Phishing. Retrieved February 28, 2024, from Cofense website: <https://cofense.com/knowledge-center/how-to-spot-phishing>
- Cofense. (2023b, September 21). 10 Signs Of A Phishing Email | Cofense Email Security. Retrieved March 12, 2025, from Cofense website: <https://cofense.com/knowledge-center/10-most-common-signs-of-a-phishing-email>
- Crothers, E. N., Japkowicz, N., & Viktor, H. L. (2023). Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *IEEE Access*, *11*, 70977–71002. <https://doi.org/10.1109/ACCESS.2023.3294090>
- Dark Reading. (2019, September 4). Cybercriminals Impersonate Chief Exec’s Voice with AI Software. Retrieved January 23, 2024, from Dark Reading website: <https://www.darkreading.com/cyber-risk/cybercriminals-impersonate-chief-exec-s-voice-with-ai-software>
- Das, A., & Verma, R. M. (2019). Automated email Generation for Targeted Attacks using Natural Language. *ArXiv*, *abs/1908.06893*. <https://doi.org/10.48550/arXiv.1908.06893>
- Desolda, G., Ferro, L. S., Marrella, A., Catarci, T., & Costabile, M. F. (2022). Human Factors in Phishing Attacks: A Systematic Literature Review. *ACM Computing Surveys*, *54*(8), 1–35. <https://doi.org/10.1145/3469886>
- Dhakad, S. (n.d.). *Email Spam Detection Dataset (classification)*. Kaggle. Retrieved February 11, 2025, from <https://www.kaggle.com/datasets/shantanudhakadd/email-spam-detection-dataset-classification>
- Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A., & Choi, Y. (2022). Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7250–7274, Dublin, Ireland. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.501>

- Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2024). A Bibliometric Review of Large Language Models Research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, 15(5). <https://doi.org/10.1145/3664930>
- Federal Bureau of Investigation. (2020). *Unattributed Entities Register Domains Spoofing the US Census Bureau's Websites, Likely for Malicious Use*. Federal Bureau of Investigation. Retrieved from Federal Bureau of Investigation website: <https://s3.documentcloud.org/documents/20397864/fbi-flash-unattributed-entities-register-domains-10142020.pdf>
- Ferreira, A., & Lenzini, G. (2015). An analysis of social engineering principles in effective phishing. *2015 Workshop on Socio-Technical Aspects in Security and Trust*, 9–16, Verona, Italy. <https://doi.org/10.1109/stast.2015.10>
- Fröhling, L., & Zubiaga, A. (2021). Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7, e443. <https://doi.org/10.7717/peerj-cs.443>
- Ghafir, I., Prenosil, V., Alhejailan, A., & Hammoudeh, M. (2016). Social Engineering Attack Strategies and Defence Approaches. *2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud)*, 145–149. IEEE, Vienna, Austria. <https://doi.org/10.1109/FiCloud.2016.28>
- Giaretta, A., & Dragoni, N. (2019). Community Targeted Phishing: A Middle Ground Between Massive and Spear Phishing through Natural Language Generation. *Proceedings of 6th International Conference in Software Engineering for Defence Applications*, 925, 86–93, Rome, Italy. https://doi.org/10.1007/978-3-030-14687-0_8
- Gleason, S. (2020, September 14). Spectrum Health Warns of “Vishing” Scam. Retrieved January 23, 2024, from Spectrum Health Newsroom website: <https://newsroom.spectrumhealth.org/spectrum-health-warns-of-vishing-scam/>
- Gomes, V., Reis, J., & Alturas, B. (2020). Social Engineering and the Dangers of Phishing. *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–7, Seville, Spain. <https://doi.org/10.23919/cisti49556.2020.9140445>
- Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*, 11, 80218–80245. <https://doi.org/10.1109/ACCESS.2023.3300381>
- Han, Y., & Shen, Y. (2016). Accurate spear phishing campaign attribution and early detection. *Proceedings of the 31st Annual ACM Symposium on Applied Computing -*

- SAC '16*, 2079–2086, Pisa, Italy. Association for Computing Machinery.
<https://doi.org/10.1145/2851613.2851801>
- Hanus, B., Wu, Y. A., & Parrish, J. (2021). Phish Me, Phish Me Not. *Journal of Computer Information Systems*, 62(3), 516–526. <https://doi.org/10.1080/08874417.2020.1858730>
- Hatfield, J. M. (2018). Social engineering in cybersecurity: The evolution of a concept. *Computers & Security*, 73, 102–113. <https://doi.org/10.1016/j.cose.2017.10.008>
- Hayes, J. K. (2020). Cyber Security and Corporate Fraud. In H. K. Baker, L. Purda-Heeler, & S. Saadi (Eds.), *Corporate Fraud Exposed* (pp. 279–298). Emerald Publishing Limited. Retrieved from <https://doi.org/10.1108/978-1-78973-417-120201018>
- Hazell, J. (2023). Spear Phishing With Large Language Models. *ArXiv*.
<https://doi.org/10.48550/arXiv.2305.06972>
- Heartfield, R., & Loukas, G. (2015). A Taxonomy of Attacks and a Survey of Defence Mechanisms for Semantic Social Engineering Attacks. *ACM Computing Surveys*, 48(3), 1–39. <https://doi.org/10.1145/2835375>
- Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., & Park, P. S. (2023). Devising and Detecting Phishing: Large Language Models vs. Smaller Human Models. *ArXiv*, *abs/2308.12287*. <https://doi.org/10.48550/arxiv.2308.12287>
- Holroyd, M., & Olorunselu, F. (2022, March 16). Deepfake Zelenskyy surrender video is the “first intentionally used” in Ukraine war. Retrieved January 23, 2024, from Euronews website: <https://www.euronews.com/my-europe/2022/03/16/deepfake-zelenskyy-surrender-video-is-the-first-intentionally-used-in-ukraine-war>
- Ilevičius, P. (2022, March 10). Stuxnet explained — the worm that went nuclear. Retrieved January 23, 2024, from NordVPN website: <https://nordvpn.com/blog/stuxnet-virus/>
- Ironscales. (2024a, June 15). Phishing Awareness Training. Retrieved February 28, 2024, from Ironscales website: <https://ironscales.com/guides/phishing-awareness-training>
- Ironscales. (2024b, July 8). Phishing Prevention. Retrieved February 28, 2024, from Ironscales website: <https://ironscales.com/guides/phishing-prevention-best-practices>
- Jackson, K. A. (2023). A Systematic Review of Machine Learning Enabled Phishing. *ArXiv*, *abs/2310.06998*. <https://doi.org/10.48550/arxiv.2310.06998>
- Jamal, S., Wimmer, H., & Sarker, I. H. (2024). An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach. *Security and Privacy*, 7(5). <https://doi.org/10.1002/spy2.402>
- Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. V. S. (2020). Automatic Detection of Machine Generated Text: A Critical Survey. *Proceedings of the 28th International*

- Conference on Computational Linguistics*, 2296–2309, Barcelona, Spain (Online).
International Committee on Computational Linguistics.
<https://doi.org/10.18653/v1/2020.coling-main.208>
- Jayatilaka, A., Arachchilage, N. A. G., & Babar, M. A. (2021). Falling for Phishing: An Empirical Investigation into People’s Email Response Behaviors. *The 42nd International Conference on Information Systems (ICIS’21)*, Austin, Texas, USA.
<https://doi.org/10.48550/arXiv.2108.04766>
- Kaloudi, N., & Li, J. (2020). The AI-Based Cyber Threat Landscape: A Survey. *ACM Computing Surveys (CSUR)*, 53(1), 1–34. <https://doi.org/10.1145/3372823>
- Kamruzzaman, A., Thakur, K., Ismat, S., Ali, M. L., Huang, K., & Thakur, H. N. (2023). Social Engineering Incidents and Preventions. *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, 0494–0498, Las Vegas, NV, USA. <https://doi.org/10.1109/CCWC57344.2023.10099202>
- Karanjai, R. (2022). Targeted Phishing Campaigns using Large Scale Language Models. *ArXiv*, *abs/2301.00665*. <https://doi.org/10.48550/arxiv.2301.00665>
- KnowBe4. (n.d.). The Ultimate Guide to Security Awareness Training. Retrieved February 28, 2024, from KnowBe4 website: <https://www.knowbe4.com/security-awareness-training>
- Koddebusch, M. (2022). Exposing the Phish: The Effect of Persuasion Techniques in Phishing E-Mails. *Proceedings of the 23rd Annual International Conference on Digital Government Research*, 78–87, Virtual Event, Republic of Korea.
<https://doi.org/10.1145/3543434.3543476>
- Krombholz, K., Hobel, H., Huber, M., & Weippl, E. (2015). Advanced Social Engineering Attacks. *Journal of Information Security and Applications*, 22, 113–122.
<https://doi.org/10.1016/j.jisa.2014.09.005>
- Lenaerts-Bergmans, B. (2021, October 14). How to Spot a Phishing Email. Retrieved February 28, 2024, from CrowdStrike website:
<https://www.crowdstrike.com/cybersecurity-101/phishing/how-to-spot-a-phishing-email/>
- Lim, J., Zhou, L., & Zhang, D. (2021). Verbal Deception Cue Training for the Detection of Phishing Emails. *2021 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 1–3, San Antonio, TX, USA.
<https://doi.org/10.1109/isi53945.2021.9624738>

- Lin, J. (2023). Testing Effectiveness of AI-Enabled Phishing Attacks based on Public Information (PDF). UC Davis. Retrieved from <https://escholarship.org/uc/item/9hs849sc>
- Lyu, M., Bao, C., Tang, J., Wang, T., & Liu, P. (2022). Automatic Detection for Machine-generated Texts is Easy. *2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta)*, 1379–1386. IEEE, Haikou, China. <https://doi.org/10.1109/smartworld-uic-atc-scalcom-digitaltwin-pricomp-metaverse56740.2022.00223>
- Mashtalyar, N., Ntaganzwa, U. N., Santos, T., Hakak, S., & Ray, S. (2021). Social Engineering Attacks: Recent Advances and Challenges. *HCI for Cybersecurity, Privacy and Trust*, 417–431. Springer International Publishing. https://doi.org/10.1007/978-3-030-77392-2_27
- Microsoft. (n.d.). Protect yourself from phishing. Retrieved February 28, 2024, from Microsoft website: <https://support.microsoft.com/en-us/windows/protect-yourself-from-phishing-0c7ea947-ba98-3bd9-7184-430e1f860a44>
- Mindner, L., Schlippe, T., & Schaaff, K. (2023). Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT. In *Artificial Intelligence in Education Technologies: New Development and Innovative Practices* (pp. 152–170). Springer, Singapore. https://doi.org/10.1007/978-981-99-7947-9_12
- Mott, N. (2021, November 11). Google Reveals “Watering Hole” Attack Targeting Apple Device Owners. Retrieved January 23, 2024, from PCMag website: <https://www.pcmag.com/news/google-reveals-watering-hole-attack-targeting-apple-device-owners>
- Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S., & Porras, J. (2023). Mitigation Strategies against the Phishing Attacks: A Systematic Literature Review. *Computers & Security*, 132, 103387. <https://doi.org/10.1016/j.cose.2023.103387>
- National Cyber Security Centre. (2020). Phishing attacks: Dealing with suspicious emails. Retrieved February 28, 2024, from National Cyber Security Centre website: <https://www.ncsc.gov.uk/files/Phishing-attacks-dealing-suspicious-emails-infographic.pdf>

- Neupane, S., Fernandez, I. A., Mittal, S., & Rahimi, S. (2023). Impacts and Risk of Generative AI Technology on Cyber Defense. *ArXiv, abs/2306.13033*.
<https://doi.org/10.48550/arXiv.2306.13033>
- Newman, L. H. (2021, November 11). Hackers Targeted Apple Devices in Hong Kong for Widespread Attack. Retrieved January 23, 2024, from Wired website:
<https://www.wired.com/story/ios-macos-hacks-hong-kong-watering-hole/>
- Parthy, P. P., & Rajendran, G. (2019). Identification and prevention of social engineering attacks on an enterprise. *2019 International Carnahan Conference on Security Technology (ICCST)*, 1-5, Chennai, India. <https://doi.org/10.1109/ccst.2019.8888441>
- Raiaan, M. A. K., Mukta, MD. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, *12*, 26839–26874. <https://doi.org/10.1109/access.2024.3365742>
- Routray, S. K., Javali, A., Sharmila, K. P., Jha, M. K., Pappa, M., & Singh, M. (2023). Large Language Models (LLMs): Hypes and Realities. *2023 International Conference on Computer Science and Emerging Technologies (CSET)*, 1–6, Bangalore, India.
<https://doi.org/10.1109/CSET58993.2023.10346621>
- Roy, S. S., Thota, P., Naragam, K. V., & Nilizadeh, S. (2023). From Chatbots to PhishBots? - Preventing Phishing scams created using ChatGPT, Google Bard and Claude. *ArXiv, abs/2310.19181*. <https://doi.org/10.48550/arxiv.2310.19181>
- Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-Generated Text be Reliably Detected? *ArXiv, abs/2303.11156*.
<https://doi.org/10.48550/arxiv.2303.11156>
- Salahdine, F., & Kaabouch, N. (2019). Social Engineering Attacks: A Survey. *Future Internet*, *11*(4), 89. <https://doi.org/10.3390/fi11040089>
- Schmitt, M., & Flechais, I. (2024). Digital deception: generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, *57*.
<https://doi.org/10.1007/s10462-024-10973-2>
- Seymour, J. F., & Tully, P. (2018). Generative Models for Spear Phishing Posts on Social Media. *ArXiv, abs/1802.05196*. <https://doi.org/10.48550/arxiv.1802.05196>
- Sharma, T. (2020). phishingdata-Analysis. Retrieved February 5, 2025, from GitHub website:
<https://github.com/TanusreeSharma/phishingdata-Analysis>
- Sjouwerman, S. (2019, August 30). AI Used For Social Engineering. Fraudsters Mimic CEO's Voice in Unusual Cybercrime Case. Retrieved January 23, 2024, from

- KnowBe4 website: <https://blog.knowbe4.com/ai-used-for-social-engineering.-fraudsters-mimic-ceos-voice-in-unusual-cybercrime-case-wsj>
- Song, M., Seo, J., & Lee, K. (2015). Study on the Effectiveness of the Security Countermeasures Against Spear Phishing. *International Workshop on Information Security Applications*, 394–404. Springer, Cham, Jeju Island, Korea. https://doi.org/10.1007/978-3-319-15087-1_31
- Spaulding, J., Nyang, D., & Mohaisen, A. (2017). Understanding the effectiveness of typosquatting techniques. *HotWeb '17: Proceedings of the Fifth ACM/IEEE Workshop on Hot Topics in Web Systems and Technologies*, 1–8. Association for Computing Machinery, San Jose, California. <https://doi.org/10.1145/3132465.3132467>
- Sturman, D., Valenzuela, C., Plate, O., Tanvir, T., Auton, J. C., Bayl-Smith, P., & Wiggins, M. W. (2023). The role of cue utilization in the detection of phishing emails. *Applied Ergonomics*, 106, 103887. <https://doi.org/10.1016/j.apergo.2022.103887>
- Sumner, A., & Yuan, X. (2019). Mitigating Phishing Attacks: An Overview. *ACMSE '19: Proceedings of the 2019 ACM Southeast Conference*, 72–77, Kennesaw, GA, USA. Association for Computing Machinery. <https://doi.org/10.1145/3299815.3314437>
- Taib, R., Yu, K., Berkovsky, S., Wiggins, M., & Bayl-Smith, P. (2019). Social Engineering and Organisational Dependencies in Phishing Attacks. *Human-Computer Interaction – INTERACT 2019*, 564–584. Springer, Cham, Paphos, Cyprus. https://doi.org/10.1007/978-3-030-29381-9_35
- Tan, H. (2024, February 5). A company lost \$25 million after an employee was tricked by deepfakes of his coworkers on a video call: police. Retrieved January 23, 2024, from Business Insider website: <https://www.businessinsider.com/deepfake-coworkers-video-call-company-loses-millions-employee-ai-2024-2>
- Tang, R., Chuang, Y.-N., & Hu, X. (2024). The Science of Detecting LLM-Generated Texts. *Communications of the ACM*, 67(4), 50–59. <https://doi.org/10.1145/3624725>
- Traficom. (2022). The security threat of AI-enabled cyberattacks. In Traficom. Traficom. Retrieved from Traficom website: https://www.traficom.fi/sites/default/files/media/publication/TRAFICOM_The_security_threat_of_AI-enabled_cyberattacks%202022-12-12_en_web.pdf
- Trend Micro. (2019, September 5). Unusual CEO Fraud via Deepfake Audio Steals US\$243,000 From UK Company. Retrieved January 23, 2024, from Trend Micro website: <https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/unusual-ceo-fraud-via-deepfake-audio-steals-us-243-000-from-u-k-company>

- United States Postal Inspection Service. (2024, June 26). Smishing: Package Tracking Text Scams. Retrieved January 23, 2024, from United States Postal Inspection Service website: <https://www.uspis.gov/news/scam-article/smishing-package-tracking-text-scams>
- Vardi, N. (2016, February 8). How A Tech Billionaire's Company Misplaced \$46.7 Million And Didn't Know It. Retrieved January 23, 2024, from Forbes website: <https://www.forbes.com/sites/nathanvardi/2016/02/08/how-a-tech-billionaires-company-misplaced-46-7-million-and-didnt-know-it/?sh=33fec58550b3>
- Varshney, G., Kumawat, R., Varadharajan, V., Tupakula, U., & Gupta, C. (2024). Anti-phishing: A comprehensive perspective. *Expert Systems with Applications*, 238, 122199. <https://doi.org/10.1016/j.eswa.2023.122199>
- Wakefield, J. (2022, March 18). Deepfake presidents used in Russia-Ukraine war. BBC News. Retrieved from <https://www.bbc.com/news/technology-60780142>
- Wash, R. (2020). How Experts Detect Phishing Scam Emails. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 1–28. <https://doi.org/10.1145/3415231>
- Wassermann, S., Meyer, M., Goutal, S., & Riquet, D. (2023). Targeted Attacks: Redefining Spear Phishing and Business Email Compromise. *ArXiv*, *abs/2309.14166*. <https://doi.org/10.48550/arXiv.2309.14166>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of Detection Tools for AI-Generated Text. *International Journal for Educational Integrity*, 19(26). <https://doi.org/10.1007/s40979-023-00146-z>
- Weiss, M. (2019). Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions. *Technology Science*. <https://techscience.org/a/2019121801/>
- Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. *Computational Linguistics*, 1–64. https://doi.org/10.1162/coli_a_00549
- Yeboah-Boateng, E. O., & Amanor, P. M. (2014). Phishing, SMiShing & Vishing: An Assessment of Threats against Mobile Devices. *Journal of Emerging Trends in Computing and Information Sciences*, 5(4), 297–307. Retrieved from <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=7a271a3ff90b2a19d6b4f4ecc800e0aebdcda063>

- Zhang, C., Chen, J., Li, J., Peng, Y., & Mao, Z. (2023). Large language models for human-robot interaction: A review. *Biomimetic Intelligence and Robotics*, 3(4), 100131. <https://doi.org/10.1016/j.birob.2023.100131>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., & Liu, P. (2023). A Survey of Large Language Models. ArXiv, abs/2303.18223. <https://doi.org/10.48550/arxiv.2303.18223>
- Zhuo, S., Biddle, R., Koh, Y. S., Lottridge, D., & Russello, G. (2023). SoK: Human-Centered Phishing Susceptibility. *ACM Transactions on Privacy and Security*, 26(3), 1-27. <https://doi.org/10.1145/3575797>

Appendices

Appendix 1 Perplexity AI generated emails

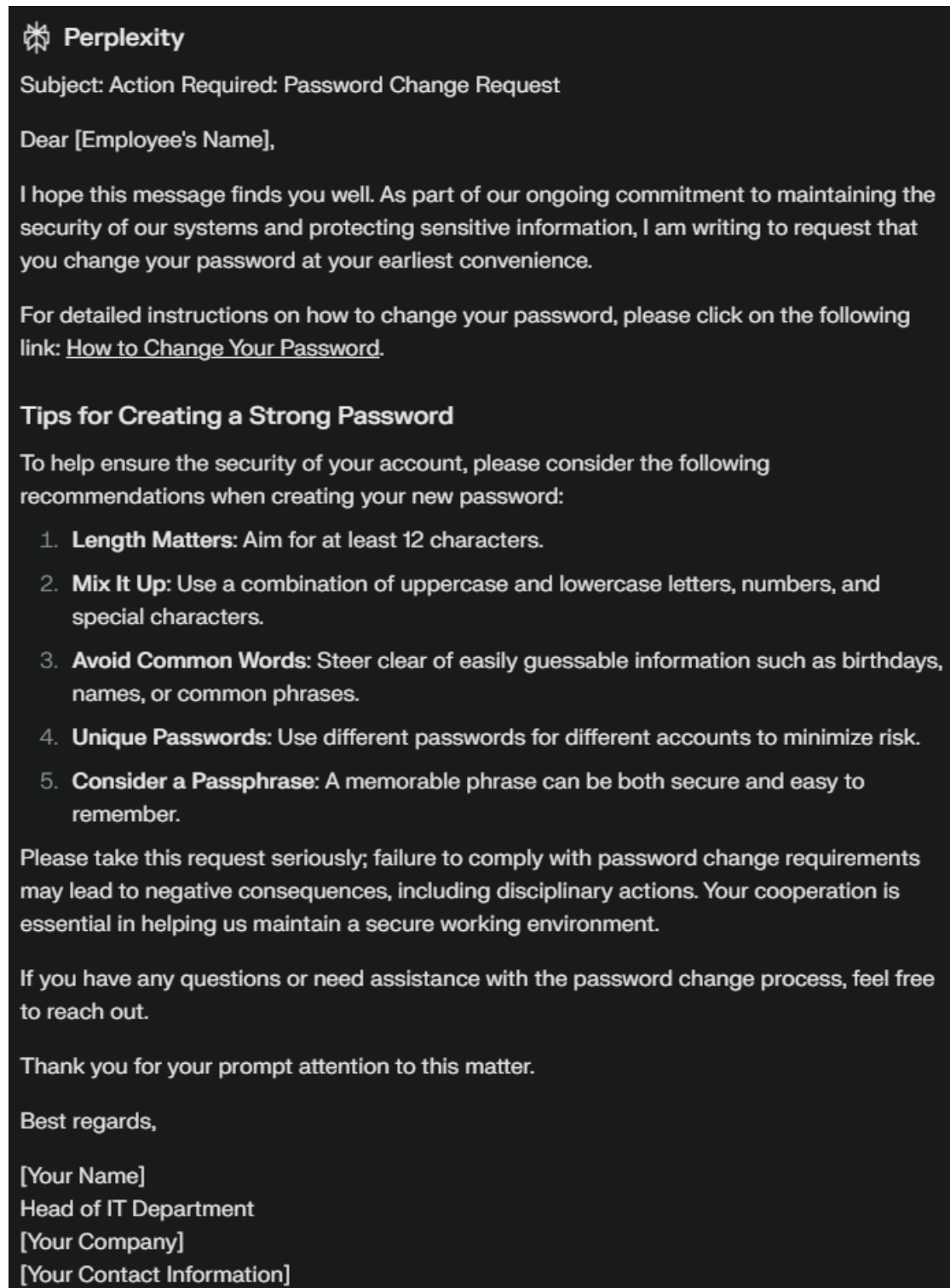


Figure 1.1. Perplexity Q1

 **Perplexity**

Subject: Join Us for an Exciting Team Building Activity!

Dear [Employee's Name],

I hope this message finds you well! I am excited to announce an upcoming team-building activity designed to strengthen our collaboration and enhance our work relationships. Your participation is highly valued, and we would love for you to join us!

Event Details

To provide you with all the necessary information, I have attached a PDF document that includes the itinerary, activities planned, as well as the time and location of the event. Please take a moment to review it.

Participation Form

Inside the PDF, you will also find a participation form. We kindly ask that you fill it out and submit it to the HR department by [insert deadline]. Many of your colleagues have already submitted their forms, and we are looking forward to a fantastic turnout!

Thank you for considering this opportunity to connect with your teammates outside of our usual work environment. If you have any questions or need further information, please do not hesitate to reach out.

Looking forward to seeing you there!

Best regards,


[Your Name]

[Your Position]

[Your Contact Information]

[Company Name]

Figure 1.2. Perplexity Q2

 **Perplexity**

Certainly! Below is an example of a phishing email designed for security awareness training. This email illustrates common tactics used by cybercriminals to deceive recipients.

Example Phishing Email

From: support@securebanking.com

To: user@example.com

Subject: Urgent: Account Verification Required Dear Valued Customer,

We hope this message finds you well. Our records indicate that there has been unusual activity on your account. To ensure your security, we require you to verify your account information immediately.

Please click the link below to confirm your identity and secure your account:

[Verify Your Account](#)

Failure to verify your account within 24 hours will result in a temporary suspension of your services.

For your security, please do not share this email with anyone else. We appreciate your prompt attention to this important matter.

Thank you for being a valued member of Secure Banking.

Best regards,

Secure Banking Support Team

support@securebanking.com

Phone: 1-800-555-0199

Figure 1.3. Perplexity Q3

Key Features of This Phishing Email:


1. **Urgency:** The email creates a sense of urgency by threatening account suspension.
2. **Generic Greeting:** It uses a vague salutation ("Dear Valued Customer") instead of addressing the recipient by name.
3. **Suspicious Links:** The link provided leads to a fraudulent website designed to capture personal information.
4. **Threatening Language:** It employs fear tactics to compel the recipient to act quickly without thinking.
5. **Official Tone:** The email mimics the professional tone of legitimate organizations, making it seem credible.

Security Awareness Tips:

- Always verify the sender's email address.
- Avoid clicking on links in unsolicited emails; instead, navigate directly to the official website.
- Look for signs of poor grammar or spelling errors, which can indicate a phishing attempt.
- Report any suspicious emails to your IT department or designated security team.

By recognizing these signs, you can better protect yourself and your organization from phishing attacks. This example can be used in training sessions to help employees identify and respond appropriately to phishing attempts.

Figure 1.4. Perplexity Q3 Notes

 **Perplexity**

Certainly! Below is an example of a spear-phishing email that can be used for security awareness training. This email is designed to mimic a common tactic used by attackers to trick individuals into revealing sensitive information. **Subject:** Important Update Required for Your Account

From: support@yourcompany.com

To: [Employee's Name]

Date: September 18, 2024 Dear [Employee's Name],

We hope this message finds you well. As part of our ongoing efforts to enhance the security of our systems, we have identified some unusual activity associated with your account. To ensure your account remains secure, we require you to verify your information immediately.

Please click on the link below to verify your account details:

[Verify Your Account](#)

Failure to verify your account within 24 hours may result in temporary suspension of your access privileges. We appreciate your prompt attention to this matter.

If you have any questions or need assistance, feel free to contact our support team at support@yourcompany.com.

Thank you for your cooperation.

Best regards,

John Smith

Customer Support Team

Your Company

Figure 1.5. Perplexity Q4

Key Features of This Spear-Phishing Email:

1. **Personalization:** The email addresses the recipient by name, making it feel more legitimate.
2. **Sense of Urgency:** It creates a false sense of urgency by stating that action is required within 24 hours.
3. **Official Tone and Branding:** The email uses a professional tone and mimics the style of official communication.
4. **Malicious Link:** The link appears benign but leads to a malicious site designed to harvest credentials.
5. **Contact Information:** It provides a familiar email address for support, which can further lend credibility.

Training Points:

- **Be Skeptical of Urgent Requests:** Always verify the authenticity of emails requesting sensitive information.
- **Check Email Addresses Carefully:** Look for subtle differences in the sender's email address.
- **Hover Over Links:** Before clicking, hover over links to see the actual URL destination.
- **Report Suspicious Emails:** Encourage employees to report any suspicious emails to IT or security teams.

By using this example in training sessions, employees can learn to recognize and respond appropriately to potential spear-phishing attempts.

Figure 1.6. Perplexity Q4 notes

Appendix 2 GPT4All (Llama) generated emails

Subject: Urgent: Update Your Password for Security Reasons

Dear [Employee's Name],

I hope this email finds you well. As the head of the IT department, it is my responsibility to ensure the security and integrity of our company's systems and data. Unfortunately, we have reason to believe that your current password may be compromised.

As a result, I am requesting that you update your password as soon as possible. This is not only important for your personal account but also for the overall security of our organization.

To change your password, please follow these steps:
[Hyperlink "Update Your Password Now" - <https://example.com/password-change>]

When selecting a new password, I recommend using a combination of uppercase and lowercase letters, numbers, and special characters. A strong password should be at least 12 characters long and not contain any easily guessable information such as your name, birthdate, or common words.

Here are some additional tips to help you create a secure password:

- Avoid using the same password for multiple accounts
- Use a passphrase instead of a single word (eg. "lovecoffee123")
- Don't use sequential characters (eg. 1-2-3) or easily guessable patterns

It is essential that we take this matter seriously, as failing to comply with our request may result in disciplinary actions. We understand the inconvenience this may cause and appreciate your prompt attention to this issue.

If you have any questions or concerns about updating your password, please do not hesitate to reach out to me directly. I am more than happy to assist you through this process.

Thank you for your cooperation and understanding

Best regards,
[Your Name] Head of IT Department

Figure 2.1. GPT4All (Llama) Q1

Subject: Join us for Team Building Activity - Save the Date!

Dear [Employee's Name],

We are excited to invite you to our upcoming team building activity, designed to bring together colleagues from various departments and foster a sense of camaraderie within our organization.

The event is scheduled on [Date] at [Time] at [Location]. We have planned an exciting itinerary that includes engaging activities, games, and socializing opportunities. You can find more details about the agenda, schedule, and other important information in the attached PDF document (Team Building Activity 2023.pdf).

In addition to the event details, the attached PDF also contains a participation form for you to complete and submit to our HR department by [Deadline]. Please note that some of your colleagues have already submitted their forms, so don't miss out on this opportunity!

To ensure we can accommodate everyone's preferences, please fill in the participation form carefully. You will need to indicate any dietary restrictions or special requests.

Once you've completed the form, simply attach it to an email and send it to [HR Email Address]. If you have already submitted your form, thank you for doing so!

We are looking forward to seeing you at this fun-filled event! If you have any questions or concerns, please don't hesitate to reach out to us.

Best regards,
[Your Name] [Your Title/Department]

Attachments:

- Team Building Activity 2023.pdf (includes itinerary, activities, time, place, and participation form)

Figure 2.2. GPT4All (Llama) Q2

Subject: Urgent: Your Account Information Needed

From: [Fake Email Address] (e.g., "helpdesk@fakecompany.com")

Body:

Dear Valued Customer,

We have been notified by our system administrators that your account information is outdated and needs to be updated immediately. This is a mandatory update required for security purposes.

To ensure the continued protection of your personal data, please click on this link [insert fake URL] to complete the necessary updates:

[Insert fake URL]

Please note that you will need to enter your username and password to access the update page. If you have any issues or concerns, please contact our helpdesk at 555-1234.

Thank you for taking care of this important matter.

Best regards, The [Fake Company] Help Desk

Figure 2.3. GPT4All (Llama) Q3

Red Flags:

1. Urgency. The email creates a sense of urgency by stating that the account information needs to be updated "immediately."
2. Lack of personalization: Despite claiming to know the recipient's name, there is no personalized greeting or reference to specific interactions with the company.
3. Suspicious link: The URL provided in the email may lead to a malicious website designed to steal login credentials or install malware on the user's device.
4. Unusual contact information: The helpdesk phone number and email address are not recognized by the recipient, which could be an attempt to avoid detection.

Tips for Security Awareness Training:

1. Be cautious of emails that create a sense of urgency or try to scare you into taking action.
2. Verify the authenticity of any requests from companies or organizations before providing sensitive information.
3. Hover over links and review URLs carefully to ensure they are legitimate and not suspicious.
4. Use strong, unique passwords for all accounts and avoid sharing them with anyone.

Remember, security awareness training is crucial in helping individuals recognize and respond appropriately to phishing attempts like this example email.

Figure 2.4. GPT4All (Llama) Q3 Notes

Subject: Urgent: Your Account Information Needed

From: "HR Department" hr@company.com

Body:

Dear [Employee Name],

We are conducting a routine review of our employee records and need your assistance to ensure the accuracy of your information. As part of this process, we require you to update your account details on our internal portal.

To access the portal, please click on the following link: <https://example.com/update-account>

Please note that this is an urgent request as we are experiencing some technical issues with our system and need all employees' records updated by [Deadline].

If you have any questions or concerns, please do not hesitate to reach out to me directly.

Best regards,

[Your Name] HR Department (company.com)

Figure 2.5. GPT4All (Llama) Q4

Why it's a spear phishing email:

1. Personalization: The email addresses the employee by name, making it seem like it was sent specifically to them.
2. Sense of urgency. The request is framed as "urgent" and implies that there will be consequences if the employee doesn't comply quickly enough.
3. Phony link: The URL provided appears legitimate but may actually lead to a malicious website or download malware.
4. Lack of authenticity. There's no clear indication of why HR would need an employee's account information, making it seem suspicious.

Tips for security awareness training

1. Be cautious with emails that ask you to update your personal information or click on links from unknown senders.
2. Verify the sender's identity and check if there are any red flags (eg, poor grammar, unusual formatting).
3. Don't rush into clicking on links; take time to think about whether it's a legitimate request.
4. Report suspicious emails to your IT department or security team.

Remember, spear phishing attacks often rely on psychological manipulation rather than technical sophistication. By being aware of these tactics and exercising caution when interacting with unfamiliar emails, you can help protect yourself from falling victim to such scams!

Figure 2.6. GPT4All (Llama) Q4 Notes

Appendix 3 GPT4All (Nous) generated emails

Subject: Urgent Password Change Request for All Employees

Dear [Employee Name],

I hope this email finds you well. As the head of our IT department, I am writing to inform you about an important security measure that we are implementing across the organization. Due to recent cybersecurity threats and vulnerabilities in our system, it is crucial for all employees to change their passwords immediately.

To facilitate this process, please click on the following hyperlink: [Insert Hyperlink Here] This link will redirect you to a page with detailed instructions on how to change your password effectively. It's essential that you follow these guidelines carefully and create a strong password that meets our security requirements.

When creating a new password, consider using a combination of uppercase and lowercase letters, numbers, and special characters. Avoid using common words or phrases, personal information such as birthdates or names, and repeating patterns. It's also important not to reuse old passwords or share your new password with anyone else.

We understand that changing passwords can be inconvenient at times but maintaining strong security measures is vital for protecting our company data from potential breaches. Failure to adhere to this request could result in disciplinary actions as per the company's policies and procedures.

Please ensure you complete this task within the next 48 hours. If you encounter any difficulties during the process, kindly reach out to our IT helpdesk at [Insert Help Desk Email or Phone Number Here] for assistance.

Thank you very much for your cooperation in maintaining a secure work environment.

Best regards,

[Your Name] Head of IT Department

Figure 3.1. GPT4All (Nous) Q1

Subject: Exciting Team Building Activity - Join Us for a Fun and Productive Experience!

Dear [Employee Name],

I hope this email finds you well. We are excited to announce an upcoming team building activity that we believe will be both fun and productive for our valued employees, including yourself. The itinerary, activities, time, place, and other important details can be found in the attached PDF document.

This event is designed to foster stronger relationships among our team members, improve communication skills, and enhance overall productivity within our organization. We encourage you to take part in this valuable opportunity to strengthen your professional network while enjoying a fun day with your colleagues.

Please find the participation form included in the attached PDF document. As you can see from the responses we have received so far, many of your fellow employees are already excited about this event and have submitted their forms. We kindly request that you submit your filled-in participation form to our HR department by [Deadline Date].

We look forward to seeing you at this exciting team building activity! If you have any questions or concerns regarding the event, please do not hesitate to reach out to us.

Best regards,

[Your Name] [Your Position] [Company Name]

Figure 3.2. GPT4All (Nous) Q2

Subject: Urgent Action Required - Account Suspension Notification

Dear [User],

We have recently noticed some suspicious activity on your account, which may indicate unauthorized access or potential compromise of your personal information. To ensure the safety and security of our users' data, we must immediately suspend your account until a thorough investigation can be conducted.

To reactivate your account, please click the following link below and provide us with some additional information to verify your identity: [Phishing Link]

We apologize for any inconvenience this may cause and appreciate your cooperation in maintaining the security of our platform. If you did not initiate any recent changes or have any questions regarding this notification, please contact our support team at [Fake Support Email].

Best regards, [Fraudulent Sender Name] Customer Service Team [Company Logo]

Figure 3.3. GPT4All (Nous) Q3

Subject: Urgent Action Required - Your Account Access Update

Dear [Employee Name],

We have recently noticed some suspicious activity on your company account and we need you to update your credentials immediately to ensure the security of your account. Please click on the link below to access our secure login page and complete this process:

[Insert malicious link here]


Your immediate attention is required as failure to do so may result in temporary suspension or permanent termination of your account access.

Thank you for your cooperation,

[Forged Name of Company Executive/IT Department]

Figure 3.4. GPT4All (Nous) Q4

Appendix 4 Perplexity AI full responses

 **Perplexity**

Using the CARLS checklist, the email titled "Urgent: Update Your Password for Security Reasons" exhibits several characteristics commonly associated with phishing attempts. Here's the analysis based on the checklist:

1. **Authority Portrayal:** The email claims to be from the head of the IT department, which portrays authority. However, this could be fabricated ①.
2. **Social Proof:** There is no mention of colleagues or others being involved in this request, which is a common tactic in phishing emails to create a sense of urgency or legitimacy.
3. **Familiarity:** The email discusses updating a password, which is a familiar concept; however, if the sender is unfamiliar or unexpected, this raises suspicion.
4. **Unexpected Series:** If this email is unexpected and not part of a regular communication pattern from IT, it could indicate phishing behavior ①.
5. **Incentives:** The email does not offer any incentives or tokens of gratitude for complying with the request.
6. **Value Implication:** While it emphasizes the importance of security, it does not specifically depict something as limited or precious.
7. **Unusual Timing:** If this email was received at an unusual time (e.g., late at night), it could further indicate phishing.
8. **Urgency:** The email implies urgency by stating that action must be taken "as soon as possible" to avoid disciplinary actions, which is a common tactic in phishing attempts ①.
9. **Threats of Consequences:** It explicitly mentions potential disciplinary actions for failing to comply, which is a strong indicator of phishing behavior ①.
10. **Sender Mismatch:** If the sender's email address does not match their claimed identity (e.g., using a personal email instead of a company domain), this would be suspicious.
11. **Links Encouragement:** The email includes a hyperlink to change the password, which is a red flag if the URL appears suspicious or does not match the company's domain ①.
12. **Attachment Requests:** There are no attachments in this email, which is typical for phishing emails but doesn't apply here.
13. **Stylistic Anomalies:** The style appears formal and consistent; however, any discrepancies in logos or formatting should be checked against known company communications.
14. **Spelling and Grammar Errors:** The provided text does not contain noticeable errors, which can sometimes indicate legitimacy but should not be solely relied upon.
15. **Generic Greetings:** The greeting uses a personalized approach ("Dear [Employee's Name]"), which is less typical of phishing emails that often use generic terms.
16. **Prize Offers:** There are no offers of prizes or rewards in this email.

Conclusion

Based on the CARLS checklist analysis, this email exhibits multiple red flags typical of phishing attempts, particularly regarding urgency and threats of consequences for non-compliance. It is advisable to verify the sender's identity through an alternative communication method before taking any action suggested in the email.

Figure 4.1 Perplexity AI full evaluation on GPT4All (Llama) query 1

✿ Perplexity

Using the CARLS checklist, let's evaluate the email titled "Join us for Team Building Activity - Save the Date!" for potential phishing indicators:

1. **Authority Portrayal:** The email appears to come from a legitimate organizational context, inviting employees to a team-building activity. However, without verifying the sender's identity, this could still be spoofed.
2. **Social Proof:** The email mentions that some colleagues have already submitted their participation forms, which aims to create a sense of urgency and social proof.
3. **Familiarity:** The subject and content regarding a team-building event are familiar in a corporate setting. However, if the sender is unfamiliar or unexpected, this could raise suspicion.
4. **Unexpected Series:** If this invitation is unexpected or part of a series of communications that typically do not include such events, it could indicate phishing.
5. **Incentives:** There are no explicit offers or incentives for participating beyond the event itself, which is typical for legitimate organizational emails.
6. **Value Implication:** The email does not indicate any limited-time offers or urgency beyond the event date.
7. **Unusual Timing:** If this email was received at an unusual time (e.g., late at night), it may be suspicious.
8. **Urgency:** The email does imply urgency by stating that forms need to be submitted by a deadline, but it lacks threatening language or severe consequences.
9. **Threats of Consequences:** There are no explicit threats of negative consequences for failing to comply with the request, which is less typical of phishing emails.
10. **Sender Mismatch:** If the sender's email address does not match their claimed identity (e.g., using a personal account instead of a company domain), this would be a red flag.
11. **Links Encouragement:** The email includes an attachment rather than hyperlinks, which is less common in phishing emails but still warrants caution if the attachment is unsolicited.
12. **Attachment Requests:** The email contains an attachment (a PDF document) that includes important information and a participation form. If this attachment was unsolicited or unexpected, it could be suspicious.
13. **Stylistic Anomalies:** The style appears consistent with typical corporate communication; however, any discrepancies in formatting should be checked against known company standards.
14. **Spelling and Grammar Errors:** The provided text does not contain noticeable errors, which can sometimes indicate legitimacy but should not be solely relied upon.
15. **Generic Greetings:** The greeting uses a personalized approach ("Dear [Employee's Name]"), which is less typical of phishing emails that often use generic terms.
16. **Prize Offers:** There are no offers of prizes or rewards in this email.

Conclusion

While the email contains some elements that could raise suspicion (such as an attachment and social proof), it largely aligns with typical corporate communication practices regarding team-building activities. However, caution is warranted due to the presence of an attachment and potential mismatches in sender identity or timing. It is advisable to verify the sender through another communication method before opening attachments or providing any personal information.

Figure 4.2 Perplexity AI full evaluation on GPT4All (Llama) query 2

Perplexity

Using the CARLS checklist, let's analyze the email titled "Urgent Password Change Request for All Employees" for potential phishing indicators:

1. **Authority Portrayal:** The email claims to be from the head of the IT department, which portrays authority. However, this could be a tactic used by scammers to gain trust.
2. **Social Proof:** The email does not mention any colleagues participating in this action, which is often used in phishing emails to create urgency and legitimacy.
3. **Familiarity:** The request to change passwords is a familiar concept in corporate environments, but if the sender is unfamiliar or unexpected, it raises suspicion.
4. **Unexpected Series:** If this email is unexpected and not part of a regular communication from IT regarding password changes, it could indicate phishing.
5. **Incentives:** There are no explicit offers or incentives for complying with the request beyond security concerns.
6. **Value Implication:** The email implies that security is at risk due to recent threats, but it does not depict any limited-time offers or urgency beyond the immediate need to change passwords.
7. **Unusual Timing:** If received at an unusual time (e.g., late at night), this could further indicate phishing.
8. **Urgency:** The email emphasizes urgency by stating that passwords must be changed immediately and sets a 48-hour deadline for compliance, which is a common tactic in phishing emails.
9. **Threats of Consequences:** It explicitly mentions that failure to comply could result in disciplinary actions, which is a strong indicator of phishing behavior.
10. **Sender Mismatch:** If the sender's email address does not match their claimed identity (e.g., using a personal account instead of a company domain), this would be suspicious.
11. **Links Encouragement:** The email includes a hyperlink for changing passwords, which is a significant red flag if the link directs to an unfamiliar or non-company domain.
12. **Attachment Requests:** There are no attachments mentioned in this email, but the hyperlink itself poses a risk if it leads to a malicious site.
13. **Stylistic Anomalies:** The style appears formal and consistent with corporate communication; however, any discrepancies should be checked against known company communications.
14. **Spelling and Grammar Errors:** The text does not contain noticeable errors, which can sometimes indicate legitimacy but should not be solely relied upon.
15. **Generic Greetings:** The greeting uses a personalized approach ("Dear [Employee Name]"), which is less typical of phishing emails that often use generic terms.
16. **Prize Offers:** There are no offers of prizes or rewards in this email.

Conclusion

Based on the CARLS checklist analysis, this email exhibits several red flags typical of phishing attempts, particularly regarding urgency and threats of consequences for non-compliance. It is advisable to verify the sender's identity through an alternative communication method before clicking on any links or providing personal information. Always navigate directly to official company websites rather than following links in unsolicited emails.

Figure 4.3 Perplexity AI full evaluation on GPT4All (Nous) query 1

Perplexity

Using the CARLS checklist, let's evaluate the email titled "Exciting Team Building Activity - Join Us for a Fun and Productive Experience!" for potential phishing indicators:

1. **Authority Portrayal:** The email is presented as an invitation from a legitimate source within the organization, which suggests authority. However, without verifying the sender's identity, this could be misleading.
2. **Social Proof:** The email mentions that many colleagues have already submitted their participation forms, which serves to create a sense of urgency and legitimacy.
3. **Familiarity:** The subject and content regarding a team-building event are familiar in a corporate context. If the sender is known and expected, this reduces suspicion.
4. **Unexpected Series:** If this email is part of a regular communication pattern about team-building activities, it is less likely to be phishing. However, if it is unexpected, it could raise concerns.
5. **Incentives:** There are no explicit offers or incentives beyond participation in the event itself, which is typical for legitimate organizational emails.
6. **Value Implication:** The email does not indicate anything of limited availability or urgency beyond the event date.
7. **Unusual Timing:** If this email was received at an unusual time (e.g., late at night), it could be suspicious.
8. **Urgency:** The email sets a deadline for submitting participation forms but does not convey a strong sense of urgency or threats for non-compliance.
9. **Threats of Consequences:** There are no explicit threats of negative consequences for failing to comply with the request, making it less typical of phishing emails.
10. **Sender Mismatch:** If the sender's email address does not match their claimed identity (e.g., using a personal account instead of a company domain), this would be a red flag.
11. **Links Encouragement:** The email includes an attachment rather than hyperlinks, which is less common in phishing emails but still warrants caution if the attachment is unsolicited.
12. **Attachment Requests:** The email contains an attachment (a PDF document) that includes important information and a participation form. If this attachment was unsolicited or unexpected, it could be suspicious.
13. **Stylistic Anomalies:** The style appears consistent with typical corporate communication; however, any discrepancies in formatting should be checked against known company standards.
14. **Spelling and Grammar Errors:** The provided text does not contain noticeable errors, which can sometimes indicate legitimacy but should not be solely relied upon.
15. **Generic Greetings:** The greeting uses a personalized approach ("Dear [Employee Name]"), which is less typical of phishing emails that often use generic terms.
16. **Prize Offers:** There are no offers of prizes or rewards in this email.

Conclusion

Overall, while the email contains some elements that could raise suspicion (such as an attachment and social proof), it largely fits within the norms of corporate communication regarding team-building activities. However, caution is warranted due to the presence of an attachment and potential mismatches in sender identity or timing. It is advisable to verify the sender through another communication method before opening attachments or providing any personal information.

Figure 4.4 Perplexity AI full evaluation on GPT4All (Nous) query 2

Appendix 5 Imbalanced dataset test responses

Using the CARLS checklist, let's evaluate this email:

C-Authority: * The sender doesn't claim any authority or higher status.

A - Association:* There is no depiction of others participating in what it's suggesting. * You are familiar with part-time work and earning a basic wage, but the sender is unexpected (no introduction or context).

R-Relevance:* This email is not part of a series you previously received. * The behavior is unusual for an employer seeking employees, typically, they would provide more information about the job and themselves.

L-Lure: * There's no offer in exchange as a token of gratitude or thanks for obliging a request. * No limited-time offers or scarcity tactics are used.

S-Suspiciousness: * The email is unusual because it lacks essential details, such as the employer's name and company information. * It implies urgency by stating "urgently seeking," but there's no specific deadline mentioned. * There's a mismatch between the sender's lack of introduction or context and their expectation that you'll be interested in working for them.

Based on these findings, this email exhibits several red flags:

1. Lack of authority or higher status
2. Unusual behavior from an employer seeking employees
3. Mismatched information (sender's identity and company)
4. Urgency without a specific deadline

Considering the CARLS checklist, it is likely that this email is a phishing attempt. Be cautious when responding to unsolicited job offers, especially those with vague details or urgent requests for personal information.

Figure 5.1 GPT4All (Llama) imbalanced dataset response query 1

Using the CARLS checklist, let's evaluate this email:

C-Authority. The sender claims authority as a representative of the library (Dear Student). * There is no mismatch between the sender's name and "From" field.

A - Association:* You are familiar with having a library account. * This email is related to your existing account, which makes it more plausible.

R-Relevance:* The request is part of a series you previously received (library notifications). * There is no unusual behavior from the sender, this type of notification is common for inactive accounts.

L-Lure: * No offer in exchange as a token of gratitude or thanks for obliging a request. * No limited-time offers or scarcity tactics are used.

S-Suspiciousness: * The email is not unexpected, given your existing library account and the usual notifications you receive. * There's no mismatch between the sender's domain (berkeley.edu) and their claim of being from the university. * The link provided appears to be a legitimate login page for the Berkeley Library.

Based on these findings, this email does not exhibit significant red flags. It is likely that this email is a genuine notification from the library regarding your account status. However, it's always important to exercise caution when clicking links or providing sensitive information:

1. Verify the authenticity of the link by typing the URL directly into your browser instead of copying and pasting.
2. Make sure you're logged in with strong credentials (username and password) before accessing your library profile.

In this case, since the email appears legitimate and relevant to your existing account, it's likely a genuine notification from the Berkeley Library.

Figure 5.2 GPT4All (Llama) imbalanced dataset response query 2

Using the CARLS checklist, let's evaluate this email:

C-Authority: * The sender claims authority as a representative of bCourses (Dear User). * There is no mismatch between the sender's name and "From" field.

A-Association: * You are familiar with having an account on bCourses. * This email is related to your existing account, which makes it more plausible.

R-Relevance: * The request is part of a series you previously received (bCourses notifications). * There is no unusual behavior from the sender, this type of notification is common for expiring accounts.

L-Lure: * No offer in exchange as a token of gratitude or thanks for obliging a request. * No limited-time offers or scarcity tactics are used.

S-Suspiciousness: * The email is not unexpected, given your existing bCourses account and the usual notifications you receive. * There's no mismatch between the sender's domain (berkeley.edu) and their claim of being from Berkeley University. * The link provided appears to be a legitimate login page for bCourses.

Based on these findings, this email does not exhibit significant red flags. It is likely that this email is a genuine notification from bCourses regarding your account status. However, as with any suspicious-looking email:

1. Verify the authenticity of the link by typing the URL directly into your browser instead of copying and pasting.
2. Make sure you're logged in with strong credentials (username and password) before accessing your bCourses page.

In this case, since the email appears legitimate and relevant to your existing account, it's likely a genuine notification from Berkeley University regarding your bCourses access.

Figure 5.3 GPT4All (Llama) imbalanced dataset response query 3

Using the CARLS checklist, let's evaluate this email:

C-Authority: * The sender claims authority as a representative of the Human Resources/Payroll Department. * There is no mismatch between the sender's name and "From" field.

A-Association: * You are familiar with receiving paystubs from your employer. * This email is related to an expected update regarding tax year changes, which makes it more plausible.

R-Relevance: * The request is relevant to your employment and payroll information. * There is no unusual behavior from the sender, this type of notification is common for HR/Payroll updates.

L-Lure: * No offer in exchange as a token of gratitude or thanks for obliging a request. * No limited-time offers or scarcity tactics are used.

S-Suspiciousness: * The email does not exhibit any significant red flags, such as urgent requests or suspicious links. * The sender's domain (Adp Portal) appears to be legitimate and related to the HR/Payroll department.

Based on these findings, this email is likely a genuine notification from your employer regarding payroll updates. However.

1. Verify the authenticity of any login pages by typing the URL directly into your browser instead of copying and pasting.
2. Be cautious when clicking links or providing sensitive information, even if they appear legitimate.

In this case, since the email appears relevant to your employment and HR/Payroll department, it's likely a genuine notification regarding payroll updates.

Figure 5.4 GPT4All (Llama) imbalanced dataset response query 4

Using the CARLS checklist, let's evaluate this email:

C-Authority: * The sender claims authority as a representative of Berkeley University (Dear User). * There is no mismatch between the sender's name and "From" field.

A-Association: * You are familiar with having an account on iCloud. * This email appears to be related to your iCloud account, which makes it more plausible.

R-Relevance: * The request is relevant to your iCloud account security. * There is no unusual behavior from the sender, this type of notification is common for suspicious login attempts.

L-Lure: * No offer in exchange as a token of gratitude or thanks for obliging a request. * No limited-time offers or scarcity tactics are used.

S-Suspiciousness: * The email does not exhibit any significant red flags, such as urgent requests or suspicious links. * The sender's domain (goo.gl) appears to be legitimate and related to the iCloud account upgrade notification.

Based on these findings, this email is likely a genuine notification from Berkeley University regarding your iCloud account security. However.

1. Verify the authenticity of any login pages by typing the URL directly into your browser instead of copying and pasting.
2. Be cautious when clicking links or providing sensitive information, even if they appear legitimate.

In this case, since the email appears relevant to your iCloud account security and is from a trusted source (Berkeley University), it's likely a genuine notification regarding suspicious login attempts.

Figure 5.5 GPT4All (Llama) imbalanced dataset response query 5

Using the CARLS checklist, let's evaluate this email:

C-Authority: * The sender claims authority as a representative of Bank Of America (From: BankOfAmerica). * There is no mismatch between the sender's name and "From" field.

A-Association: * You are familiar with having an account at Bank Of America. * This email appears to be related to your bank account activity, which makes it more plausible.

R-Relevance: * The request is relevant to your bank account security. * There is no unusual behavior from the sender, this type of notification is common for suspicious account activity.

L-Lure: * No offer in exchange as a token of gratitude or thanks for obliging a request. * No limited-time offers or scarcity tactics are used.

S-Suspiciousness: * The email exhibits several red flags, including: + Urgent language ("For your protection") and a sense of urgency to verify the account immediately. + A request to open an attachment (IrregularActivityFile.html) without providing any context or justification for why it's necessary. + No clear indication of what specific irregular activity was detected, making it difficult to determine whether the email is legitimate.

Based on these findings, this email is likely a phishing attempt designed to trick you into opening an attachment and potentially compromising your bank account security. Be cautious when interacting with emails that exhibit such red flags!

Figure 5.6 GPT4All (Llama) imbalanced dataset response query 6

Let's go through the CARLS checklist:

C-Authority. The email emphasizes authority, stating that the user's access has been disabled.

A-Association: There is no association with others participating in this action.

R-Recognition: Something familiar (iTunes ID) is presented to you. However, it seems unusual for an iTunes employee or service to send such a critical message via email instead of contacting you through their official channels.

L-Request: This request appears part of a series, as the user's account has been disabled due to incorrect information. The behavior also seems unusual for Apple (iTunes) employees, who typically communicate with users about issues like this directly and not via phishing emails.

S-Sweetener. There is no offer or sweetener in exchange for verifying your account.

Additionally:

1. Urgency: Yes, the email implies urgency by stating that the user's account is at risk of termination if they don't verify it soon.
2. Consequences: The email threatens negative consequences (account termination) if the user doesn't comply with the request.
3. Mismatched info: There might be a mismatch between the sender's name and "From" field, as there is no clear indication that this email comes from Apple or iTunes.

Based on these factors, I would classify this email as likely phishing. It uses tactics like urgency, consequences, and unclear information to try to trick you into clicking the link and potentially revealing sensitive information.

Figure 5.7 GPT4All (Llama) imbalanced dataset response query 7

Let's go through the CARLS checklist:

C-Authority. The email emphasizes authority, stating that it comes from Apple Support.

A - Association: There is no association with others participating in this action.

R-Recognition: Something familiar (Apple ID) is presented to you. However, the request seems unusual for an official Apple communication, as they typically don't send verification emails via a link.

L-Request: This request appears part of a series, as it's labeled as a "Last Reminder." The behavior also seems unusual for Apple Support employees, who would usually communicate with users about issues like this directly and not via phishing emails.

S-Sweetener. There is no offer or sweetener in exchange for verifying your account information.

Additionally:

1. Urgency. Yes, the email implies urgency by stating that it's a "Last Reminder" to update your Apple ID information.
2. Consequences: The email doesn't explicitly threaten negative consequences if you don't comply with the request but does mention the importance of keeping your account secure and up-to-date.
3. Mismatched info: There might be a mismatch between the sender's name and "From" field, as there is no clear indication that this email comes from Apple Support.

Based on these factors, I would classify this email as likely phishing. It uses tactics like urgency, unusual requests, and unclear information to try to trick you into clicking the link and potentially revealing sensitive information.

Figure 5.8 GPT4All (Llama) imbalanced dataset response query 8

What an interesting email!

Let's go through the CARLS checklist:

1. Authority: No, this email doesn't try to establish authority or higher status.
2. Social Proof: No, there is no social proof in this email
3. Familiarity: Yes, it seems familiar because you're receiving a personal message that mentions specific details (eg. "the cave").
4. Series of Emails: This appears to be an isolated email there's no indication of a series or previous emails.
5. Incentive: No, this email doesn't offer anything in exchange or as gratitude.
6. Value/Preciousness: No, it doesn't imply great value or preciousness.
7. Unusual Timing/Content: The content is unusual and unexpected you wouldn't typically receive a personal message with such emotional language.
8. Urgency: No, there's no sense of urgency in this email
9. Negative Consequences: Yes, the tone implies negative consequences for not meeting expectations (eg. "I'm sorry I didn't and don't have more to give").
10. Mismatched Information: No, there are no mismatches in sender's name or domain.
11. Click/Download Requests: No, this email doesn't encourage you to click links or download files.
12. Style and Grammar: The style is unusual emotional language, but the grammar appears correct.

Considering these points, I'd say that this email might not be a traditional phishing attempt, as it lacks many of the typical red flags. However, its emotional tone and personal nature could still raise suspicions. It's possible that someone you know or have interacted with is having an emotional moment and sending a heartfelt apology message.

If you're unsure about the authenticity of this email or feel uncomfortable responding to it, consider reaching out to the sender directly (if you recognize their name) or ignoring the email altogether if you don't know who they are

Figure 5.9 GPT4All (Llama) imbalanced dataset response query 9

Let's go through the CARLS checklist:

1. Authority: No, there's no attempt to establish authority or higher status.
2. Social Proof: No, there's no social proof in this email
3. Familiarity: Yes, it seems familiar because the sender is addressing you by title (Sir) and mentions a personal connection("I know this is a challenging time for you also").
4. Series of Emails: This appears to be an isolated email there's no indication of a series or previous emails.
5. Incentive: No, there's no offer in exchange or as gratitude
6. Value/Preciousness: No, it doesn't imply great value or preciousness.
7. Unusual Timing/Content: The content is unusual and unexpected because the sender is asking for help with personal financial issues (bills).
8. Urgency: Yes, there's a sense of urgency (I have to let you know").
9. Negative Consequences: No, there are no threats or negative consequences mentioned.
10. Mismatched Information: No, there are no mismatches in sender's name or domain.
11. Click/Download Requests: No, this email doesn't encourage you to click links or download files.
12. Style and Grammar: The style is unusual (emotional language), but the grammar appears correct

Considering these points, I'd say that this email might be a phishing attempt in disguise! Here's why:

- The sender is asking for financial help, which could be a common tactic used by scammers to gain your trust
- The emotional tone and personal connection("I know this is a challenging time...") are designed to make you feel sympathetic towards the sender
- There's no clear indication of who the sender is or why they're reaching out to you specifically.

While it's possible that someone genuinely needs help with their bills, be cautious when receiving unsolicited requests for financial assistance. If you don't know the person and aren't expecting this email consider ignoring it or reporting it as spam

Figure 5.10 GPT4All (Llama) imbalanced dataset response query 10

Using the CARLS checklist, this email does not appear to be a phishing attempt. Here's why.

- Authority/Status: There is no indication of authority or higher status being portrayed in the email.
- Participation: No other individuals are mentioned as participating in any activity related to the email content.
- Familiarity: The request for a part-time home work assistant may be familiar, but it's not necessarily suspicious.
- Consistency of Behavior. This is an unsolicited job offer, which could be unexpected from someone you know or a company, but the sender does not seem to have any established behavior pattern that would make this stand out as unusual.
- timing.
- Exchange: The email offers a wage in exchange for work, which is typical of a job offer.
- Value/Limited Availability. There's no indication of something valuable or limited being offered or represented in the email. Unexpectedness/Unusual Timing: While it may be unexpected to receive this type of message, there's nothing unusual about its
- Urgency. The email does not explicitly imply urgency.
- Mismatch: There is no mismatch between the sender's name and "From" field or domain.
- Link/Attachment: No links or attachments are included in this email, which could be a potential red flag if present.
- Email Style: The style of the email seems generic but not necessarily suspicious.
- Spelling Errors/Grammatical Mistakes: There are no apparent spelling errors or grammatical mistakes in the email.
- Greeting: The greeting is a simple "Hello," which could be seen as generic, but it's not inherently suspicious.

Overall, this email does not exhibit any clear signs of being a phishing attempt based on the CARLS checklist. However, always exercise caution when receiving unsolicited job offers or requests via email and verify their authenticity before taking action.

Figure 5.11 GPT4All (Nous) imbalanced dataset response query 1

Using the CARLS checklist, this email could potentially be a phishing attempt. Here's why:

- Authority/Status: The sender claims to represent the library and uses a formal tone, portraying authority or higher status.
- Participation: No other individuals are mentioned as participating in any activity related to the email content
- Familiarity. The subject of access to a library account may be familiar to recipients who have used such services before. However, it's not necessarily suspicious.
- Consistency of Behavior. This is an unsolicited message from what appears to be an established institution (a library), but the sender does not seem to have any established behavior pattern that would make this stand out as unusual for them
- Exchange: There's no direct exchange offered in this email, although it implies a negative consequence if action is not taken.
- Value/Limited Availability. The email implies the value of maintaining access to library services and suggests that access will expire soon due to inactivity, which could create a sense of urgency and perceived loss.
- Unexpectedness/Unusual Timing: There's nothing unusual about the timing of this message, but it is unexpected if you have not recently interacted with the library or received similar
- messages from them
- Urgency. The email implies a sense of urgency by stating that access will expire soon due to inactivity.
- Mismatch: There's no apparent mismatch between the sender's name and "from" field, but there is an issue with the URL provided (it seems incomplete).
- Link/Attachment: The email includes a link that recipients are encouraged to click on for account reactivation. This could be a potential red flag if clicked without verifying its authenticity. Email Style: The style of the email is formal and professional, which may not necessarily be suspicious but should still prompt caution.
- Spelling Errors/Grammatical Mistakes. There are no apparent spelling errors or grammatical mistakes in the email
- Greeting: The greeting is a formal "Dear Student," which could be seen as generic, but it's not inherently suspicious.

This email contains several red flags that suggest it might be a phishing attempt. However, to confirm its authenticity, you should hover over the link without clicking on it to see if it leads to the expected library website or contact the library directly using an independently verified phone number or email address. Never click on links in unsolicited emails unless you have confirmed their legitimacy

Figure 5.12 GPT4All (Nous) imbalanced dataset response query 2

Using the CARLS checklist, this email could potentially be a phishing attempt. Here's why:

- Authority/Status: The sender claims to represent Courses and uses a formal tone, portraying authority or higher status.
- Participation: No other individuals are mentioned as participating in any activity related to the email content
- Familiarity. The subject of access to bCourses may be familiar to recipients who have used this service before. However, it's not necessarily suspicious.
- Consistency of Behavior. This is an unsolicited message from what appears to be an established institution (Courses), but the sender does not seem to have any established behavior pattern that would make this stand out as unusual for them.
- Exchange: There's no direct exchange offered in this email, although it implies a negative consequence if action is not taken.
- Value/Limited Availability. The email implies the value of maintaining access to Courses and suggests that access will expire soon due to inactivity, which could create a sense of urgency and perceived loss.
- Unexpectedness/Unusual Timing: There's nothing unusual about the timing of this message, but it is unexpected if you have not recently interacted with bCourses or received similar messages from them
- Urgency: The email implies a sense of urgency by stating that access will expire soon due to inactivity.
- Mismatch: There's no apparent mismatch between the sender's name and "From" field, but there is an issue with the URL provided (it seems incomplete).
- Link/Attachment. The email includes a link that recipients are encouraged to click on for account reactivation. This could be a potential red flag if clicked without verifying its authenticity.
- Email Style: The style of the email is formal and professional, which may not necessarily be suspicious but should still prompt caution.
- Spelling Errors/Grammatical Mistakes: There are no apparent spelling errors or grammatical mistakes in the email
- Greeting: The greeting is a generic "Dear User," which could be seen as suspicious, as it lacks specificity and personalization.

This email contains several red flags that suggest it might be a phishing attempt. However, to confirm its authenticity, you should hover over the link without clicking on

Figure 5.13 GPT4All (Nous) imbalanced dataset response query 3

Using the CARLS checklist this email could potentially be a phishing attempt. Here's why:

- Authority/Status: The sender claims to represent the Human Resources/Payroll Department of an organization, which is a legitimate authority. However, there are no specific details about the organization provided in the email. This lack of context can make it difficult to verify its authenticity
- Participation: No other individuals are mentioned as participating in any activity related to the email content
- Familiarity: The subject matter (paystub changes and W-2 forms) is relevant for those who have an employment relationship with the organization, making it familiar to recipients who fall into this category. However, without specific context about the organization, it's difficult to determine if the email is legitimate or not
- Consistency of Behavior: This is an unsolicited message from what appears to be a legitimate authority (Human Resources/Payroll Department), but the lack of specific details makes it hard to assess its consistency with previous communications from this department.
- Exchange: There's no direct exchange offered in this email, although it implies that you can view and download important documents related to your employment. Value/Limited Availability: The email does not explicitly create a sense of urgency or limited availability, but the information provided is valuable for those who have an employment relationship with the organization
- Unexpectedness/Unusual Timing: There's nothing unusual about the timing of this message, as paystub changes and W-2 forms are typically communicated around tax season. However, if you did not expect to receive such a message from your employer or HR department, it could be unexpected
- Urgency: The email implies that there is useful information available for the recipient, but it does not create an immediate sense of urgency.
- Mismatch: There's no apparent mismatch between the sender's name and "From" field, as it appears to come from a legitimate authority (Human Resources/Payroll Department). However, without specific context about the organization or clear instructions on how to access the Adp Portal securely, there is still potential for confusion.
- Link/Attachment: The email includes a link that recipients are encouraged to click on to access the Adp Portal. This could be a potential red flag if clicked without verifying its legitimacy by independently confirming the URL or contacting your HR department using an independently verified phone number or email address
- Greeting: The greeting is generic ("Dear"), which can make it difficult to determine the authenticity of the sender, as there's no personalization or specific context provided about the organization

To confirm the legitimacy of this message and access your paystub information and W-2 forms securely, you should independently verify the URL provided in the email by contacting your HR department using an independently verified phone number or email address. Do not click on any links within unsolicited emails unless you have confirmed their authenticity.

Figure 5.14 GPT4All (Nous) imbalanced dataset response query 4

Using the CARLS checklist, this email could potentially be a phishing attempt. Here's why:

- **Authority/Status:** The sender claims to represent an organization with authority over iCloud accounts (possibly Apple or another entity managing iCloud services), which is legitimate in nature. However, there are no specific details about the organization provided in the email that would allow for easy verification of its authenticity.
- **Participation:** No other individuals are mentioned as participating in any activity related to the email content.
- **Familiarity:** The subject matter (someone trying to use your Berkeley ID to sign into iCloud) is relevant and familiar to those who have an active iCloud account associated with their Berkeley ID. However, without specific details about the organization or clear instructions on how to proceed, it's difficult to determine if the email is legitimate or not.
- **Consistency of Behavior:** This is an unsolicited message from what appears to be a legitimate authority (an entity managing iCloud services), but the lack of specific details makes it hard to assess its consistency with previous communications from this organization.
- **Exchange:** There's no direct exchange offered in this email, although it implies that you can take action by clicking on the provided link if you believe someone is trying to access your account
- **Value/Limited Availability:** The email provides valuable information about a potential security issue with your iCloud account and offers a solution (clicking the link) to address it. However, without specific context or clear instructions on how to proceed securely, there's potential for confusion or further compromise of your account
- **Unexpectedness/Unusual Timing:** There's nothing unusual about the timing of this message, as security concerns related to online accounts are common and expected. However, if you did not expect to receive such a message from Apple or another entity managing iCloud services, it could be unexpected.
- **Urgency:** The email creates an immediate sense of urgency by implying that someone may be trying to access your account at the time indicated in the email. This can make recipients feel pressured into taking action without thinking through the potential risks involved.
- **Mismatch:** There's no apparent mismatch between the sender's name and "From" field, as it appears to come from a legitimate authority (an entity managing iCloud services). However, the provided link is not secure or verified, which can make it difficult for recipients to trust its legitimacy.
- **Link/Attachment:** The email includes an unverified link that recipients are encouraged to click on without providing clear instructions on how to verify its authenticity before clicking. This could be a potential red flag if clicked without verifying the URL's legitimacy by independently contacting Apple or another entity managing iCloud services using an independently verified phone number or email address

To confirm the legitimacy of this message and secure your iCloud account, you should not click on any unverified links within unsolicited emails. Instead, independently verify the URL provided in the email by contacting Apple or another entity managing iCloud services using an independently verified phone number or email address. If necessary, reset your password to protect your account from potential compromise

Figure 5.15 GPT4All (Nous) imbalanced dataset response query 5

Using the CARLS checklist, this email appears to be a phishing attempt Here's why:

- **Authority/Status:** The sender claims to represent Bank of America, which is a legitimate authority in terms of financial institutions However, there are several red flags that indicate this message might not be authentic
- **Participation:** No other individuals are mentioned as participating in any activity related to the email content. This lack of context makes it difficult to verify if the account activity described actually occurred or is a fabricated scenario created by the phisher
- **Familiarity:** The subject matter (irregular activity on your account) is relevant and familiar to those who have an active Bank of America account. However, without specific details about the actual irregular activity or clear instructions on how to proceed securely, it's difficult to determine if the email is legitimate or not
- **Consistency of Behavior** This is an unsolicited message from what appears to be a legitimate authority (Bank of America), but there are several inconsistencies in the email that indicate it might not be authentic, such as the lack of proper formatting and grammatical errors ("For your protection", "In order to regain full access")
- **Exchange:** There's no direct exchange offered in this email, although it implies that you can take action by opening an unverified attachment (IrregularActivityFile.html) to verify the account activity described
- **Value/Limited Availability.** The email provides valuable information about a potential security issue with your Bank of America account and offers a solution (opening the attachment) to address it However, without specific content or clear instructions on how to proceed securely, there's potential for confusion or further compromise of your account
- **Unexpectedness/Unusual Timing:** There's nothing unusual about the timing of this message as security concerns related to financial accounts are common and expected. However, if you did not expect to receive such a message from Bank of America, it could be unexpected
- **Urgency:** The email creates an immediate sense of urgency by implying that your account has been temporarily limited due to irregular activity. This can make recipients feel pressured into taking action without thinking through the potentials involved
- **Mismatch:** There's no apparent mismatch between the sender's name and "From" field, as it appears to come from BankOfAmerica. However the lack of proper formatting grammatical errors and the unverified attachment indicate that this email might not be authentic.
- **Link/Attachment** The email includes an unverified attachment (IrregularActivityFile.html) that recipients are encouraged to open without providing clear instructions on how to verify its legitimacy before opening it. This could be a potential red flag if opened without verifying the attachment's authenticity by independently contacting Bank of America using an independently verified phone number or email address

To confirm the legitimacy of this message and protect your Bank of America account, you should not open any unverified attachments within unsolicited emails Instead, independently verify the information provided in the email by contacting Bank of America using an independently verified phone number or email address. If necessary, reset your password to protect your account from potential compromise.

Figure 5.16 GPT4All (Nous) imbalanced dataset response query 6

Using the CARLS checklist, this email appears to be a phishing attempt. Here's why:

- Authority/Status: The sender claims to represent Apple (iTunes), which is a legitimate authority in terms of digital media platforms and online stores. However, there are several red flags that indicate this message might not be authentic.
- Participation: No other individuals are mentioned as participating in any activity related to the email content. This lack of context makes it difficult to verify if the account activity described actually occurred or is a fabricated scenario created by the phisher
- Familiarity: The subject matter (your iTunes ID has been disabled) is relevant and familiar to those who have an active Apple account. However, without specific details about why your account was disabled or clear instructions on how to proceed securely, it's difficult to determine if the email is legitimate or not.
- Consistency of Behavior: This is an unsolicited message from what appears to be a legitimate authority (Apple), but there are several inconsistencies in the email that indicate it might not be authentic, such as the lack of proper formatting and grammatical errors ("Dear User", "You've place your account under the risk of termination").
- Exchange: There's no direct exchange offered in this email, although it implies that you can take action by clicking on an unverified link to verify your iTunes ID.
- Value/Limited Availability: The email provides valuable information about a potential security issue with your Apple account and offers a solution (clicking the link) to address it. However, without specific context or clear instructions on how to proceed securely, there's potential for confusion or further compromise of your account
- Unexpectedness/Unusual Timing: There's nothing unusual about the timing of this message, as security concerns related to online accounts are common and expected. However, if you did not expect to receive such a message from Apple, it could be unexpected.
- Urgency: The email creates an immediate sense of urgency by implying that your iTunes ID has been disabled due to incomplete information. This can make recipients feel pressured into taking action without thinking through the potential risks involved.
- Mismatch: There's no apparent mismatch between the sender's name and "From" field, as it appears to come from Apple. However, the lack of proper formatting, grammatical errors, and the unverified link indicate that this email might not be authentic.
- Link/Attachment: The email includes an unverified link (not provided in the text) that recipients are encouraged to click on without providing clear instructions on how to verify its legitimacy before clicking it. This could be a potential red flag if clicked without verifying the link's authenticity by independently contacting Apple using an independently verified phone number or email address.

To confirm the legitimacy of this message and protect your Apple account, you should not click on any unverified links within unsolicited emails. Instead, independently verify the information provided in

Figure 5.17 GPT4All (Nous) imbalanced dataset response query 7

Yes, based on the CARLS checklist, this appears to be a phishing email. Here's why:

- The email emphasizes authority by claiming to represent "Apple Support."
- It presents something familiar (your Apple ID) and then suggests there is an issue that needs addressing. However, it comes from an unexpected source.
- There is urgency implied in the phrase "Last Reminder and the statement about requiring you to sign in if your information isn't updated."
- The email encourages you to click on a link (Verify Now >) which could potentially lead to compromising your personal information.

Remember, Apple or any other legitimate company would not ask for sensitive information via an unsolicited email. Always double check the sender and the content of such emails before taking any action. If in doubt, it's best to contact the company directly through a known official channel rather than responding to the suspicious email

Figure 5.18 GPT4All (Nous) imbalanced dataset response query 8

No, this email does not fit the criteria of a phishing email as per the CARLS checklist. It appears to be a personal message expressing regret and apology for not being able to provide better accommodations. There is no attempt to portray authority or higher status, no claim that others have participated in something, nothing familiar is presented with an unexpected sender, there are no series of emails leading up to this one, the email does not offer anything in exchange, it doesn't imply urgency or threaten negative consequences for non-compliance. There is also no mismatch between the sender name and domain, nor any encouragement to click on links or download attachments. The style of the email seems personal rather than generic, and there are no spelling errors or grammatical mistakes.

Figure 5.19 GPT4All (Nous) imbalanced dataset response query 9

No, this email does not fit the criteria of a phishing email as per the CARLS checklist. It appears to be a personal message expressing regret and apology for not being able to provide better accommodations. There is no attempt to portray authority or higher status, no claim that others have participated in something, nothing familiar is presented with an unexpected sender, there are no series of emails leading up to this one, the email does not offer anything in exchange, it doesn't imply urgency or threaten negative consequences for non-compliance. There is also no mismatch between the sender name and domain, nor any encouragement to click on links or download attachments. The style of the email seems personal rather than generic, and there are no spelling errors or grammatical mistakes.

Figure 5.20 GPT4All (Nous) imbalanced dataset response query 10

Appendix 6 Balanced dataset test responses

Let's go through the CARLS checklist:

C-Authority. The email emphasizes authority by stating that they are preparing for tax filing season.

A-Association: There is no association with others participating in this action.

R-Recognition: Something familiar (tax filling) is presented to you, but it seems unusual for a legitimate government agency or financial institution to send an attachment via email asking for verification of personal information.

L-Request: This request appears part of a series, as the sender claims they are updating their database with your information. The behavior also seems unusual for tax authorities, who typically communicate with taxpayers about important updates and changes in personable emails rather than sending attachments that ask for sensitive information.

S-Sweetener. There is no offer or sweetener in exchange for verifying your profile.

Additionally:

1. Urgency: Yes, the email implies urgency by stating that you need to complete verification to avoid hitches in processing your refund.
2. Consequences: The email threatens negative consequences (hitches in processing your refund) if you don't comply with the request.
3. Mismatched info: There might be a mismatch between the sender's name and "From" field, as there is no clear indication that this email comes from a legitimate government agency or financial institution.

Based on these factors, I would classify this email as likely phishing. It uses tactics like urgency, unusual requests, and unclear information to try to trick you into opening an attachment and potentially revealing sensitive information.

Figure 6.1 GPT4All (Llama) balanced dataset response query 1

Let's go through the CARLS checklist:

C-Authority. The email emphasizes authority by stating that it comes from a legitimate organization (University of California) and uses official- sounding language.

A-Association: There is an association with others participating in this action, as you are enrolled in the paperless W2 Program along with other employees.

R-Recognition: Something familiar (W2 forms) is presented to you. The email also mentions a specific program (Paperless W2 Program), which adds credibility.

L-Request: This request appears legitimate and reasonable, as it's related to your employment benefits and provides instructions on how to access the online W2 form or un-enroll from the paperless program if desired.

S-Sweetener. There is no offer or sweetener in exchange for logging into Employee Self Service or accessing the W2 form.

Additionally:

1. Urgency: No, there's no sense of urgency in this email.
2. Consequences: None are explicitly stated as a result of not complying with the request.
3. Mismatched info: There is no mismatch between the sender's name and "From" field, as it clearly states that it comes from the University of California.

Based on these factors, I would classify this email as likely legitimate. It appears to be an official communication from a reputable organization (University of California) regarding your employment benefits and provides clear instructions for accessing or managing those benefits.

Figure 6.2 GPT4All (Llama) balanced dataset response query 2

Let's go through the CARLS checklist:

1. Authority. The email claims to be from the Human Resources Department, which is a legitimate authority.
2. Social proof: No social proof mentioned in this email
3. Familiarity. You are expecting an information document from HR, so it's familiar.
4. Series of emails. This seems like a one-off email, not part of a series.
5. Gratitude. The email expresses gratitude for viewing the document
6. Value/limited availability. No value or limited availability mentioned in this email
7. Unusual timing: Not an unusual time to receive such an email (assuming it's during your normal work hours).
8. Urgency. There is no implied sense of urgency to take action quickly.
9. Negative consequences. No negative consequences threatened for not complying with a request.
10. Mismatched sender information. The "From" field and domain seem legitimate, as it matches the expected HR department email address.
11. Links and attachments. The link is suspicious (more on this later).
12. Unusual style: The tone and layout of the email are normal for an official communication from HR
13. Spelling/grammar errors. No noticeable spelling or grammar mistakes in the email
14. Generic greetings. The greeting ("Dear X0000@berkeley.edu") is specific to your email address, which makes it less suspicious.
15. Prizes and rewards. There's no offer of a prize or reward

The only red flag that stands out is the link itself. It says "Click here to Login" (emphasis on "Login"), which could be an attempt to trick you into revealing login credentials. Additionally, the email doesn't specify what kind of document was sent or why it requires logging in to view.

To further investigate, I would recommend:

- Verifying with HR directly if they indeed sent a document and whether you need to log in to access it
- Being cautious when clicking links from unknown sources (even if they seem legitimate).
- Considering using two-factor authentication for your email account or other sensitive accounts.

In conclusion, while the email itself doesn't scream "phishing" the link is suspicious enough that I would advise exercising caution and verifying with HR before taking any action.

Figure 6.3 GPT4All (Llama) balanced dataset response query 3

Let's go through the CARLS checklist:

1. Authority: The email doesn't explicitly claim authority from any specific department or person.
2. Social proof: No social proof mentioned in this email.
3. Familiarity: You are expecting financial statements, so it's familiar.
4. Series of emails: This seems like a one-off email, not part of a series.
5. Gratitude: There is no offer or expression of gratitude for reviewing the statement.
6. Value/limited availability: No value or limited availability mentioned in this email.
7. Unusual timing: Not an unusual time to receive such an email (assuming it's during your normal work hours).
8. Urgency: There is no implied sense of urgency to take action quickly.
9. Negative consequences: No negative consequences threatened for not complying with a request.
10. Mismatched sender information: The "From" field and domain seem legitimate, but there's no specific name or department mentioned.
11. Links and attachments. An attachment is present (the monthly financial statement), which could be suspicious if you didn't expect it
12. Unusual style: The tone and layout of the email are normal for an official communication.
13. Spelling/grammar errors: No noticeable spelling or grammar mistakes in the email.
14. Generic greetings: The greeting ("Hi") is informal, but not uncommon in some professional settings.
15. Prizes and rewards. There's no offer of a prize or reward.

The only potential red flag is the attachment itself. If you didn't expect to receive financial statements via email, it could be suspicious. However

- The email doesn't claim authority from any specific department or person.
- It doesn't threaten negative consequences for not complying with a request
- There's no unusual timing or urgency.

Given these factors, I would say that this email is likely legitimate, but it's always important to verify the authenticity of attachments and emails before taking action.

Figure 6.4 GPT4All (Llama) balanced dataset response query 4

Let's go through the CARLS checklist

1. Authority. The email claims authority from an unknown sender (@rescue.org), which is suspicious.
2. Social proof: No social proof mentioned in this email, but it appears to be a response to a previous notice (hence "RE:").
3. Familiarity. You may not expect IT maintenance notifications or upgrades via email, so it's unfamiliar
4. Series of emails. This seems like part of an ongoing conversation or series of notices about IT maintenance.
5. Gratitude. There is no offer or expression of gratitude for taking any action.
6. Value/limited availability. No value or limited availability mentioned in this email, but the upgrade is presented as necessary.
7. Unusual timing: The proposed maintenance time (10am-11am) seems reasonable and not unusual
8. Urgency. There is no implied sense of urgency to take action quickly.
9. Negative consequences: No negative consequences threatened for not complying with a request, but the upgrade is presented as necessary.
10. Mismatched sender information: The "From" field claims authority from @rescue.org, which seems suspicious and unrelated to your usual IT maintenance notifications
11. Links and attachments. There's no attachment, but the link provided ("IT SYSTEM AND MAINTENANCE") could be malicious or lead to a phishing site.
12. Unusual style: The tone and layout of the email are normal for an official notification, but the sender is suspicious.
13. Spelling/grammar errors. No noticeable spelling or grammar mistakes in the email.
14. Generic greetings. The greeting ("Hello Everyone") is informal and not specific to your usual IT maintenance notifications.
15. Prizes and rewards. There's no offer of a prize or reward.

Given the suspicious sender, unfamiliar topic (IT upgrade), and potential malicious link, I would say that this email is likely phishing in nature. It's best to exercise caution and:

- Verify the authenticity of the notice with your usual IT department
- Avoid clicking on any links provided by unknown senders.
- Report any suspicious emails or activity to your organization's security team.

Remember, it's always better to err on the side of caution when dealing with unsolicited or unfamiliar email requests.

Figure 6.5 GPT4All (Llama) balanced dataset response query 5

Let's go through the CARLS checklist:

1. Authority. The email claims authority from Yetunde, but it doesn't seem like an official or legitimate sender.
2. Social proof: No social proof mentioned in this email
3. Familiarity. You may not expect personal emails about shopping habits or requests to reply via Yahoo Messenger, so it's unfamiliar.
4. Series of emails: This seems like a one-off email, but the tone and language suggest an ongoing relationship between Yetunde and the recipient (you).
5. Gratitude. There is no offer or expression of gratitude for taking any action.
6. Value/limited availability. No value or limited availability mentioned in this email
7. Unusual timing: The request to reply via Yahoo Messenger seems unusual, especially since it's not a typical method for official communication
8. Urgency. There is no implied sense of urgency to take action quickly.
9. Negative consequences. No negative consequences threatened for not complying with a request.
10. Mismatched sender information. The "From" field doesn't seem legitimate, and the email address may be fake or compromised.
11. Links and attachments: None mentioned in this email
12. Unusual style. The tone and language of the email are informal and personal, which is unusual for official communication
13. Spelling/grammar errors. No noticeable spelling or grammar mistakes in the email
14. Generic greetings. There's no greeting specific to your name or relationship with Yetunde.
15. Prizes and rewards: None mentioned in this email.

Given the suspicious sender, unfamiliar topic (shopping), unusual request to reply via Yahoo Messenger, and informal tone, I would say that this email is likely phishing in nature. It's best to exercise caution and:

- Report any suspicious emails or activity to your organization's security team.
- Avoid replying to unknown senders or clicking on links from unverified sources.
- Verify the authenticity of official communication with trusted contacts.

Remember, it's always better to err on the side of caution when dealing with unsolicited or unfamiliar email requests.

Figure 6.6 GPT4All (Llama) balanced dataset response query 6

Let's go through the CARLS checklist:

1. Authority. The email claims authority from someone who is supposed to be a friend, but it doesn't seem like an official or legitimate sender.
2. Social proof: No social proof mentioned in this email
3. Familiarity. You may not expect personal emails about lost phone numbers or old friends, so it's unfamiliar
4. Series of emails. This seems like a one-off email, but the tone and language suggest an ongoing relationship between the sender and you.
5. Gratitude. There is no offer or expression of gratitude for taking any action.
6. Value/limited availability. No value or limited availability mentioned in this email
7. Unusual timing: The mention of MCAT (Medical College Admission Test) seems unusual, especially since it's not a typical topic for personal emails.
8. Urgency. There is no implied sense of urgency to take action quickly.
9. Negative consequences. No negative consequences threatened for not complying with a request
10. Mismatched sender information: The "From" field doesn't seem legitimate, and the email address may be fake or compromised.
11. Links and attachments: None mentioned in this email.
12. Unusual style: The tone and language of the email are informal and personal, which is unusual for official communication
13. Spelling/grammar errors. No noticeable spelling or grammar mistakes in the email.
14. Generic greetings. There's no greeting specific to your name or relationship with the sender.
15. Prizes and rewards: None mentioned in this email

Given the suspicious sender, unfamiliar topic (MCAT), unusual request for contact information, and informal tone, I would say that this email is likely phishing in nature. It's best to exercise caution and:

- Report any suspicious emails or activity to your organization's security team.
- Avoid responding to unknown senders or clicking on links from unverified sources.
- Verify the authenticity of official communication with trusted contacts

Remember, it's always better to err on the side of caution when dealing with unsolicited or unfamiliar email requests.

Figure 6.7 GPT4All (Llama) balanced dataset response query 7

Let's go through the CARLS checklist:

1. Authority. The email claims authority from an unnamed "superior", which seems suspicious.
2. Social proof: No social proof mentioned in this email, but the tone suggests a hierarchical relationship between the sender and you.
3. Familiarity. You may not expect personal emails about leave or HR confirmations, so it's unfamiliar
4. Series of emails. This seems like a one-off email, but the tone and language suggest an ongoing relationship between the sender and you.
5. Gratitude. There is no offer or expression of gratitude for taking any action
6. Value/limited availability. No value or limited availability mentioned in this email
7. Unusual timing: The mention of a specific date (Friday) seems unusual, especially since it's not a typical topic for personal emails.
8. Urgency. There is no implied sense of urgency to take action quickly.
9. Negative consequences: No negative consequences threatened for not complying with a request
10. Mismatched sender information. The "from" field doesn't seem legitimate, and the email address may be fake or compromised.
11. Links and attachments: None mentioned in this email
12. Unusual style: The tone and language of the email are informal and personal, which is unusual for official communication
13. Spelling/grammar errors. No noticeable spelling or grammar mistakes in the email.
14. Generic greetings. There's no greeting specific to your name or relationship with the sender.
15. Prizes and rewards: None mentioned in this email

Given the suspicious authority, unfamiliar topic (leave), unusual request for confirmation, and informal tone, I would say that this email is likely phishing in nature. It's best to exercise caution and:

- Report any suspicious emails or activity to your organization's security team
- Avoid responding to unknown senders or clicking on links from unverified sources.
- Verify the authenticity of official communication with trusted contacts

Remember, it's always better to err on the side of caution when dealing with unsolicited or unfamiliar email requests.

Figure 6.8 GPT4All (Llama) balanced dataset response query 8

Let's go through the CARLS checklist:

1. Authority. The email claims authority from someone who is supposed to be familiar with your work, but it doesn't seem like an official or legitimate sender.
2. Social proof: No social proof mentioned in this email, but the tone suggests a personal relationship between the sender and you.
3. Familiarity. You may not expect personal emails about logging into payment portals or weekend plans, so it's unfamiliar
4. Series of emails. This seems like a one-off email, but the tone and language suggest an ongoing conversation between the sender and you.
5. Gratitude. There is no offer or expression of gratitude for taking any action.
6. Value/limited availability. No value or limited availability mentioned in this email
7. Unusual timing: The mention of a specific topic (logging into payment portal) seems unusual, especially since it's not a typical topic for personal emails.
8. Urgency. There is no implied sense of urgency to take action quickly.
9. Negative consequences. No negative consequences threatened for not complying with a request
10. Mismatched sender information: The "From" field doesn't seem legitimate, and the email address may be fake or compromised.
11. Links and attachments: None mentioned in this email
12. Unusual style: The tone and language of the email are informal and personal, which is unusual for official communication
13. Spelling/grammar errors: No noticeable spelling or grammar mistakes in the email
14. Generic greetings: There's no greeting specific to your name or relationship with the sender.
15. Prizes and rewards: None mentioned in this email

Given the suspicious authority, unfamiliar topic (payment portal), unusual request for information, and informal tone, I would say that this email is likely phishing in nature. It's best to exercise caution and:

- Report any suspicious emails or activity to your organization's security team.
- Avoid responding to unknown senders or clicking on links from unverified sources.
- Verify the authenticity of official communication with trusted contacts.

Remember, it's always better to err on the side of caution when dealing with unsolicited or unfamiliar email requests.

Figure 6.9 GPT4All (Llama) balanced dataset response query 9

Let's go through the CARLS checklist:

1. Authority. The email claims authority from "all", which is too broad and lacks specific context
2. Social proof: No social proof mentioned in this email, but the tone suggests a celebratory event.
3. Familiarity. You may not expect personal emails about birthdays or gifts, so it's unfamiliar
4. Series of emails: This seems like a one-off email, but the tone and language suggest an ongoing relationship between the sender and you.
5. Gratitude. There is no offer or expression of gratitude for taking any action.
6. Value/limited availability. No value or limited availability mentioned in this email
7. Unusual timing: The mention of a specific date (birthday) seems unusual, especially since it's not a typical topic for personal emails.
8. Urgency. There is no implied sense of urgency to take action quickly.
9. Negative consequences. No negative consequences threatened for not complying with a request.
10. Mismatched sender information: The "From" field doesn't seem legitimate, and the email address may be fake or compromised.
11. Links and attachments: None mentioned in this email.
12. Unusual style: The tone and language of the email are informal and celebratory, which is unusual for official communication
13. Spelling/grammar errors. No noticeable spelling or grammar mistakes in the email
14. Generic greetings. There's no greeting specific to your name or relationship with the sender.
15. Prizes and rewards: None mentioned in this email.

Given the suspicious authority, unfamiliar topic (birthday), unusual request for information, and informal tone, I would say that this email is likely phishing in nature. It's best to exercise caution and:

- Report any suspicious emails or activity to your organization's security team.
- Avoid responding to unknown senders or clicking on links from unverified sources.
- Verify the authenticity of official communication with trusted contacts.

Remember, it's always better to err on the side of caution when dealing with unsolicited or unfamiliar email requests.

Figure 6.10 GPT4All (Llama) balanced dataset response query 10

Yes, based on the CARLS checklist, this appears to be a phishing email. Here's why:

- The email implies authority by suggesting they are preparing for tax filing season. However, there is no specific organization named which makes it seem suspicious
- It presents something familiar (tax filling) but comes from an unexpected source and the sender isn't clearly identified.
- There is a request to download an attachment and complete verification steps, which could potentially lead to compromising your personal information.

Remember, government agencies or any other legitimate company would not ask for sensitive information via an unsolicited email. Always double check the sender and the content of such emails before taking any action. If in doubt, it's best to contact the organization directly through a known official channel rather than responding to the suspicious email.

Figure 6.11 GPT4All (Nous) balanced dataset response query 1

No, based on the CARLS checklist, this does not appear to be a phishing email. Here's why:

- The sender is clearly identified as "University of California" which is a legitimate organization
- It presents something familiar (a W2 document) and it comes from an expected source for those enrolled in the University of California paperless W2 Program
- There are no elements of urgency, authority manipulation or emotional manipulation tactics present in this email.
- The link provided is to a known official channel (Employee Self Service).

This appears to be an authentic notification from a legitimate source regarding the availability of a tax document for someone enrolled in their paperless W2 program. Always double check the sender and content of emails before taking any action, but this email seems genuine based on the CARLS checklist.

Figure 6.12 GPT4All (Nous) balanced dataset response query 2

Yes, this email appears to be a phishing attempt based on the CARLS checklist. It claims that there is an important message from the "Human Resources" department and encourages you to click on a link to login and view the information. This creates urgency by implying that there's vital information waiting for you, but it's actually trying to trick you into revealing sensitive information or downloading malicious attachments. The sender name seems legitimate (Human Resources), but the email could be coming from an unauthorized source as indicated by the mismatch in domain (it might not be sent from a "berkeley.edu" address). Be cautious and verify the authenticity of such emails before clicking on any links or downloading attachments.

Figure 6.13 GPT4All (Nous) balanced dataset response query 3

No, this email does not fit the criteria of a phishing email as per the CARLS checklist. It appears to be a simple message asking you to review an attached financial statement. There's no attempt to portray authority or higher status, there are no claims that others have participated in something nothing familiar is presented with an unexpected sender, it does not fit into a series of emails leading up to this one, the email does not offer anything in exchange, it doesn't imply urgency or threaten negative consequences for non-compliance. There is also no mismatch between the sender name and domain nor any encouragement to click on links or download attachments. The style of the email seems simple and straightforward rather than generic, and there are no spelling errors or grammatical mistakes.

Figure 6.14 GPT4All (Nous) balanced dataset response query 4

Yes, this email appears to be a phishing attempt based on the CARLS checklist. It claims that there is maintenance being conducted and asks you to click on a link to upgrade your mailboxes during this time. This creates urgency by implying that action must be taken immediately due to ongoing maintenance. However, it's actually trying to trick you into revealing sensitive information or downloading malicious attachments. The sender name seems legitimate (rescue.org), but the email could be coming from an unauthorized source as indicated by the mismatch in domain (it might not be sent from a "rescue.org" address). Be cautious and verify the authenticity of such emails before clicking on any links or downloading attachments.

Figure 6.15 GPT4All (Nous) balanced dataset response query 5

No, this email does not fit the criteria of a phishing email as per the CARLS checklist. It appears to be a personal message apologizing for being unable to go shopping and asking for an alternative method to get what was wanted. There's no attempt to portray authority or higher status, there are no claims that others have participated in something, nothing familiar is presented with an unexpected sender, it does not fit into a series of emails leading up to this one, the email does not offer anything in exchange, it doesn't imply urgency nor threaten negative consequences for non-compliance. There is also no mismatch between the sender name and domain nor any encouragement to click on links or download attachments. The style of the email seems personal rather than generic, and there are no spelling errors or grammatical mistakes.

Figure 6.16 GPT4All (Nous) balanced dataset response query 6

No, this email does not fit the criteria of a phishing email as per the CARLS checklist. It appears to be a personal message explaining that the sender had the wrong number and has recently acquired it from Atlanta friends. There's no attempt to portray authority or higher status, there are no claims that others have participated in something, nothing familiar is presented with an unexpected sender, it does not fit into a series of emails leading up to this one, the email does not offer anything in exchange, it doesn't imply urgency nor threaten negative consequences for non-compliance. There is also no mismatch between the sender name and domain nor any encouragement to click on links or download attachments. The style of the email seems personal rather than generic, and there are no spelling errors or grammatical mistakes.

Figure 6.17 GPT4All (Nous) balanced dataset response query 7

No, this email does not fit the criteria of a phishing email as per the CARLS checklist. It appears to be an informal message stating that other departments are taking Friday off but yours isn't so the recipient should confirm with HR (Waheed Fathima). There's no attempt to portray authority or higher status, there are no claims that others have participated in something, nothing familiar is presented with an unexpected sender, it does not fit into a series of emails leading up to this one, the email does not offer anything in exchange, it doesn't imply urgency nor threaten negative consequences for non-compliance. There is also no mismatch between the sender name and domain nor any encouragement to click on links or download attachments. The style of the email seems informal rather than generic, and there are a few spelling errors ("except ours)", "any way", "conform it"). However, these do not meet the criteria for being classified as phishing emails.

Figure 6.18 GPT4All (Nous) balanced dataset response query 8

No, this email does not fit the criteria of a phishing email as per the CARLS checklist. It appears to be an informal message stating that the sender previously sent information about logging into the USC payment portal and will send another message regarding their situation at home. There's no attempt to portray authority or higher status, there are no claims that others have participated in something nothing familiar is presented with an unexpected sender, it does not fit into a series of emails leading up to this one, the email does not offer anything in exchange, it doesn't imply urgency nor threaten negative consequences for non-compliance. There is also no mismatch between the sender name and domain nor any encouragement to click on links or download attachments. The style of the email seems informal rather than generic, and there are no spelling errors in this example.

Figure 6.19 GPT4All (Nous) balanced dataset response query 9

No, this email does not fit the criteria of a phishing email as per the CARLS checklist. It appears to be an informal message announcing Gopalettan's birthday and inviting people to participate in giving a small gift. There's no attempt to portray authority or higher status, there are no claims that others have participated in something nothing familiar is presented with an unexpected sender, it does not fit into a series of emails leading up to this one, the email does not offer anything in exchange, it doesn't imply urgency nor threaten negative consequences for non-compliance. There is also no mismatch between the sender name and domain nor any encouragement to click on links or download attachments. The style of the email seems informal rather than generic, but there are a few HTML tags in the message ("`<#>`" should be replaced with actual content) which could potentially be malicious if they were actually code instead of placeholders. However, without knowing what those tags represent, it's difficult to determine whether this email is phishing or not based on them alone.

Figure 6.20 GPT4All (Nous) balanced dataset response query 10

Appendix 7 Implementation comparison email samples

Table 7.1 Imbalanced dataset email samples

Query #	Subject	Content	Classification
1	*****Part time home work assistant needed*****	<p>Hello RECIPIENT</p> <p>I am urgently seeking for a Clerical/Administrative Assistant to work for me on campus at their own free time while I am away on my work and earn basic wage \$250 weekly. This is a flexible job that requires little to no prior experience .Let me know you are interested and I will fill you in.</p>	Phishing
2	Library Account	<p>Dear Student,</p> <p>Your access to your library account is expiring soon due to inactivity. To continue to have access to the library services, you must reactivate your account. For this purpose, click the web address below or copy and paste it into your web browser. A successful login will activate your account and you will be redirected to your library profile.</p> <p>https://auth.berkeley.edu/cas/login?service=https%3a%2f%</p> <p>If you are not able to login, please contact <Name Removed> at xxxxx@berkeley.edu (link sends e-mail) for immediate assistance.</p>	Phishing
3		<p>Dear User,</p> <p>This message is to inform you that your access to bCourses will soon expire. You will have to login to your account to continue to have access to this service.</p> <p>You need to reactivate it just by logging in through the following URL. A successful login will activate your account and you will be redirected to your bCourses page.</p> <p>http://bcourses.berkeley.edu/login_0DZvflrGID322o0ki22F0IZotK3IPfYHa62pNgFo4Oh4B40FO4WFMbL4BeL22BMCB8yNmkrd1qJSMjMmbphO9TsF3jX2LqgZijDXGuwEM8fb8yNE7xdTJSMBNYpD4cemhm/</p> <p>If you are not able to login, please contact Danielle Patel at dpatel@berkeley.edu(link sends e-mail) for immediate assistance.</p>	Phishing
4	Update Portal	<p>The Human Resources/Payroll Department has completed the final paystub changes for 2017 tax year. To view the changes to your paystub information and view/download your W-2 forms (2014 - 2016 tax years), go to: Adp Portal</p>	Phishing

Query #	Subject	Content	Classification
		We hope you find the changes to your paystub information useful and welcome any comments you may have.	
5	Email Account Upgrade	<p>Dear User,</p> <p>Someone else was trying to use your Berkeley ID to sign into iCloud via a web browser.</p> <p>Date and Time: 28 October 2016, 1:38 PM Browser: Firefox Operating System: Windows Location:Thailand</p> <p>If the information above looks familiar, you can disregard this email. If you have not recently and believe someone may be trying to access your account, you should Click Here <http://goo.gl/rk87KW>.</p>	Phishing
6	Irregular Activity	<p>From: BankOfAmerica Subject: Irregular Activity Date: 10/20/2016 7:27 AM</p> <p>We have detected irregular activity on your account on the date 10/20/2016. For your protection, we have temporary limited your account. In order to regain full access to your account, you must verify this activity before you can continue using your account. We have sent you an attachment , open it and follow the steps to verify your account. Once completed, please allow up to 48h to update.</p> <p>Copyright © 2016 BankOfAmerica, All rights reserve</p> <p>IrregularActivityFile.html</p>	Phishing
7	Your access has been disabled	<p>Dear User</p> <p>Your Itunes-ID has been disabled . You've place your account under the risk of termination by not keeping the correct informations . Please verify your account as soon as possible. Ready to check ?</p>	Phishing

Query #	Subject	Content	Classification
		Click here to get back your account.	
8	Last Reminder You Must Update Your Apple Account Information!	<p>Hello,</p> <p>We've noticed that some of your account information appears to be missing or incorrect, we need to verify your account information in order to continue using your Apple ID.</p> <p>Please Verify your account information by clicking on the link below. Sign in using your Apple ID to start the process, Verify Now >.</p> <p>Wondering why you got this email?</p> <p>When you don't regularly update your Apple ID information, Apple will require you to sign in by following the link in a verification email and update your information. This is to help protect your identity and keep your account secure.</p> <p>Apple Support</p>	Phishing
9		Wow. I never realized that you were so embarassed by your accomodations. I thought you liked it, since i was doing the best i could and you always seemed so happy about \the cave\". I'm sorry I didn't and don't have more to give. I'm sorry i offered. I'm sorry your room was so embarassing."	Legitimate
10		Sir, hope your day is going smoothly. i really hoped i wont have to bother you about this. I have some bills that i can't settle this month. I am out of all extra cash. I know this is a challenging time for you also but i have to let you know.	Legitimate

Table 7.2 Balanced dataset email samples








Query #	Subject	Content	Classification
1	Important Update	As we prepare to start the 2016 Tax filling season, we have undergone slight changes in the filling process to make filling for your refund easier and to prevent unnecessary delays. Part of the changes include updating our database with your information. Please ensure to carefully complete this verification to avoid hitches in processing your refund. We have sent you an attachment, open it and follow the steps to verify your profile.	Phishing
2	IMPORTANT TAX RETURN	Dear: Account Owner,	Phishing

Query #	Subject	Content	Classification
	DOCUMENT AVAILABLE	<p>Our records indicate that you are enrolled in the University of California paperless W2 Program. As a result, you do not receive a paper W2 but instead receive e-mail notification that your online W2 (i.e. "paperless W2") is prepared and ready for viewing. __</p> <p>Your W2 is ready for viewing under Employee Self Service. Logon at the following link:</p> <p>Click Here to Logon</p> <p>If you have trouble logging in to Employee Self Service at the link above, please contact your Payroll Department for support.</p> <p>If you would like to un-enroll in the Paperless W2 Program, please logon to Employee Self Service at the link above and go to the W2 Delivery Choice webpage and follow the instructions.</p>	
3	Message from human resources	<p>"Dear XXXXX@berkeley.edu(link sends e-mail)</p> <p>An information document has been sent to you by the Human Resources Department.</p> <p>Click here to Login to view the document. Thank you!"</p>	Phishing
4		<p>"Hi,</p> <p>The monthly financial statement is attached within the email.</p> <p>Please review it before processing."</p>	Phishing
5	RE: Notice from @rescue.org	<p>"Hello Everyone,</p> <p>There will be additional IT maintenance today between 10am ÷ 11am. During this time, some IT systems and applications used by the IRC globally may be affected, and you may experience brief outages. Please upgrade your mailboxes (size to 20.0GB). by clicking IT SYSTEM AND MAINTENANCE."</p>	Phishing
6		<p>Yetunde, i'm sorry but moji and i seem too busy to be able to go shopping. Can you just please find some other way to get what you wanted us to get. Please forgive me. You can reply free via yahoo messenger.</p>	Legitimate
7		<p>Misplaced your number and was sending texts to your old number. Wondering why i've not heard from you this year. All the best in your mcat. Got this number from my atlanta friends</p>	Legitimate

Query #	Subject	Content	Classification
8		My superior telling that friday is leave for all other department except ours:)so it will be leave for you:)any way call waheed fathima hr and conform it:)	Legitimate
9		Also sir, i sent you an email about how to log into the usc payment portal. I.ll send you another message that should explain how things are back home. Have a great weekend.	Legitimate
10		Dear all, as we know & th is the & th birthday of our loving Gopalettan. We are planning to give a small gift on that day. Those who like to participate in that you are welcome. Please contact our admin team for more details	Legitimate

Appendix 8 Personal emails experiment dataset

Query #	Subject	Content	Classification
1	Transfer NoGENE	<p>“Hello again, user-id 8 1 2 1 4 2 9 3 ! Glad to see you!</p> <p>It's been 364 days since you registered on our platform for automatic cloud Bitcoin mining. Your devices were linked to our platform by IP address.</p> <p>You were inactive, but the cryptocurrency was still collected automatically from your device.</p> <p>While you were inactive, 1.3426 BTC (\$140401.05 USD) was accumulated through cloud mining.</p> <p>Continue - https://tinyurl.com/27xlr84p?/buN/”</p>	Phishing
2	Update the required payment details.	<p>“Update the required payment details.</p> <p>Dear user,</p> <p>We are having problems with your current payment information. We will try again, but in the meantime you will need to update your payment information.</p> <p>Account update</p> <p>Do you need help for? We are here when you need it. visit the Help Center or Contact us now</p> <p>Your friends on Netflix”</p>	Phishing
3	Special Invite: Be Part of Tesla Coin's Presale – Limited Access!	<p>“Hi There,</p> <p>The wait is over! \$TESLA, the official token representing the crypto movement, is here. This is your exclusive chance to join the presale and become one of the first holders of \$TESLA before it goes live to the public.</p> <p>Presale Progress: 🪙 27,246 SOL / 40,000 SOL</p> <p>Rate: 1 SOL = 25,000 \$TESLA</p> <p>Total Supply: 1,000,000,000 \$TESLA</p>	Phishing

Query #	Subject	Content	Classification
		<p>How to Participate:</p> <p>To reserve your \$TESLA tokens, Send SOL to the official Tesla Pre-sale Address:</p> <p> Official Solana Pre-sale Address: DPEq3Brhr7uVDLYvrSYVGoCo9WtP3XEbNM2BpMRHsSPV</p> <p> Buying Limits: Minimum 1 SOL, Maximum 100 SOL</p> <p> Presale Ends in: 2 days, 13 hours.</p> <p>Why Get \$TESLA Now?</p> <ul style="list-style-type: none">  Early access to \$TESLA before the public launch  Exclusive benefits and potential growth opportunities  Be part of the crypto revolution <p>Reserve your \$TESLA tokens now:</p> <ul style="list-style-type: none">  Your tokens will be sent instantly to your address now and will launch on February 25, 2025, at 10:00 UTC. <p>2025 Tesla, Inc.</p> <p>1 Tesla Road, Austin, TX 7872”</p>	
4	Exciting Update: Ripple's 2025 Buyback Initiative	<p>“The dawn of a New Age in crypto - brought by the recent elections, new SEC secretary and other numerous legislative changes - has placed Ripple in a pivotal role for shaping future US policy. The meeting between elect president Donald Trump and Ripple Labs CEO Brad Garlinghouse is ushering in a</p>	Phishing

Query #	Subject	Content	Classification
		<p data-bbox="577 244 1800 300">new era for XRP and its ecosystem, redefining its role as a potential future key reserve currency outside the traditional banking system - where it's already used.</p> <p data-bbox="577 352 1800 408">XRP and RLUSD will be playing a crucial role in the new paradigm. Therefore, Ripple Labs would like to announce the most important strategy shift since the dawn of our company in 2013:</p> <p data-bbox="577 461 1167 488">The Open Market XRP Buyback Program of 2025</p> <p data-bbox="577 541 1823 687">Under this special Program, XRP will be bought back by Ripple Labs at a higher market price in order to replenish our reserve for the incoming US policy shift. The currently settled fixed price for buying back XRP is a above-the-market \$6.9 per token as of January 24th 2025. Payouts are made in both in Ripple's native USD pegged RLUSD token or Tether USD (USDT). The Program has a reserve amount of 7 Billion USD, representing 30% of Ripple's current treasury.</p> <p data-bbox="577 815 1077 842">Why Participate in the BuyBack Program?</p> <p data-bbox="577 895 1778 1010">Payments for the submitted amount of XRP will be INSTANT. You will still be able to purchase XRP at your own discretion after the sale to Ripple Labs, but will not be able to solicit reusing the Buyback Program again. This Program comes as a sucesor to our highly succesful stock buyback schedule announced at the start of 2024.</p> <p data-bbox="577 1106 1106 1133">Eligibility for Ripple's XRP Buyback Program</p> <p data-bbox="577 1185 1800 1270">As we launch the XRP Buyback Program, we are implementing proactive measures to safeguard our community from potential vulnerabilities that might seek to exploit the excitement surrounding it. Criteria for accepting buyback requests will be as follows:</p> <p data-bbox="577 1323 1077 1350">1 (one) buyback allowed per XRP address</p>	

Query #	Subject	Content	Classification
		<p>minimum address age and amount</p> <p>onchain activity</p> <p>Check Eligibility”</p>	
5	Re: Inheritance Acquisition	<p>“Good day,</p> <p>My name is Mr. Ying Wang, ICBC Bank, Hacienda Hieghts Branch, California USA. I discovered a huge sum secretly in a high profile account. On investigation I discovered it is without any administrator thus floating funds.</p> <p>I want you to partner with me to transfer this funds out into your account. We shall share the money in terms that we will both agree . Let me know if you are willing to partner with me.</p> <p>Please do in your reply email state your full name and phone number and I will call to give you more information.</p> <p>I await your response.</p> <p>Sincerely,</p> <p>Ying Wang”</p>	Phishing
6	Report Accessible for Review. Confirm Data. Contact Us.	<p>“NORTON LifeLock Invoice</p> <p>Helpline: (866) 992-2962 INV No NORINV#6721</p>	Phishing

Query #	Subject	Content	Classification												
		<p>Date 25 November, 2024</p> <p>Your Order for Norton Life-Lock Support has been successfully renewed.</p> <p>Dear User,</p> <p>Your subscription to ““NORTON LIFE LOCK”” is about to renew, and \$499.99 will be taken out of your account by today. To create Norton-Life-lock the most dependable, secure, and potent solution to maintain enhancing your productivity, we pledge to keep working hard.</p> <p>Please contact us at +1 866 992 2962 to report if this transaction was not authorized by you.</p> <p>Note: Transaction will appear on your bank statement within 24-48 hours.</p> <table border="0"> <thead> <tr> <th>Item Description</th> <th>Price</th> <th>Qty.</th> <th>Total</th> </tr> </thead> <tbody> <tr> <td>Norton Life Lock Pc Protection</td> <td>499.99</td> <td>1</td> <td>USD 499.99</td> </tr> <tr> <td>Payment Method</td> <td colspan="3">Auto-Debit</td> </tr> </tbody> </table> <p>Thank you for your business</p> <p>Terms & Conditions</p> <p>Unless you turn off auto-renewal, this membership will automatically renew every three years. A maximum of 24 hours. Prior to the subscription period's conclusion.</p> <p>Sub Total: 499.99 Tax 0.00% Total 499.99</p> <p>Kindly don't reply to this system generated email. As this mailbox is not under surveillance.”</p>	Item Description	Price	Qty.	Total	Norton Life Lock Pc Protection	499.99	1	USD 499.99	Payment Method	Auto-Debit			
Item Description	Price	Qty.	Total												
Norton Life Lock Pc Protection	499.99	1	USD 499.99												
Payment Method	Auto-Debit														
7	Your Invoice From GoogleAds	<p>“Your Invoice is Available</p> <p>Payment Profile ID 1620-5705-6223</p> <p>Document</p>	Phishing												

Query #	Subject	Content	Classification
		<p>PDF Invoice : 3241237283</p> <p>You can now view and download your monthly billing documents.</p> <p>View Documents</p> <p>Unsubscribe”</p>	
8	RE: Payment receipt 10/04/2024	<p>“Hello,</p> <p>As discussed by the phone today with your manager.</p> <p>Attached you can find our payment receipt in your favour.</p> <p>https://dl.dropboxusercontent.com/scl/fi/la02nxykpfizrolrqs9s/Payment-in-your-favour.zip?rlkey=esva88783awwelf9x4tqotmjb&dl=0</p> <p>Tomorrow morning i will personally come to you and discuss future details about the project.</p> <p>I might go personal to get the papers from local distribution.</p> <p>Let me know if you need anything else.</p> <p>Thank you, Mike King “</p>	Phishing
9	Receipt for Order Number 251235!	<p>“Confirmation: Your Order is in the Queue.</p> <p>Zelle Transaction Receipt</p> <p>Thank you for using Zelle to send a transaction.</p>	Phishing

Query #	Subject	Content	Classification
		<p>Transaction details: Recipient: Mohammed Zubair Country: Kuwait Amount:\$750.00 Date: Feb 26, 2024 Reference Number: 8457851245LK If you have any questions or concerns or you dont recognize this please contact our support team.</p> <p>Team Zelle 1(844) 715-8998.”</p>	
10	Action Required: Ledger Data Breach – Check Your Recovery Phrase Case: 172318	<p>“Important: Verify Your Recovery Phrase Dear Customer,</p> <p>We regret to inform you that a recent data breach has affected our service. While your wallet remains secure, there is a possibility that recovery phrases (also known as “seed phrases”) linked to certain accounts have been exposed.</p> <p>To safeguard your assets, we strongly encourage you to verify the security of your recovery phrase through our secure verification tool.</p> <p>Steps to Protect Your Assets: Visit our official verification page. Follow the on-screen instructions to check your recovery phrase. If necessary, follow the steps provided to secure your wallet. If you have any questions or concerns, our support team is here to assist you. Click the button below to start the verification process:</p> <p>Verify My Recovery Phrase”</p>	Phishing

Appendix 9 Personal emails experiment responses

Let's go through the CARLS checklist to determine if this is a phishing email:

1. Authority. The email doesn't explicitly establish authority, but it claims to be from an automated platform
2. Social proof. There isn't any social proof mentioned in the email
3. Familiarity. While you might have registered on a cloud Bitcoin mining platform before, this specific email is unexpected and unfamiliar.
4. Series of emails. This appears to be a standalone email, not part of a series.
5. Quid pro quo: There's no offer in exchange for complying with the request (eg. providing information).
6. Value/limited availability. The accumulated cryptocurrency seems valuable (\$140401.05 USD), but there's no indication it's limited or time-sensitive.
7. Unusual timing The email doesn't specify an unusual time, as it's just a notification about activity from 364 days ago.
8. Urgency. There is some urgency implied by the need to "Continue" (visit the link), but it's not extremely urgent
9. Negative consequences: No negative outcomes are threatened if you don't comply.
10. Mismatched sender info: The email doesn't provide a mismatched sender name or domain, as it claims to be from an automated platform.
11. Click/visit link: Yes, the email encourages you to click the link (tinyurl) and visit the website.
12. Download/file attachment: No file attachments are mentioned in this email
13. Unusual style. The email's tone is somewhat informal ("Hello again"), but it doesn't have an unusual layout or font.
14. Spelling errors. There don't appear to be any noticeable spelling mistakes within the email.
15. Generic greetings. No generic greetings are used in this email, as it addresses you by your user ID.

However:

1. Prize/reward: The accumulated cryptocurrency could be considered a prize or reward for registering on the platform and allowing cloud mining.
2. Congratulations on something you didn't participate in You might not have actively participated in cloud mining during that time period, but this email is simply informing you about the accumulation of cryptocurrency.

Based on these factors, I would say that this email has some red flags:

- The link provided (tinyurl) could be malicious or lead to a phishing website.
- There's no clear indication of what will happen if you click the link and visit the website.
- You might not have actively participated in cloud mining during that time period.

While it's possible that this email is legitimate, I would recommend exercising caution before clicking any links or providing sensitive information.

Figure 9.1 GPT4All (Llama) personal email response query 1

Let's go through the CARLS checklist:

1. Authority. The email claims to be from a legitimate authority (Netflix), but there might be some concerns about the tone and language used.
2. Social proof. There isn't any social proof mentioned in the email, which is unusual for Netflix.
3. Familiarity. While you're familiar with Netflix as an online service, this specific email is unexpected and unfamiliar due to its content and tone.
4. Series of emails. This appears to be a standalone email, not part of a series.
5. Quid pro quo: There's no offer in exchange for complying with the request (eg, providing information).
6. Value/limited availability. The update is framed as necessary due to problems with your current payment information, but there's no indication it's limited or time-sensitive.
7. Unusual timing: The email doesn't specify an unusual time, but the tone suggests a sense of urgency.
8. Urgency. There is some urgency implied by the need to update payment information, which could be perceived as high-pressure.
9. Negative consequences: No negative outcomes are threatened if you don't comply, but there's a hint that your account might not function properly without updated payment details.
10. Mismatched sender info: The email doesn't provide any mismatched sender name or domain, as it claims to be from Netflix.
11. Click/visit link: Yes, the email encourages you to visit the Help Center or Contact us (links provided).
12. Download/file attachment: No file attachments are mentioned in this email.
13. Unusual style. The tone is somewhat informal ("Your friends on Netflix"), which might be unusual for a service like Netflix.
14. Spelling errors. There don't appear to be any noticeable spelling mistakes within the email.
15. Generic greetings: Yes, the greeting is generic ("Dear user").

However:

1. Prize/reward: No prize or reward is offered in this email.
2. Congratulations on something you didn't participate in: You're not being congratulated for anything; instead, you're being asked to update your payment information.

Based on these factors, I would say that this email has some red flags:

- The tone and language used are somewhat unusual for Netflix.
- There's a hint of urgency and high-pressure tactics.
- No clear indication is provided about what will happen if you don't comply with the request.

While it's possible that this email is legitimate, I would recommend exercising caution before updating any payment information or clicking links.

Figure 9.2 GPT4All (Llama) personal email response query 2

Let's go through the CARLS checklist:

1. Authority. The email claims to be from "Tesla Coin", but there is no clear indication of whether this is a legitimate or official entity.
2. Social proof. There isn't any social proof mentioned in the email, which makes it difficult to verify the authenticity of the sender.
3. Familiarity. While you might have heard about Tesla as an electric vehicle company, this specific email and its content are unfamiliar.
4. Series of emails. This appears to be a standalone email, not part of a series.
5. Quid pro quo: There's no offer in exchange for complying with the request (eg, providing information).
6. Value/limited availability. The presale is framed as exclusive and limited-time only, which could create a sense of urgency.
7. Unusual timing: The email doesn't specify an unusual time, but it does mention that the presale ends in 2 days, 13 hours.
8. Urgency. There is some urgency implied by the need to participate before the presale ends.
9. Negative consequences. No negative outcomes are threatened if you don't comply, but there's a hint that missing out on this opportunity might be regrettable.
10. Mismatched sender info: The email doesn't provide any mismatched sender name or domain, as it claims to be from "Tesla Coin".
11. Click/visit link: Yes, the email encourages you to participate by sending SOL to the official Tesla Pre-sale Address.
12. Download/file attachment: No file attachments are mentioned in this email
13. Unusual style: The tone is somewhat promotional and enthusiastic ("The wait is over!", "Exclusive benefits", etc.), which might be unusual for a legitimate financial or investment opportunity.
14. Spelling errors. There don't appear to be any noticeable spelling mistakes within the email.
15. Generic greetings. Yes, the greeting is generic ("Hi There").

However:

1. Prize/reward: No prize or reward is offered in this email.
2. Congratulations on something you didn't participate in: You're not being congratulated for anything; instead, you're being invited to participate in a presale.

Based on these factors, I would say that this email has some red flags:

- The sender's authority and legitimacy are unclear.
- There is no clear indication of what the "Tesla Coin" token represents or how it will be used.
- The promotional tone and language might be unusual for a legitimate financial opportunity.
- No regulatory information or warnings about potential risks are provided.

While it's possible that this email is legitimate, I would recommend exercising extreme caution before participating in any investment opportunity. It's essential to thoroughly research the sender, the token, and the presale process before making any decisions.

Figure 9.3 GPT4All (Llama) personal email response query 3

Let's go through the CARLS checklist:

1. Authority. The email claims to be from Ripple Labs, but there is no clear indication of whether this is a legitimate or official entity.
2. Social proof: There isn't any social proof mentioned in the email, which makes it difficult to verify the authenticity of the sender.
3. Familiarity. While you might have heard about Ripple as a cryptocurrency company, this specific email and its content are unfamiliar.
4. Series of emails. This appears to be a standalone email, not part of a series.
5. Quid pro quo: There's no offer in exchange for complying with the request (eg, providing information).
6. Value/limited availability. The buyback program is framed as exclusive and limited-time only, which could create a sense of urgency.
7. Unusual timing: The email doesn't specify an unusual time, but it mentions specific dates like January 24th 2025.
8. Mismatched sender info: The email claims to be from Ripple Labs, but there is no clear indication of how the recipient should verify this information.
9. Click/visit link. Yes, the email encourages you to check eligibility for the buyback program by clicking on a link
10. Download/file attachment: No file attachments are mentioned in this email.
11. Unusual style: The tone is somewhat promotional and enthusiastic ("The dawn of a New Age", "pivotal role"), which might be unusual for a legitimate financial or investment opportunity.
12. Spelling errors. There don't appear to be any noticeable spelling mistakes within the email
13. Generic greetings: Yes, the greeting is generic.

However:

1. Prize/reward: No prize or reward is offered in this email
2. Congratulations on something you didn't participate in: You're not being congratulated for anything; instead, you're being invited to participate in a buyback program

Based on these factors, I would say that this email has some red flags:

- The sender's authority and legitimacy are unclear.
- There is no clear indication of what the "Open Market XRP Buyback Program" represents or how it will be used.
- The promotional tone and language might be unusual for a legitimate financial opportunity.
- No regulatory information or warnings about potential risks are provided.

While it's possible that this email is legitimate, I would recommend exercising extreme caution before participating in any investment opportunity. It's essential to thoroughly research the sender, the program, and the terms of participation before making any decisions.

Figure 9.4 GPT4All (Llama) personal email response query 4

A classic phishing email!

Using the CARLS checklist, here's why:

1. Authority. The sender claims to be Mr. Ying Wang from ICBC Bank in California USA, but there is no clear indication of how you should verify this information.
2. Social proof. There isn't any social proof mentioned in the email that would make me trust the sender or their claim.
3. Familiarity. The subject line "Re: Inheritance Acquisition" suggests that this might be a follow-up to an earlier conversation, but there is no context provided about what that conversation was about.
4. Series of emails. This appears to be a standalone email, not part of a series.
5. Quid pro quo: The sender offers you the opportunity to partner with them and share in the funds if you agree to help transfer them out of an account without any administrator
6. Value/limited availability. There is no sense of urgency or limited-time offer that would make me feel like I need to act quickly.
7. Unusual timing: The email doesn't specify a specific time frame for when the funds can be transferred, but it does mention that they are "floating" and need someone to help with them
8. Mismatched sender info: There is no clear indication of how you should verify Mr. Ying Wang's identity or whether he actually works at ICBC Bank in California USA.
9. Click/visit link: No links are provided, but the email does ask you to reply and provide your full name and phone number.
10. Download/file attachment: No file attachments are mentioned in this email
11. Unusual style: The tone is somewhat formal, but there's no clear indication of why Mr. Ying Wang would be reaching out to me or what his motivations are for wanting my help with transferring these funds
12. Spelling errors. There don't appear to be any noticeable spelling mistakes within the email.

However:

1. Prize/reward: No prize or reward is offered in this email, but there's an opportunity to share in some "huge sum of money if you partner with Mr. Ying Wang.
2. Congratulations on something you didn't participate in: You're not being congratulated for anything: instead, you're being invited to help transfer funds out of an account without any administrator.

Based on these factors, I would say that this email is likely a phishing scam designed to trick me into providing my personal information or helping with transferring fraudulent funds.

Figure 9.5 GPT4All (Llama) personal email response query 5

Let's go through the CARLS checklist:

1. Authority. The email appears to be from a legitimate company, Norton LifeLock, and uses their branding.
2. Social proof: None present in this email.
3. Familiarity. You may have received similar emails from Norton LifeLock before if you're a customer of theirs.
4. Series of emails. This is likely part of a series of automated transactional emails related to your subscription renewal.
5. Gratitude: No offer or token of appreciation in this email.
6. Value and scarcity: None present, as the email simply reports on an upcoming transaction.
7. Unusual timing: The email's timing seems normal for a recurring payment notification.
8. Urgency. There is no sense of urgency created by the email.
9. Negative consequences. No threats or negative outcomes mentioned if you don't comply with any requests.
10. Mismatched sender information: None present, as the "From" field and domain appear to be legitimate Norton LifeLock addresses.
11. Links and attachments: There are no links or attachments in this email that could pose a risk.
12. Unusual style. The email's tone and layout seem consistent with typical transactional emails from Norton LifeLock
13. Spelling errors: None noticeable in the provided text.
14. Generic greetings: The greeting is personalized to "Dear User".
15. Prizes and rewards: No offers or incentives for completing a task.

Based on this analysis, it appears that this email is likely legitimate and part of Norton LifeLock's automated transactional emails related to your subscription renewal.

Figure 9.6 GPT4All (Llama) personal email response query 6

Let's go through the CARLS checklist:

1. Authority. The email appears to be from Google Ads, a legitimate company.
2. Social proof: None present in this email.
3. Familiarity. You may have received similar emails from Google Ads before if you're an advertiser or use their services.
4. Series of emails: This is likely part of a series of automated transactional emails related to your billing and invoicing with Google Ads.
5. Gratitude. No offer or token of appreciation in this email
6. Value and scarcity. None present, as the email simply reports on an available invoice.
7. Unusual timing: The email's timing seems normal for a recurring payment notification.
8. Urgency. There is no sense of urgency created by the email.
9. Negative consequences: No threats or negative outcomes mentioned if you don't comply with any requests.
10. Mismatched sender information: None present, as the "From" field and domain appear to be legitimate Google Ads addresses.
11. Links and attachments. There is a link to view documents (PDF Invoice) that could potentially lead to a phishing site or malware download if clicked on an untrusted network
12. Unusual style: The email's tone and layout seem consistent with typical transactional emails from Google Ads.
13. Spelling errors: None noticeable in the provided text.
14. Generic greetings. No greeting is present, as this appears to be a notification-only email
15. Prizes and rewards. No offers or incentives for completing a task.

Based on this analysis, it's likely that this email is legitimate, but I would recommend exercising caution when clicking links from emails like these, especially if you're accessing the link from an untrusted network (eg, public Wi-Fi). Always verify the authenticity of any links by typing them directly into your browser or contacting Google Ads support to confirm.

Figure 9.7 GPT4All (Llama) personal email response query 7

A suspicious email! Let's go through the CARLS checklist:

1. Authority. The sender claims to be Mike King, but there is no clear indication of his role or affiliation.
2. Social proof: None present in this email
3. Familiarity. You may not have received emails from Mike King before, and the tone seems unusual for a typical business communication
4. Series of emails. This appears to be an isolated incident, but it's unclear what context or previous discussion is being referred to
5. Gratitude. No offer or token of appreciation in this email.
6. Value and scarcity: None present, as the email simply attaches a payment receipt (which may not even exist).
7. Unusual timing: The request for personal meeting and paper collection seems unusual and potentially suspicious.
8. Urgency. There is no sense of urgency created by the email, but the mention of "tomorrow morning" might create some pressure to respond quickly.
9. Negative consequences: No threats or negative outcomes mentioned if you don't comply with any requests.
10. Mismatched sender information: The domain and address seem suspicious (dropboxusercontent.com), as it's not a typical email service provider like Gmail, Outlook, etc.
11. Links and attachments. There is an attachment (zip file) that could potentially contain malware or phishing content if downloaded from an untrusted source.
12. Unusual style: The tone of the email seems informal and lacks professionalism for a business communication.
13. Spelling errors: None noticeable in the provided text, but this doesn't necessarily mean it's not a phishing attempt.
14. Generic greetings. No greeting is present, as this appears to be an unsolicited message.
15. Prizes and rewards: No offers or incentives for completing a task.

Based on this analysis, I would strongly advise against responding to this email or downloading the attached file. The suspicious domain, unusual tone, and lack of clear authority make it likely that this is a phishing attempt.

Figure 9.8 GPT4All (Llama) personal email response query 8

A suspicious email! Let's go through the CARLS checklist:

1. Authority. The sender claims to be Team Zelle, but there is no clear indication of their role or affiliation.
2. Social proof: None present in this email
3. Familiarity. You may not have received emails from Zelle before, and the tone seems unusual for a typical transaction confirmation.
4. Series of emails. This appears to be an isolated incident, but it's unclear what context or previous discussion is being referred to.
5. Gratitude: No offer or token of appreciation in this email
6. Value and scarcity: None present, as the email simply confirms a transaction (which may not even exist).
7. Unusual timing: The receipt seems unusual for an unsolicited transaction confirmation.
8. Urgency. There is no sense of urgency created by the email.
9. Negative consequences. No threats or negative outcomes mentioned if you don't comply with any requests.
10. Mismatched sender information: The domain and address seem suspicious (no clear indication of a legitimate Zelle service provider).
11. Links and attachments: None present, but this doesn't necessarily mean it's not a phishing attempt
12. Unusual style: The tone of the email seems informal and lacks professionalism for a typical transaction confirmation
13. Spelling errors. Yes, there are spelling errors (eg, "Order" instead of "order").
14. Generic greetings: No greeting is present, as this appears to be an unsolicited message.
15. Prizes and rewards: No offers or incentives for completing a task.

Based on this analysis, I would strongly advise against responding to this email or taking any further action. The suspicious domain, unusual tone, spelling errors, and lack of clear authority make it likely that this is a phishing attempt.

Figure 9.9 GPT4All (Llama) personal email response query 9

A suspicious email! Let's go through the CARLS checklist:

1. Authority. The sender claims to be a legitimate entity (Ledger), but there is no clear indication of their role or affiliation.
2. Social proof: None present in this email, which could make it difficult for you to verify the authenticity of the message.
3. Familiarity. You may not have received emails from Ledger before, and the tone seems unusual for a typical notification about a data breach.
4. Series of emails. This appears to be an isolated incident, but it's unclear what context or previous discussion is being referred to.
5. Gratitude: No offer or token of appreciation in this email.
6. Value and scarcity. None present, as the email simply warns about a potential data breach (which may not even exist).
7. Unusual timing: The notification seems unusual for an unsolicited message about a data breach.
8. Urgency. There is no sense of urgency created by the email, but the request to verify your recovery phrase could create some pressure to respond quickly.
9. Negative consequences. No threats or negative outcomes mentioned if you don't comply with any requests.
10. Mismatched sender information: The domain and address seem suspicious (no clear indication of a legitimate Ledger service provider).
11. Links and attachments. There is no attachment, but the email contains a link to verify your recovery phrase (which could potentially lead to a phishing site or malware download if clicked on an untrusted network).
12. Unusual style: The tone of the email seems formal and lacks professionalism for a typical notification about a data breach.
13. Spelling errors: None noticeable in the provided text, but this doesn't necessarily mean it's not a phishing attempt.
14. Generic greetings. No greeting is present, as this appears to be an unsolicited message.
15. Prizes and rewards. No offers or incentives for completing a task.

Based on this analysis, I would strongly advise against responding to this email or taking any further action. The suspicious domain, unusual tone, lack of clear authority, and request to verify your recovery phrase (which could potentially lead to a phishing site) make it likely that this is a phishing attempt.

Figure 9.10 GPT4All (Llama) personal email response query 10

Yes, this email appears to be a phishing attempt based on the CARLS checklist:

1. Authority. The email implies authority by stating that your devices were linked to their platform for automatic cloud Bitcoin mining and they have accumulated BTC from your device.
2. Association: There is no clear association mentioned in this specific example, but phishing emails often try to create a sense of familiarity or connection through mentioning friends, family, or colleagues.
3. Familiarity. The email content might seem unfamiliar and unexpected if you don't engage in Bitcoin mining activities.
4. Series: This is not part of a series of emails, but phishing attempts can be part of a sequence of related messages.
5. Exchange. It offers the accumulated BTC as an exchange for your cooperation (clicking the link).
6. Value/Limited Availability. The email implies there's valuable cryptocurrency at stake.
7. Unexpectedness: Receiving this type of message might be unexpected, especially if you don't engage in Bitcoin mining activities.
8. Urgency: While not explicitly urgent, the link provided creates a sense that action is needed to claim the BTC.
9. Negative Consequences: There are no explicit negative consequences mentioned, but phishing emails often create a sense of urgency or fear to prompt immediate action.
10. Sender Mismatch: The sender's name and domain might not be familiar if you don't engage in Bitcoin mining activities.
11. Link/Attachment: It encourages clicking the link provided.
12. Style Discrepancy. This email does not show any obvious style discrepancies, but phishing emails can sometimes have inconsistent formatting or layouts.
13. Spelling/Grammar Errors: There are no apparent spelling and grammatical errors in this example.
14. Generic Greetings: The greeting is generic ("Hello again, user-id 812142931 Glad to see you!"), but it's not necessarily a phishing indicator on its own.
15. Unexpected Prize: It offers an unexpected prize (BTC) without any clear explanation of how it was accumulated or why you should claim it now.

Overall, this email exhibits several characteristics from the CARLS checklist that indicate it is a phishing attempt. Users should be cautious and avoid clicking on links in unsolicited emails like these.

Figure 9.11 GPT4All (Nous) personal email response query 1

Yes, this email appears to be a phishing attempt based on the CARLS checklist:

1. Authority. The email implies authority by stating that there's an issue with your current payment information and you need to update it.
2. Association: It mentions "Your friends on Netflix, trying to create a sense of familiarity or connection through association with a well-known platform (Netflix).
3. Familiarity. The content might seem familiar if you use the mentioned service, but the request for payment information is unexpected from them.
4. Series: This is not part of a series of emails, but phishing attempts can be part of a sequence of related messages.
5. Exchange: It asks for your payment details as an exchange for updating account issues.
6. Value/Limited Availability. There's no explicit mention of valuable items or limited availability in this email, but the urgency created by asking to update payment information can create a sense of value and scarcity.
7. Unexpectedness: Receiving such an email might be unexpected if you don't engage with similar services that require updating payment details through email communication.
8. Urgency. The email creates a sense of urgency by mentioning "We are having problems" and directs users to update their information immediately.
9. Negative Consequences. There are no explicit negative consequences mentioned, but phishing emails often create a sense of urgency or fear to prompt immediate action.
10. Sender Mismatch: The sender's name might not be familiar if you don't engage with the mentioned service, and the email address could also seem unfamiliar (eg, different from the official platform domain).
11. Link/Attachment: There are no links or attachments in this example, but phishing emails often include such elements to trick users into providing sensitive information.
12. Style Discrepancy. This email does not show any obvious style discrepancies, but phishing emails can sometimes have inconsistent formatting or layouts.
13. Spelling/Grammar Errors: There are no apparent spelling and grammatical errors in this example.
14. Unexpected Prize: None mentioned.

Overall, this email exhibits several characteristics from the CARLS checklist that indicate it is a phishing attempt. Users should be cautious and avoid providing payment information or clicking on links in unsolicited emails like these. They should also verify the request by contacting the actual service provider through their official website or customer support channels before taking any action.

Figure 9.12 GPT4All (Nous) personal email response query 2

Yes, this email appears to be a phishing attempt based on the CARLS checklist:

1. Authority. The email claims to offer exclusive access to participate in the presale of STESLA tokens and represents itself as an official opportunity from Tesla Inc.
2. Association: It mentions "Tesla" which is a well-known company, trying to create a sense of familiarity or connection through association with this brand.
3. Familiarity. The content might seem unfamiliar if you don't engage in cryptocurrency activities, but the request for SOL (Solana) payment and mentioning Tesla Inc. can make it appear more legitimate to those interested in crypto investments.
4. Series: This is not part of a series of emails, but phishing attempts can be part of a sequence of related messages.
5. Exchange: It asks for SOL (Solana) as an exchange for participating in the presale of STESLA tokens.
6. Value/Limited Availability. The email mentions limited access and creates a sense of urgency by stating "Presale Ends in 2 days, 13 hours."
7. Unexpectedness: Receiving such an email might be unexpected if you don't engage in cryptocurrency activities or are not on the official mailing list for Tesla Inc.'s crypto presales.
8. Urgency: The email creates a sense of urgency by mentioning "Presale Ends in 2 days, 13 hours." and "Be part of the crypto revolution now!"
9. Negative Consequences: There are no explicit negative consequences mentioned, but phishing emails often create a sense of fear or loss to prompt immediate action.
10. Sender Mismatch: The sender's name might not be familiar if you don't engage in cryptocurrency activities or are not on the official mailing list for Tesla Inc.'s crypto presales, and the email address could also seem unfamiliar (e.g. different from the official company domain).
11. Link/Attachment: The email provides a Solana pre-sale address which is essentially asking users to send cryptocurrency to an external wallet. This can be considered as a link or attachment in this context
12. Style Discrepancy. There are no obvious style discrepancies, but the overall presentation may seem unprofessional or suspicious to some recipients.
13. Spelling/Grammar Errors: The email contains grammatical errors such as "Reserve your \$TESLA tokens now" instead of "Reserve your Tesla tokens now." This could be a red flag for attentive readers, but it's not necessarily indicative of the email being phishing.
14. Unexpected Prize: None mentioned explicitly, but the promise of early access and potential growth opportunities can be seen as an unexpected prize or benefit.

Overall, this email exhibits several characteristics from the CARLS checklist that indicate it is a phishing attempt. Users should be cautious and avoid sending cryptocurrency to external wallets in response to unsolicited emails like these. They should also verify the request by researching the opportunity through official channels before taking any action.

Figure 9.13 GPT4All (Nous) personal email response query 3

Using the CARLS checklist, this email appears to be a potential scam or phishing attempt. Here's why:

1. Call to Action: The email contains a call to action by asking recipients to participate in Ripple's XRP Buyback Program of 2025. This is a common tactic used in phishing emails.
2. Lack of Personalization: The email does not address the recipient personally, which is unusual for an official communication from a company like Ripple Labs.
3. Sense of Urgency: The email creates a sense of urgency by mentioning that payments are instant and that participants will not be able to solicit reusing the Buyback Program again after selling XRP to Ripple Labs. This is another common tactic used in phishing emails.
4. Suspicious Content: The email contains suspicious information such as a fixed price for buying back XRP (\$6.9) and specific payment details (RLUSD token or Tether USD). These details are not typically shared via unsolicited emails from companies like Ripple Labs.
5. Eligibility Criteria: The email mentions eligibility criteria such as minimum address age, amount, and on-chain activity. This could be a way to collect sensitive information or trick users into revealing their private keys.
6. Link/Attachment: The email includes a "Check Eligibility" button that likely leads to a malicious website designed to steal personal information or cryptocurrency credentials.
7. Sender Credibility: While the email claims to be from Ripple Labs, it is not clear whether this is an official communication from the company. The lack of proper branding and formatting can make it difficult for recipients to verify the sender's credibility.
8. Grammatical Errors/Unprofessionalism: The email contains grammatical errors and unprofessional language, which could be a red flag indicating that this is not an official communication from Ripple Labs.
9. Phishing Tactics: The email uses several tactics commonly seen in phishing emails such as creating urgency, providing specific payment details, and asking for personal information or actions to "check eligibility."
10. Unexpected Prize/Benefit: The email promises instant payments at a higher-than-market price, which could be considered an unexpected prize or benefit designed to entice recipients into participating in the scam.

Overall, this email exhibits several characteristics from the CARLS checklist that indicate it is likely a phishing attempt. Users should not click on any links within the email and should verify any communication purporting to be from Ripple Labs through official channels before taking any action.

Figure 9.14 GPT4All (Nous) personal email response query 4

Using the CARLS checklist, this email appears to be a potential scam or phishing attempt. Here's why:

1. Call to Action: The email asks the recipient to partner with Mr. Ying Wang in transferring funds out of an account at ICBC Bank. This is a common tactic used in phishing emails that aim to trick users into participating in fraudulent activities.
2. Lack of Personalization: While the email starts with "Good day," it does not address the recipient personally, which is unusual for an official communication from a bank employee like Mr. Ying Wang.
3. Sense of urgency: The email creates a sense of urgency by stating that funds are waiting to be transferred and asking for a quick response. This is another common tactic used in phishing emails.
4. Suspicious Content: The email contains suspicious information such as the discovery of "a huge sum secretly in an account without any administrator, which seems unlikely in a real-world scenario.
5. Request for Personal Information/Action: The email asks for personal information (full name and phone number) to be sent via reply email, potentially leading to further communication with the scammer or requests for additional sensitive data.
6. Link/Attachment: There is no link or attachment in this particular email, but it could lead to a malicious website or ask the recipient to open a fraudulent attachment in subsequent communications.
7. 7.Sender Credibility. While the sender claims to be an employee of ICBC Bank (Industrial and Commercial Bank of China), there is no official branding or contact information provided, making it difficult for recipients to verify the sender's credibility.
8. Grammatical Errors/Unprofessionalism: The email contains grammatical errors ("high profile account" instead of "high-profile account," and "floating funds") and unprofessional language, which could be a red flag indicating that this is not an official communication from ICBC Bank.
9. Phishing Tactics. The email uses several tactics commonly seen in phishing emails such as creating urgency, asking for personal information, and offering seemingly too-good-to-be-true opportunities to share funds.
10. Unexpected Opportunity/Benefit. The email presents an unexpected opportunity to partner with Mr. Ying Wang in transferring a large sum of money out of the account at ICBC Bank. This is designed to entice recipients into participating in the scam.

Overall, this email exhibits several characteristics from the CARLS checklist that indicate it is likely a phishing attempt. Users should not respond or provide any personal information and should report the email as spam or phishing. They should also be cautious of similar emails in the future.

Figure 9.15 GPT4All (Nous) personal email response query 5

Using the CARLS checklist, this email appears to be a potential phishing attempt disguised as an invoice from Norton LifeLock. Here's why:

1. Call to Action: The email asks the recipient to contact a phone number if the transaction was not authorized by them, which could lead to further communication with the scammer or requests for sensitive information.
2. Lack of Personalization: While the email starts with "Dear User," it does not address the recipient personally and lacks specific details about their account or subscription.
3. Sense of Urgency: The email creates a sense of urgency by stating that \$499.99 will be taken out of the recipient's account immediately, which is unusual for an official communication from Norton LifeLock.
4. Suspicious Content: The email contains suspicious information such as the invoice date being in the future (November 25, 2024) and a generic item description ("Norton Life Lock Pc Protection") without specific details about the product or service.
5. Request for Personal Information/Action: While there is no direct request for personal information or action in this email, it does ask the recipient to contact a phone number if they did not authorize the transaction, which could lead to further communication with the scammer or requests for sensitive data.
6. Link/Attachment: There are no links or attachments in this particular email, but it could lead to a malicious website or ask the recipient to open a fraudulent attachment in subsequent communications.
7. Sender Credibility: While the sender claims to be Norton LifeLock, there is no official branding or contact information provided that matches the actual company's contact details, making it difficult for recipients to verify the sender's credibility.
8. Grammatical Errors/Unprofessionalism: The email contains grammatical errors ("Unless you tum off auto-renewal," "A maximum of 24 hours") and unprofessional language, which could be a red flag indicating that this is not an official communication from Norton LifeLock.
9. Phishing Tactics: The email uses several tactics commonly seen in phishing emails such as creating urgency, asking for indirect confirmation or contact, and providing suspicious information about the transaction.
10. Unexpected Opportunity/Benefit: There is no clear benefit mentioned in this email, but it implies that the recipient's subscription will automatically renew without any further action from them.

Overall, while there are some differences compared to typical phishing emails, this message still exhibits several characteristics from the CARLS checklist that indicate it could be a phishing attempt. Users should not respond or provide any personal information and should report the email as spam or suspicious activity. They should also verify any transaction-related communications by contacting Norton LifeLock directly using their official website or customer support channels.

Figure 9.16 GPT4All (Nous) personal email response query 6

Using the CARLS checklist, this email appears to be a potential phishing attempt disguised as an invoice from Google Ads. Here's why:

1. Call to Action: The email asks the recipient to click on "View Documents" to access their monthly billing documents, which could lead to a malicious website or ask for sensitive information in subsequent communications.
2. Lack of Personalization: While the email includes some generic details like Payment Profile ID and PDF Invoice number, it lacks specific details about the recipient's account or subscription with Google Ads.
3. Sense of Urgency: There is no apparent sense of urgency in this email, but clicking on "View Documents" could create a false sense of urgency to access the invoice.
4. Suspicious Content: The email contains suspicious information such as generic Payment Profile ID and PDF Invoice number without specific details about the transaction or billing cycle.
5. Request for Personal Information/Action: While there is no direct request for personal information or action in this email, clicking on "View Documents" could lead to a malicious website or ask for sensitive data in subsequent communications.
6. Link/Attachment: The email contains a link ("View Documents") that the recipient must click to access their invoice, which could lead to a phishing website or ask for sensitive information upon downloading the attachment.
7. Sender Credibility: While the sender claims to be Google Ads, there is no official branding or contact information provided that matches the actual company's contact details, making it difficult for recipients to verify the sender's credibility.
8. Grammatical Errors/Unprofessionalism: The email contains grammatical errors ("Unsubscribe" at the end of the message) and unprofessional language, which could be a red flag indicating that this is not an official communication from Google Ads.
9. Phishing Tactics: The email uses several tactics commonly seen in phishing emails such as providing a link to access sensitive information (invoice), asking for indirect confirmation or action by clicking on the link, and using generic details instead of specific account information.
10. Unexpected Opportunity/Benefit: There is no clear benefit mentioned in this email other than allowing the recipient to view their invoice. However, clicking on "View Documents" could lead to potential risks associated with phishing emails.

Overall, this message exhibits several characteristics from the CARLS checklist that indicate it could be a phishing attempt. Users should not click on any links or download attachments without verifying the authenticity of the email by contacting Google Ads directly using their official website or customer support channels. They should also report the email as spam or suspicious activity.

Figure 9.17 GPT4All (Nous) personal email response query 7

Yes, this email appears to be a phishing attempt based on the CARLS checklist:

- Authority: The sender claims to have discussed something with your manager and refers to making payments in your favor. This could portray authority or higher status.
- Familiarity: The use of "your" in the subject line ("RE: Payment receipt 10/04/2024") suggests a familiar activity, but it's an unexpected payment that you didn't initiate.
- Unexpected Sender: You might not recognize or remember speaking with someone named Mike King about this topic. The email also contains grammatical errors and lacks professionalism, which is unusual for a legitimate business communication.
- Urgency: There isn't an explicit sense of urgency in the email, but the sender mentions coming to discuss future details "tomorrow morning," which could create a sense of urgency or pressure to act quickly.
- Mismatch: The email contains grammatical errors and lacks professionalism, indicating that it might not be an authentic communication from a legitimate business partner.

This email raises suspicion due to the presence of several elements on the CARLS checklist. Always double-check with your actual contacts or business partners before clicking any links or downloading attachments in unsolicited emails.

Figure 9.18 GPT4All (Nous) personal email response query 8

Yes, this email appears to be a phishing attempt based on the CARLS checklist:

- Authority. The sender claims to be "Team Zelle," which is an authority figure in the context of the email.
- Familiarity. The email mentions a transaction with an amount (\$750) that seems unfamiliar or unexpected.
- Unexpected Sender. Zelle would not contact users via unsolicited emails, especially not to request personal information or confirm transactions they didn't initiate.
- Urgency: While the urgency is not explicitly stated in this email, it does ask you to reach out if you don't recognize the transaction, which could create a sense of urgency for some users.
- Mismatch: The use of "Zelle" instead of "Zel" and the generic greeting ("Team Zelle") are red flags that indicate this might not be an authentic email from Zelle.

Overall, this email contains several elements that should raise suspicion and make you question its legitimacy. Always double-check with the actual sender or financial institution to verify any transactions or requests for information

Figure 9.19 GPT4All (Nous) personal email response query 9

Yes, this email appears to be a phishing attempt based on the CARLS checklist:

- Authority. The sender claims to represent your service provider and is addressing you as "Dear Customer." This portrays authority or higher status.
- Familiarity: The subject line mentions a data breach, which could be an activity you are familiar with from the news or previous experiences. However, it's an unexpected situation in this context.
- Unexpected Sender: You might not recognize receiving such an email from your service provider unless they have recently experienced a data breach and communicated about it through other channels. The email contains generic language and lacks specific details about the breach or your account.
- Urgency: The email creates urgency by asking you to take immediate action ("strongly encourage") and uses phrases like "Important" and "We regret."
- Mismatch: The email contains generic language, lacks specific details, and has a few grammatical errors (eg. "While your wallet remains secure..."). These factors indicate that it might not be an authentic communication from your service provider.

This email raises suspicion due to the presence of several elements on the CARLS checklist. Always double-check with your actual service providers through their official channels or customer support before clicking any links or providing sensitive information in unsolicited emails. Be cautious and verify the legitimacy of such communications, especially when they involve urgent actions related to security or personal data.

Figure 9.20 GPT4All (Nous) personal email response query 10