Highlights

Securing the Weakest Link: Exploring Affective States Exploited in Phishing Emails with Large Language Models

Faithful Chiagoziem Onwuegbuche,Rajesh Titung,Esa Rantanen,Anca Delia Jurcut,Cecilia O. Alm,Liliana Pasquale

- Explored emotional responses targeted by attackers using Large Language Models
- Identified comprehensive phishing-specific affective states
- Used zero-shot learning to detect the exploitation of these states in over 5,000 phishing emails
- Results suggest that phishing mostly exploits curiosity, urgency in email subject lines; fear, urgency, trust in email bodies
- Findings imply that considering affective states may improve phishing email detectors

Securing the Weakest Link: Exploring Affective States Exploited in Phishing Emails with Large Language Models

Faithful Chiagoziem Onwuegbuche[a,b,c,*,1], Rajesh Titung[d,1], Esa Rantanen[d], Anca Delia Jurcut[a], Cecilia O. Alm[d] and Liliana Pasquale[a,c]

[a]School of Computer Science, University College Dublin, Dublin, Ireland
[b]SFI Center for Research Training in Machine Learning, Dublin, Ireland
[c]SFI Research Centre for Software, Lero, Ireland
[d]Rochester Institute of Technology, Rochester, New York, USA

ARTICLE INFO

Keywords:
Phishing Susceptibility
Affective States
Large Language Models (LLMs)
Cyberattacks
Human Factors in Cybersecurity

ABSTRACT

Cyberattacks typically begin with a phishing email aimed at compromising recipients' security by exploiting inherent human traits that are often the weakest link in cybersecurity defences. While prior research has identified human susceptibility factors to scams and fraud, it remains unclear how these factors are exploited in phishing emails to induce emotional responses and how effective Large Language Models (LLMs) are in detecting these factors automatically. This study identified comprehensive psychological phishing-specific factors of human susceptibility inspired by interdisciplinary fields such as security, human-computer interaction (HCI), and psychology to explain the emotional responses targeted by attackers in phishing emails. We detected the exploitation of these affective responses in over 5,000 phishing emails targeting six universities with three state-of-the-art LLMs - GPT-4, Llama 2, and Gemini Pro. Our analysis compared LLMs' performance with human annotators using reliability statistics, mapped exploited affect states to the valence-arousal cartesian space, examined LLMs' performance in detecting phishing emails with special characters, and analyzed LLM hallucinations. The study findings indicate that attackers exploit curiosity and urgency in phishing email subjects, while fear, urgency, and trust emerge as key affective states exploited in the body of phishing emails. Affective exploits in email influence arousal levels, with fear and urgency having the highest arousal levels in both subjects and bodies, while the valence of the text tends to be neutral. In terms of paired affect exploitation in phishing emails, we observed that the following combinations: fear-urgency, urgency-trust, and trust-curiosity are among the top four pairings in both phishing email subjects and bodies. Comparatively, GPT-4 slightly performs slightly better than Gemini Pro and Llama 2. Additionally, there is a fair agreement between the LLMs and human annotators. These findings have important implications for further research, suggesting that considering affective states can enhance the development of phishing email detectors.

1. Introduction

Humans are considered the weakest link in securing systems as the human element introduces unpredictability, susceptibility to manipulation, and the potential for errors (Mitnick and Simon, 2003; Mansfield-Devine, 2017; Danet, 2021; Schneier, 2000). Evidence shows that 82% of all cyberattacks involve a human element (Davies, 2023). Similarly, 91% of all cyber attacks begin with a phishing email to an unexpected victim (Deloitte, 2020).

Thus, phishing is still the most common form of cybercrime. Valimail (2019) estimates that 3.4 billion phishing emails are sent daily and that phishing accounts for nearly 36% of all data breaches (Muncaster, 2021). Moreover, the Anti-Phishing Working Group APWG (2024) Phishing Activity Trends Report observed that there were almost five million phishing attacks in 2023, making it the worst year for phishing on record. Several reports and surveys such as the one conducted by Proofpoint (2021) revealed that 83% of companies indicated that their organization had experienced a successful email-based phishing attack.

Furthermore, the number of phishing attacks is expected to rise as advancements in artificial intelligence, particularly Large Language Models (LLMs), have significantly facilitated sophisticated phishing campaigns, enabling cybercriminals to launch a higher volume of attacks more quickly and reducing the barrier to entry (Gupta, Akiri, Aryal, Parker and Praharaj, 2023; Egress, 2023; Okey, Udo, Rosa, Rodríguez and Kleinschmidt, 2023). In the past, phishing emails were often recognizable due to poor writing practices, such as spelling mistakes. However, malicious LLMs such as WormGPT and PoisonGPT can generate more convincing phishing emails that are harder to detect and exploit affective states (Picard, 2000) in individuals (Packetlabs, 2023). Also, most of the existing phishing email detectors focus on building systems that detect phishing emails based on senders' domain names, email addresses, and URLs (Sharma and Bashir, 2020; Atari and Al-Mousa, 2022). Thus, they do not consider the affective states exploited by cybercriminals in their detection system. However, phishing attackers continue to devise more sophisticated and innovative methods to capture users' attention by exploiting human affective states

*Corresponding Author
faithful.chiagoziemonwuegb@ucdconnect.ie (F.C. Onwuegbuche); rt7331@g.rit.edu (R. Titung)
ORCID(s): 0000-0001-9580-4260 (F.C. Onwuegbuche); 0000-0001-5938-7121 (R. Titung); 0000-0001-9666-4458 (E. Rantanen); 0000-0002-2705-1823 (A.D. Jurcut); 0000-0002-8730-0916 (C.O. Alm); 0000-0001-9673-3054 (L. Pasquale)
[1]Made Equal Contributions

and bypassing these detection models (Sharma and Bashir, 2020). Consequently, these attacks' increased sophistication, persuasion, and scale are expected to lead to a heightened susceptibility to phishing email attacks. This means that security researchers must effectively adapt their strategies to combat evolving phishing attacks.

Cybercriminals design phishing emails that exploit phishing susceptibility factors to increase the recipients' susceptibility to deception and trick recipients into revealing sensitive information or compromising their security (Frauenstein and Flowerday, 2020; Ribeiro, Guedes and Cardoso, 2024). These emails often mimic legitimate communications from trusted entities such as banks, social media platforms, or online services (Almomani, Gupta, Atawneh, Meulenberg and Almomani, 2013). The phishing susceptibility factors exploited can influence the victim's reasoning and undermine their ability to assess the content of the email presented to them in a rational manner (Button, Nicholls, Kerr and Owen, 2014).

Various theories have been proposed to explain human susceptibility to influence, scams, and fraud, such as the principles of influence (Cialdini, 1993), psychological triggers (Gragg, 2003), principles of scams (Stajano and Wilson, 2011), principles of persuasion in social engineering (Ferreira, Coventry and Lenzini, 2015), and cognitive hacks (Schneier, 2023). However, it is unclear how attackers exploit these susceptibility factors in phishing emails in terms of the emotional responses they induce in email users, as well as the effectiveness of LLMs in automatically detecting these susceptibility factors. Therefore, there is a need to identify phishing-specific psychological factors that explain the emotional states targeted by cybercriminals, increasing individuals' vulnerability to phishing attacks. Additionally, the identified factors should be unambiguous and classifiable by LLMs without requiring extensive fine-tuning, which can be computationally resource-intensive.

To address these gaps, our paper explores how attackers exploit susceptibility factors in phishing emails and assesses how LLMs can detect these affective states automatically.

This study makes the following contributions:

1. We identify phishing-specific psychological susceptibility factors, which we term *affective states*. They explain the emotional states targeted by cybercriminals in phishing emails. The identified factors inspired by previous research on susceptibility factors to influence, scams, and frauds in security, HCI, and psychology.
2. We conducted the first thorough comparative analysis of the effectiveness of three state-of-the-art LLMs (GPT-4, Gemini Pro, and Llama 2) in automatically classifying phishing emails based on the identified affective states. Additionally, we compared the performance of LLMs against human annotators using reliability analysis with statistics such as Cohen's κ, Fleiss' κ, and Krippendorff α, providing valuable insights into the capabilities of LLMs for this task.
3. We mapped the identified affective states to a valence-arousal cartesian space, based on the valence and arousal ratings provided by majority voting of three LLMs.
4. We investigated how cybercriminals use special characters to bypass phishing filters. We evaluated the ability of LLMs to detect such tactics, addressing the evolving nature of phishing attacks.
5. We explored LLM hallucination in detecting affective states in phishing emails.

Our results indicate that attackers commonly exploit curiosity and urgency in email subjects, while fear, urgency, and trust emerge as central exploited affect states in email bodies. Attackers also often target a single affective state for email subjects but use more strategies in email bodies, targeting multiple affective states. Our analysis further suggests that the exploited affective states in the phishing emails influence arousal levels, with fear and urgency having the highest arousal levels in both email subjects and bodies, whereas valence tends to be neutral. The findings suggest that considering affective states can enhance the development of phishing email detectors and educate targeted user groups, thereby reducing their susceptibility to phishing emails. However, they also suggest that LLMs are not yet robust enough in recognizing them.

The remainder of the paper is structured as follows: Section 2 reviews the relevant literature and emphasises the research gaps filled by this paper. In Section 3, the identified affective states are presented to form the theoretical basis for the work. The methodology employed in this study is explained in Section 4. Section 5 is dedicated to presenting and discussing the results, while Section 6 offers concluding remarks.

2. Related work

We review related work that attempts to characterize phishing susceptibility factors and automate the detection of affective states in phishing emails.

2.1. Phishing susceptibility factors

Phishing susceptibility refers to the likelihood of an individual falling prey to a phishing attack. Fan, Li, Laskey and Chang (2024) and Parrish Jr, Bailey and Courtney (2009) suggested that phishing susceptibility is influenced by various factors, which can be categorized into three main categories: demographic, experiential, and psychosocial factors, as illustrated in Figure 1.

Demographic factors encompass attributes that are either unchangeable or challenging to modify, such as gender, age, and cultural background. Simoiu, Zand, Thomas and Bursztein (2020) showed that phishing targets are sometimes localized based on age, locality, device classes, and even prior security incidents to which users fell victim. There are various and sometimes conflicting findings concerning the effect of demographic factors on phishing susceptibility. Research on age and phishing susceptibility has shown

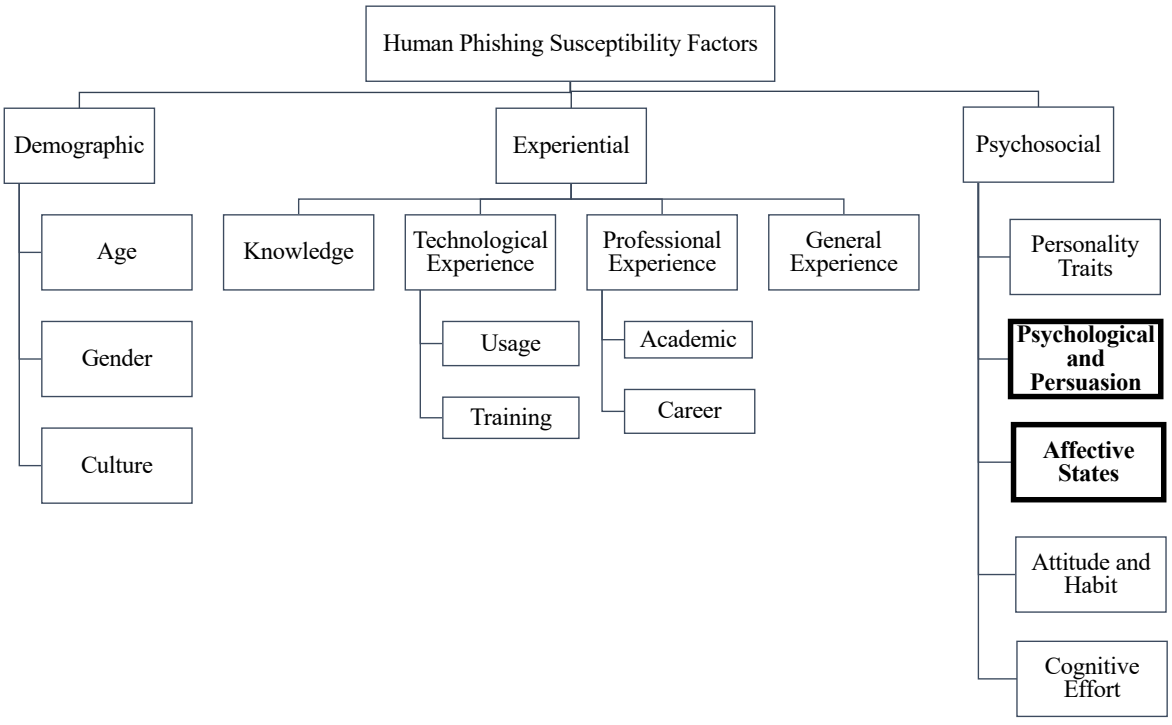


Figure 1: Categorization of human phishing susceptibility factors. This figure illustrates the different research areas on human susceptibility to phishing. Our work falls under psychosocial factors specifically affective states, and psychological and persuasion methods.

varying results, with some studies suggesting that older adults may be more susceptible due to overconfidence and lower awareness (Lin, Capecci, Ellis, Rocha, Dommaraju, Oliveira and Ebner, 2019; Li, Lee, Purl, Greitzer, Yousefi and Laskey, 2020). In contrast, others have found no significant age differences or limited strength of age effects (Ribeiro et al., 2024; Mohebzada, El Zarka, BHojani and Darwish, 2012; Baki and Verma, 2022). Gender differences have also been observed, with some studies indicating that women are more susceptible to phishing than men (Sheng, Holbrook, Kumaraguru, Cranor and Downs, 2010; Li et al., 2020), while others suggest that men may be more susceptible in specific scenarios (Mohebzada et al., 2012) and some studies indicating no significant difference (Ribeiro et al., 2024; Gopavaram, Dev, Grobler, Kim, Das and Camp, 2021; Greitzer, Li, Laskey, Lee and Purl, 2021). Meanwhile, Rocha Flores, Holm, Nohlberg and Ekstedt (2015) found that phishing susceptibility factors differ based on culture.

Experiential factors shape an individual's personality due to past events or experiences. These factors include their previous knowledge of phishing, as well as their technological, professional, and general experiences. Thomas (2018) interviewed security professionals to identify factors that distinguish vulnerable users. They found that technology-savvy users are better equipped to resist phishing attacks, and users in high-impact roles, handling sensitive data, and interacting with company leadership are more susceptible to phishing attacks. Critically, training emerged as a crucial tool to help users resist spear phishing. Similarly, Jampen, Gür, Sutter and Tellenbach (2020) provided targeted training suggestions, emphasizing segmentation based on information literacy competence, job roles, familiarity with phishing, and confidence levels. This segmentation enables the development of focused screening and training programs to thwart spear phishing attacks. Other work has considered the factors that induce users to distrust phishing emails, such as the use of a large body of text (base-rate neglect) (Arduin, 2023) or fictitious shared experiences (Wright, Jensen, Thatcher, Dinger and Marett, 2014).

Psychosocial factors deal may influence an individual's susceptibility to phishing attacks (Fan et al., 2024). These factors include their personality, psychological and persuasion methods, attitude, affective states, habit, and cognitive effort. Several studies on personality traits and phishing susceptibility adopt the Big Five Personality Traits proposed by Digman (2004). For example, Ge, Lu, Cui, Chen and Qu (2021) found that low conscientiousness, low openness, and high neuroticism were associated with increased susceptibility, with these traits indirectly influencing susceptibility through their impact on the cognitive processing of an email. Individuals' e-mail habits, rooted in conscientiousness and emotional stability, significantly influence the outcome of a phishing attack (Vishwanath, 2015; Vishwanath, Harrison and Ng, 2015). Similarly, Neupane, Rahman, Saxena and Hirshfield (2015) used neural activity and eye gaze patterns to characterize vulnerable users to phishing emails. Their

study revealed that vulnerable users tend to spend insufficient time analyzing phishing indicators. Also, metacognition plays an important role in human's ability to detect phishing emails Canfield, Fischhoff and Davis (2019). This correlates with the findings that threat detection significantly reduces phishing susceptibility (Musuva, Getao and Chepken, 2019), explaining why individuals who spend cognitive effort on processing phishing emails are less likely to be phishing victims.

In terms of psychological persuasion methods and phishing susceptibility, recent papers have studied the principles of influence, scams, and frauds such as those proposed by Cialdini (1993); Gragg (2003); Stajano and Wilson (2011) and showed they have a significant impact on phishing susceptibility (Ferreira et al., 2015; Wright et al., 2014; Burda, Chotza, Allodi and Zannone, 2020; Lin et al., 2019; Taib, Yu, Berkovsky, Wiggins and Bayl-Smith, 2019; Lawson, Pearson, Crowson and Mayhorn, 2020). However, none of these works considered how these psychological persuasion methods or factors are specifically exploited by cybercriminals in phishing emails to elicit emotional responses in individuals. Furthermore, Zhuo, Biddle, Koh, Lottridge and Russello (2023) in their recent comprehensive systematic literature review on human-centred phishing susceptibility suggested that more research is needed to keep track of the effectiveness of persuasion principles used in phishing email construction. Thus, the goal of this paper is to fill up these gaps as several studies have shown that psychosocial factors are leading risks to phishing susceptibility (Braca and Dondio, 2023; Tornblad, Jones, Namin and Choi, 2021; Bright, Wziatka and Ngaruko, 2022; Eftimie, Moinescu and Răcuciu, 2022).

Other categorizations exist such as the Phishing Susceptibility Model (PSM) proposed by Zhuo et al. (2023) that categorizes factors influencing phishing susceptibility into three temporal stages. The first includes long-term stable factors such as acquired knowledge, individual differences, demographics, personality traits, and habits. The second involves situational characteristics and access methods in which users check their emails, which can interact with long-term stable traits. The third refers to in-the-moment state factors when dealing with a specific potential phishing email, such as cognitive effort, persuasion methods, and visual presentation.

2.2. Detection of affective states

Salloum, Gaber, Vadera and Shaalan (2021) surveyed prior work on phishing detection, suggesting the need for Natural Language Processing (NLP)-based approaches. Chatterjee and Basu (2021) modeled the use of Cialdini's six principles of persuasion in phishing emails using a fine-tuned pre-trained BERT model and found that the majority of phishing emails exploit the principle of scarcity followed by the principle of reciprocity. Sharma, Kumar, Gonzalez and Dutt (2022) explored cognitive biases, highlighting the effectiveness of authority bias over hyperbolic discounting in deceiving users. Similarly, Sharma and Bashir (2020) focused on trigger elements in email subjects and bodies, suggesting that attackers exploit commonalities in users' behaviour based on fear, anticipation, and trust.

Shahriar, Mukherjee and Gnawali (2022) aimed to enhance phishing email detection by incorporating Phishing Psychological Trait (PPT) scores. They identified *A Sense of Urgency*, *Inducing Fear*, and *Enticement with Desire* as dominant affective states, using BERT, Sentence-BERT, and Character-level-CNN models for training. Results demonstrated significant improvement in phishing detection with PPT scores, particularly highlighting fear as the strongest cue. Similar work was done by Kashapov, Wu, Abuadbba and Rudolph (2022) to analyze psychological triggers and generate email summaries for discerning phishing and non-phishing emails using T5 LLM. Similarly, Van Der Heijden and Allodi (2019) employed Cialdini (1993) principles to ascertain if cognitive vulnerability triggers can predict phishing email success. They utilised NLP and econometrics to create a triaging mechanism, analyzing a dataset from a European financial organization's phishing response division. The study suggests that assessing the email body cognitively can improve response team operations and awareness campaigns.

As shown in Table 14 in the Appendix, we provide a summary of 30 relevant studies in the field. We found that previous research on detecting persuasive techniques in the phishing emails predominantly relied on Cialdini (1993) principles of persuasion. However, these principles lack specificity in addressing phishing attacks and overlook the emotional responses that cybercriminals aim to elicit in email users when sending phishing emails. In contrast to existing works, we identified comprehensive phishing-specific affective states and gave a thorough analysis of the capability of different state-of-the-art LLMs in detecting these affective states within phishing emails.

3. Identified affective states exploited in phishing emails

The identified affective states combine the findings of prominent theories to characterize human susceptibility factors to influence, scams and frauds. These theories include principles of influence Cialdini (1993), psychological triggers Gragg (2003), principles of scams Stajano and Wilson (2011), principles of persuasion in social engineering Ferreira et al. (2015) (a synthesis of the previous three), and cognitive hacks Schneier (2023). The principles in each theory are listed in the columns of Table 1.

As shown in Figure 2, we identified six main emotion-driven affective states: *Fear, Urgency, Greed, Curiosity, Trust*, and *Compassion*. These states are selected because they are the most frequently exploited by cybercriminals in phishing emails and explain the emotional responses attackers aim to elicit from individuals in phishing emails. Cybercriminals exploit them to deceive and manipulate victims into disclosing sensitive information or taking actions that jeopardise their security. In figure 2, for each affective

Table 1
Theories of influence, scams and frauds. The table should be read column-wise as each column contains the principles proposed in that theory.

S/N	Cialdini (C)	Gragg (G)	Stajano & Wilson (SW)	Fereriera et al. (F)	Schneier (S)
1	Authority	Authority	Social Compliance	Authority	Trust and Authority
2	Social Proof	Diffusion Responsibility	Herd	Social Proof	Persuasion
3	Liking and Similarity	Deceptive Relationship	Kindness	Liking, Similarity and Deception	Fear and Risk
4	Commitment and Consistency	Integrity and Consistency	Dishonesty	Commitment, Reciprocation and Consistency (CRC)	Attention and Addiction
5	Scarcity	Overloading	Time	Distraction	-
6	Reciprocation	Reciprocation	Need and Greed	-	-
7	-	Strong Affect	Distraction	-	-

state, we indicate the susceptibility factors that refer to it by providing the column's letter and the row's number of Table 1. For example, *Curiosity* is referred to by the factor "Strong Affect" proposed in Gragg's theory (G7).

Understanding these affective states is crucial in defending against phishing attacks, as it allows individuals and organizations to equip themselves with the knowledge needed to identify and thwart cybercriminals' efforts. By recognizing the affect-based psychological tactics used in phishing attacks, individuals can take steps to protect themselves, such as verifying the sender's identity, double-checking email addresses, and avoiding clicking on suspicious links or downloading attachments from unknown sources.

3.1. Fear

Fear is a complex and powerful emotional response to perceived threats or dangers. It is a natural and adaptive mechanism that has evolved to help organisms respond to situations that may pose a risk to their well-being (Gross and Canteras, 2012). Bitaab, Cho, Oest, Zhang, Sun, Pourmohamad, Kim, Bao, Wang, Shoshitaishvili et al. (2020) observed a surge in phishing attacks amid the COVID-19 pandemic, exploiting users' pandemic-related uncertainty and fear. Criminals exploit this affective state strategically, inducing fear of consequences to compel victims into revealing sensitive information or compromising their security.

Only Gragg (2003) and Schneier (2023) explicitly discussed the concept of fear. Gragg (2003) categorized fear as a *Strong Affect* (G7), a trigger leveraging heightened emotional states to facilitate hackers exceeding reasonable boundaries. Schneier (2023) extends the discussion, focusing on fear and risk (S3), illustrating how attackers leverage this emotion to achieve their goals. Attackers instil fear in phishing emails using attention-grabbing subjects like *"Your Account Security Compromised"* and alarming statements in the body, such as *"Your secrets are exposed! Pay now to prevent the release of damaging information."* These fear-inducing tactics aim to prompt hasty actions, bypassing rational scrutiny.

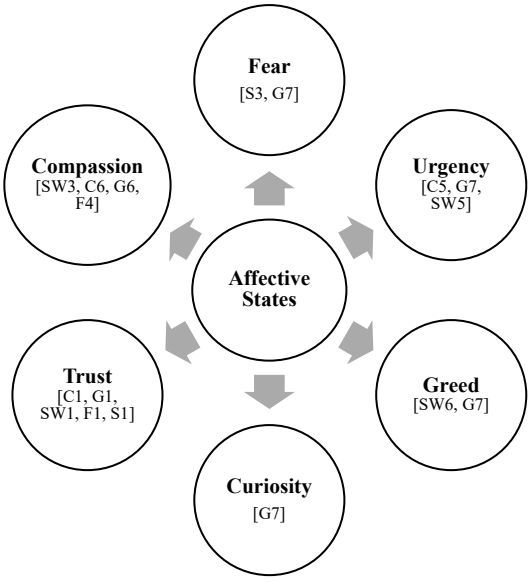


Figure 2: Identified affective states.

3.2. Urgency

Urgency entails the immediate need to act or respond without delay, often involving tasks with short expiration. Urgency represents a significant affective state exploited by attackers during phishing attempts, capitalizing on individuals' tendency to respond hastily to time-sensitive requests (Zhu, Yang and Hsee, 2018; Williams, Hinds and Joinson, 2018). For example, email subjects like *"Immediate Action Required"* and statements such as *"Urgent Request"* emphasize urgency.

Cialdini (1993) introduced *Scarcity* (C5) as a persuasion principle, compelling immediate action due to limited time or resources. This concept aligns with Gragg (2003) *Strong Affect* (G7), where urgency prompts quick responses, leaving little time for rational thought. Similarly, Stajano and Wilson (2011) *Time* (SW5) factor represents time scarcity. These

different factors are used by attackers to solicit *Urgency* in the victims.

While there may be similarities between urgency and fear, such as their ability to trigger impulsive decision-making and interfere with rational thinking, there are also differences. Urgency is often related to a perceived need to act quickly or fear of missing out, while fear is related to a perceived threat or danger (Beesdo, Knappe and Pine, 2009).

3.3. Greed

Lambie and Haugen (2019) defined greed as an excessive desire that (a) extends beyond monetary wealth or material possession, (b) includes perpetual dissatisfaction, (c) disregards the potential costs, and (d) motivates both acquisition and retention motivations. Cybercriminals exploit *Greed* in phishing emails by leveraging individuals' excessive desire for gain. Subjects like *"Exclusive Offer: Triple Your Income!"* and statements such as *"Claim Your Jackpot Prize!"* play on the recipient's desire for financial gain. The body often elaborates on lucrative opportunities, promising quick profits or exclusive rewards.

Stajano and Wilson (2011) discussed the affect *Greed* under the *principle of Need & Greed*. This concept closely aligns with Cialdini's *principle of Scarcity* Cialdini (1993), as scarcity can intensify people's desire to possess limited resources.

3.4. Curiosity

Curiosity is a fundamental human trait characterized by a cognitive and emotional desire for information, knowledge, or novelty. It triggers the brain's reward system, releasing dopamine when encountering novel stimuli (Gruber, Gelman and Ranganath, 2014). It plays a crucial role in learning, creativity, and motivation (Kang, Hsu, Krajbich, Loewenstein, McClure, Wang and Camerer, 2009), contributing to positive outcomes like enhanced cognitive abilities (Kidd and Hayden, 2015). However, criminals exploit curiosity in phishing attacks (Benenson, Gassmann and Landwirth, 2017; Moody, Galletta and Dunn, 2017; Sarno, Harris and Black, 2023), using enticing subject lines like *"You won't believe what happened!"* or intriguing file names to compromise security. Despite the prominence susceptibility, only Gragg (2003) explores curiosity in the *Strong Affect* principle (G7).

3.5. Trust

Trust is fundamentally a behavioral concept, involving confidence in others based on perceived trustworthiness and expectations (Li, Betts et al., 2003). Social engineering success, like phishing, hinges on exploiting trust, leveraging people's inclination to obey trusted authorities such as banks or police (Schneier, 2023), and taking advantage of this "suspension of suspiciousness" to manipulate people into compliance (Stajano and Wilson, 2011). Email subjects like *"Security Verification Required"* and statements such as *"Important Account Update from X Bank"* create a false sense of legitimacy. This is well-examined across all reviewed theories (see row 1 in Table 1).

3.6. Compassion

According to Perez-Bret, Altisent and Rocafort (2016), compassion could be defined as "the sensitivity shown to understand another person's suffering, combined with a willingness to help and to promote the well-being of that person, to find a solution to their situation" (p. 599). As observed by Strauss, Taylor, Gu, Kuyken, Baer, Jones and Cavanagh (2016), compassion is frequently linked with terms such as empathy, kindness, pity, and altruism. Although they acknowledge subtle distinctions between these terms, in this study, we employ the term compassion to encompass all of them collectively.

In phishing emails, emotional appeals, such as narratives of hardship, appeals to help in war-troubled zones, or donating to charity, can manipulate empathy, prompting victims to unwittingly compromise their security by clicking on malicious links, downloading malicious attachments, supplying information or transferring money to the attacker. This affective state, as discussed by Stajano and Wilson (2011) under the principle of kindness, is closely linked to the principle of reciprocation (von Bieberstein, Essl and Friedrich, 2021; Cialdini, 1993; Gragg, 2003; Ferreira et al., 2015).

However, not all the principles from Table 1 (e.g., *Distraction, Overloading, Dishonesty*) are directly linked to the identified affective states, although they may still be exploited by attackers in phishing emails. For example, attackers can exploit the *Distraction* principle to divert users' attention from their true intent while posing as a trusted entity or offering assistance. The rationale behind this choice is that we only considered affective states associated with susceptibility factors that can be identified automatically in phishing emails.

4. Methodology

This section discusses the methodology used to explore affective states in phishing emails. We discuss the dataset, the selected LLMs, the prompt provided as input to the LLMs, the measures used to evaluate inter-rater reliability, the valence-arousal space, and the use of special characters in phishing emails investigated in this study.

4.1. Dataset

This paper utilizes an email-based phishing corpus derived from the work of Ciambrone and Wilson (2023), who compiled a dataset of phishing emails from five universities. They used this dataset to perform topic modeling and visualize topic evolution over time. We expanded this dataset by adding phishing emails from a sixth university. The combined corpus, totaling 5187 emails, underwent parsing and formatting to create a tabular structure, including source link, attachment format, subject, body, and email date. Each email was assigned a unique eight-character identifier for identification purposes.

Table 1 provides descriptive statistics on the dataset. Some emails lack a subject or body, with 196 missing a subject and 12 lacking a body, suggesting phishing attempts

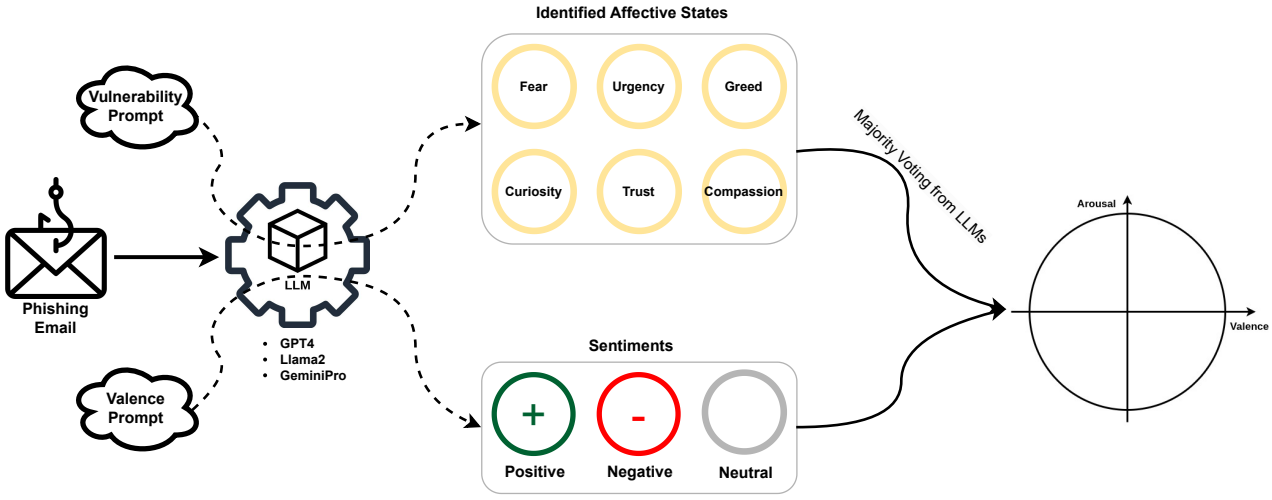


Figure 3: Experimental design: LLMs are prompted to determine the sentiment and anticipated responses the attackers are trying to elicit in receivers. For analysis, we consider majority voting by the LLMs to determine valence-arousal coordinates.

Table 2
Dataset descriptive statistics

Email	Total Number	Empty	Average character length	Maximum character length	Average word length	Maximum word length
Subject	5187	196	25	429	4	75
Body	5187	12	404	12902	76	2703

often rely on email bodies. On average, subjects contain 4 words, with a maximum of 75 words. Similarly, the average body length is 76 words, with a maximum observed count of 2,703 words.

4.2. Large language models (LLMs)

One of this study's key goals is to assess current LLMs' ability to identify affective states in phishing emails and evaluate their performance against human annotators. We used the latest LLMs as at the time of performing these experiments released by three major vendors: GPT-4 by OpenAI OpenAI (2023), Llama 2 by Meta Touvron, Martin, Stone, Albert, Almahairi, Babaei and Scialom (2023), and Gemini Pro by Google Gemini (2023).

Figure 3 illustrates the experimental design adopted in this study. We performed zero-shot classification of email subjects and bodies using the prompts described in section 4.2.1 to identify affective states as well as sentiments of the phishing emails. LLMs were also requested to analyze emotions according to valence and arousal (explained in section 4.4), in the email subject and body, similar to the affective response assessment in the Self-Assessment Manikin test (Bradley and Lang, 1994). The LLM is presented with a new task without any prior training or fine-tuning on that specific task, and generate an output (Chae and Davidson, 2023). While fine-tuning can improve the performance of the LLM on some specific tasks, it requires a large amount of labeled data and can be expensive both computationally and financially (Zhang, Talukdar, Vemulapalli, Ahn, Wang, Meng, Murtaza, Leshchiner, Dave, Joseph et al., 2024). We utilised state-of-the-art LLMs, used effective prompts, and focused on concepts such as fear, urgency, etc.

The predictions, regarded as the "silver truth" Tekumalla and Banda (2023), were compared against manually annotated labels in a subset of emails for comparative analysis.

4.2.1. Prompt

The prompts provided as input to the LLMs consist of a *context* defining the system's role and specifying the scope of the query conversation. The *content* embodies the user role and encompasses either the email subject or the body. Figure 4 illustrates three types of prompts. The *Identification Prompt* solely requests labels to identify the type of affective state. While the *Explanation Prompt* requires the LLM to locate sentences hinting at the classification and provide an arousal rating. The arousal rating offers insight into the intensity of the email text. Additionally, the *Valence Prompt* aims to determine the valence of the email's subject or body.

4.2.2. LLM hallucination

LLMs occasionally generate persuasive yet inaccurate or contextually irrelevant outputs, referred to as *hallucinations* (Ji, Lee, Frieske, Yu, Su, Xu, Ishii, Bang, Madotto and Fung, 2023). We observed *Self-contradiction* hallucination in the LLMs responses that occur when the models offer different, sometimes conflicting responses upon rephrasing prompts or introducing additional context.

Figure 5 illustrates self-contradiction observed in *Gemini Pro's* responses to the email body. While the model identifies only *Curiosity* as an affective state, its response contradicts itself when asked for an explanation and arousal rating, returning *Urgency, 3: Kindly download the attached and confirm your delivery details to avoid further delay*. To analyze this self-contradiction, we conducted comparisons

Identification Prompt

Context: You will be provided with the [subject | body] of an email, and your task is to classify the email [subject | body] into the following psychological vulnerabilities it tries to exploit: 'Fear', 'Curiosity', 'Trust', 'Compassion', 'Urgency', or 'Greed'. Only include the labels in your output.

Content: <Email subject/body>

(a)

Explanation Prompt

Context: You will be provided with the [subject | body] of an email, and your task is to classify the email [subject | body] into the following psychological vulnerabilities it tries to exploit: 'Fear', 'Curiosity', 'Trust', 'Compassion', 'Urgency', or 'Greed'. Provide the lines in the email where each vulnerability is found. Also, state the arousal level of the vulnerability: 1=very low arousal, 2=low arousal, 3=medium arousal, 4=high arousal, 5=very high arousal.
Your output should be provided in a way similar to this example - fear, 3: sentence in email where fear was found

Content: <Email subject/body>

(b)

Valence Prompt

Context: You will be provided with the [subject | body] line of an email, and your task is to classify the line into its sentiment: either 'positive', or 'negative', or 'neutral' sentiment. Only include the labels in your output.

Content: <Line of email subject/body>

(c)

Figure 4: a) Identification prompt, b) Explanation prompt and c) Valence prompt used in the study.

of responses with and without explanations, utilizing both hard and soft comparisons. In the soft comparison, two responses are not considered hallucinations if there is at least one affective state overlap, whereas hard comparison considers responses that are not a perfect match as hallucinations.

Dear Valued Customer,
Please receive your package: Your shipment document parcel have arrived at the post office. Unfortunately, our courier was unable to deliver the parcel to you due to incorrect delivery details. Kindly download the attached and confirm your delivery details to avoid further delay.

Best regard,
DHL Delivery Dept.

Identification Prompt: Curiosity

Explanation Prompt: Urgency, 3: *Kindly download the attached and confirm your delivery details to avoid further delay.*

Figure 5: Responses of Gemini Pro using the *Identification* and the *Explanation* prompts for the above email.

4.3. Reliability analysis

Reliability refers to the level of consistency among various raters or coders when interpreting a shared set of items Marchal, Scholman, Yung and Demberg (2022). Evaluating reliability becomes crucial as different LLMs and humans independently annotate the same information. Various methods are available for assessing inter-rater reliability, with considerations such as data type (single-label or multi-label; nominal or ordinal), the number of raters, and sample size playing key roles in the selection of an appropriate approach. In this study, we used the following methods for inter-rater reliability.

Cohen's κ is a metric for two raters inter-rater reliability Cohen (1960), given by the following formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where, P_o is the observed agreement, the proportion of cases where the raters agree and P_e is the expected agreement, the agreement by chance. This was used to find the pairwise inter-rater reliability for the LLMs and also for human annotators in single-label sentiment analysis.

Cohen's Kappa measures agreement only between two raters. Fleiss' Kappa (κ) measures agreement for situations involving more than two raters Fleiss (1971). This study used Fleiss' κ to ascertain the overall reliability among all three LLMs in their sentiment analysis.

Krippendorff (2011)'s α was used for both the affective states analysis's pair and overall inter-rater reliability since it involves multi-labels and multi-raters.

The values of Cohen's κ, Fleiss' κ and Krippendorff's α can range from -1 to 1, (κ or α) $= 1$ signifies perfect agreement, (κ or α) $= 0$ indicates no agreement beyond chance and (κ or α) < 0 implies there is less agreement than random chance.

We performed soft and hard comparisons to assess the percentage of agreements between raters, with higher percentages indicating better agreement. In a soft comparison, responses are considered the same if there is at least one common affective state. In contrast, a hard comparison deems them equal only if they are a perfect match (Marchal et al., 2022). For instance, if the identification prompt reveals affective states as *Trust, Curiosity*, and the explanation prompt identifies *Curiosity*, a soft comparison treats them as a match, whereas a hard comparison does not. Soft and hard comparisons yield the same result for single-label cases.

4.4. Valence-arousal space

Through *Explanation* prompt in Figure 4(b), we produced responses accompanied by arousal ratings. Additionally, a distinct prompt was utilized to extract the overall sentiment of both the email subject and body separately shown in Figure 4(c). Consequently, with the valence and arousal rating for each email, we derived arousal ratings for every affective state across all three valence labels. This implies that the tone of the email subject/body used to elicit a specific affective state may be negative, positive, or neutral.

Hence, these two values were utilized to map the proposed affective states to a valence-arousal cartesian space.

For the valence-arousal analysis, we only considered the email subjects and bodies for which the models have consensus in only a single label. In the comparison based on majority consensus, we calculated the average arousal rating for each email subject/body. This average is determined by considering the consensus labels of the models that propose the agreed-upon affective state for that specific subject/body.

4.5. Special character analysis

Minor alterations in the email subject or body can allow them to evade phishing detection systems easily (Ghazi-Tehrani and Pontell, 2022). Consequently, we sought to assess LLMs' resilience against using special characters as deceptive tactics.

We categorized special character usage in email subjects and bodies into three types. Using Gemini Pro, we obtained emails with special characters and refined the subset through manual observation, deriving the types based on visual clustering.

1. **Special character:** Attackers substitute certain letters in the email subject or body with diacritics or other non-English letters. Example: *Hí! Únförtűnately, Í have sóme bad news fór yöű*
2. **Leetspeak-like writing:** Attackers introduce additional letters or rearrange words to ensure readability by humans but without words in standard embeddings. Example: *Wde notiiced incompp6leptei sjeccuri8ty set-sutp o7n your Cohwacsfe7 banrk6ingh accouhnt, we strongely recommenid trhfat you7 ujpdnate yrouer3 banrk accountg fo3r security pzurupose3.*
3. **Punctuations:** Attackers use punctuation marks to avoid triggering highly flagged words or to elude embeddings. Example: *A d_oc_ume _nt h_as b_ee _n s _en_t t_o y_o_u*

These three groups were then analyzed.

5. Results and discussion

5.1. Comparative analysis of LLMs

5.1.1. affective states distribution

Figure 6 shows the distribution of the proposed affective states detected by the LLMs. The *Majority Consensus* (refer to Algorithm 1 in Appendix A) indicates that at least two out of three LLMs agreed on the affective state. Panels 6a and 6b depict the affect distribution, considering only a single primary label (assuming the first affective state in the email text is the most important). Panels 6c and 6d represent the same affect distribution, considering multiple affective states. For instance, if Llama 2 suggests both Urgency and Greed as possible labels, both are counted. The *Unclassified* label refers to instances where the LLM could not identify any affective states in the email subject/body. Additionally, the *Disagreement* label (not displayed) is exclusive to the Majority Consensus, indicating a lack of consensus among the three LLMs regarding any affective states.

When examining email subjects (refer to panels 6a and 6c) based on Majority Consensus, the most exploited affective state is Curiosity, followed by Urgency and then Trust. This finding is aligned with the theory that curiosity is a primary trait leveraged in phishing attacks (Moody et al., 2017). However, Llama 2 presents a slightly different distribution, with Urgency being the predominant affective state targeted, followed by fear. These findings also indicate that compassion and greed are less frequently utilized in email subjects. Both panels reveal that the affect distribution across all models remains relatively consistent even when considering all responses, with only a few additional labels.

When examining the email bodies (see panels 6b and 6d), Fear is the most exploited affective state in the primary labels followed by Urgency and Trust while in multi-labels Urgency is the most exploited closely followed by Trust and Fear. In contrast to subject distributions, there is a significant shift in label distributions between single labels and all labels, with a substantial increase in the number of labels. This is primarily because LLMs recognize multiple affective states in email bodies, which tend to be more text-heavy than subjects.

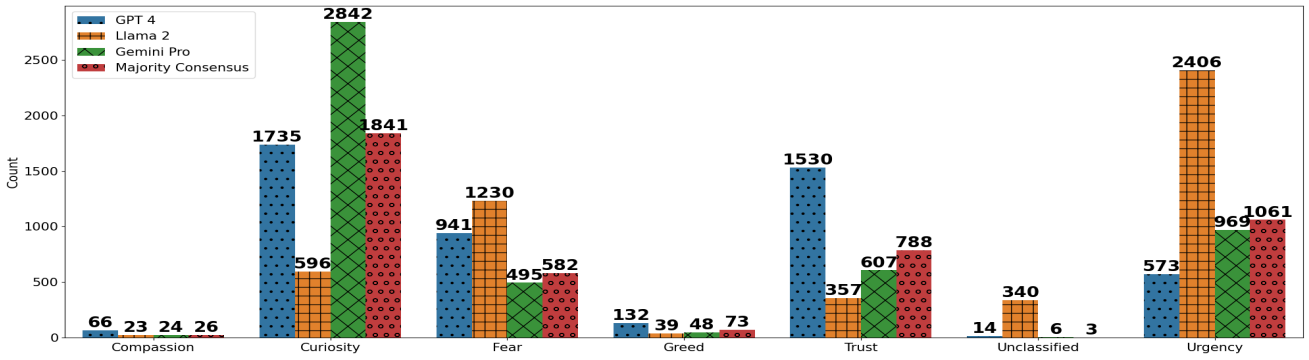
5.1.2. Email sentiment distribution

Figure 7 shows the distribution of sentiments recognized by different LLMs in the email subject and body. Panels 7a and 7b collectively suggest that attackers predominantly adopt a neutral tone when crafting phishing emails. This inclination might indicate the attacker's strategy of employing a polite tone to minimize suspicion and make the email appear normal (Turnage, 2007). Notably, there is a decrease in the use of a neutral tone in the body compared to the subject, dropping by 12.5%, while positive and negative sentiments increased by 103% and 22%, respectively. The increase in positive polarity aligns with the rise in greed when considering multi-labels for email bodies, as observed in Figure 6c and 6d.

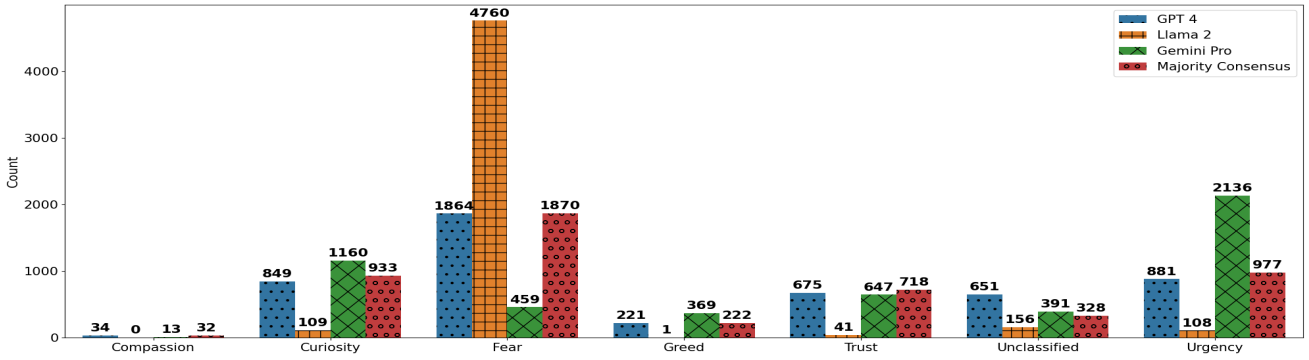
5.1.3. Paired accompanied affective states analysis

As indicated in Figure 6, LLMs often generate more than one response for email subjects and bodies, with a higher frequency observed for bodies. Figure 8 examines how frequently two affective states co-occur, leveraging the *Majority Consensus* approach to glean insights into attackers' reliance on multiple affective states for successful attacks. Nodes represent affective states, edges signify co-occurrence, and edge weights denote the frequency of these paired affective states being exploited together. A self-loop denotes instances where only one affective state is recognized.

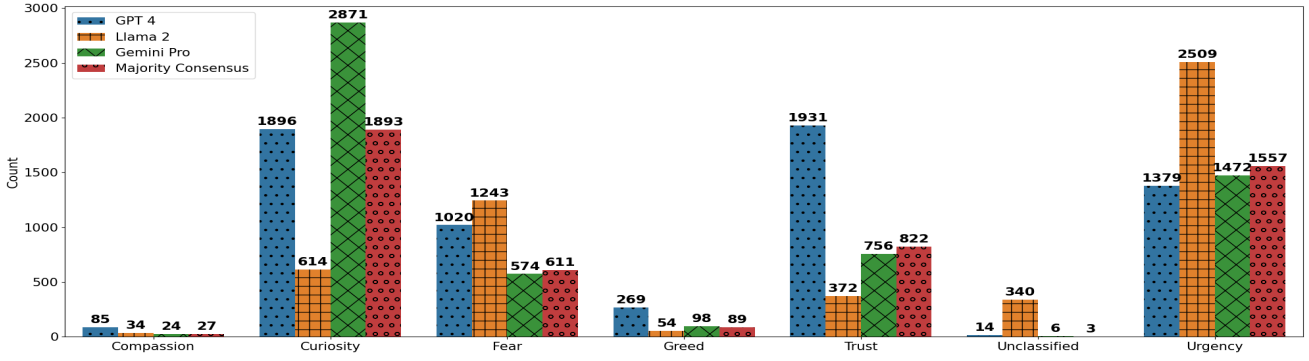
As shown in Figures 8a and 8b, attackers predominantly rely on only a single affective state for subjects due to word length limitation while utilizing multiple affective states in email bodies. Thus, the prevalence of email bodies relying on a single affective state decreases by 82% in comparison



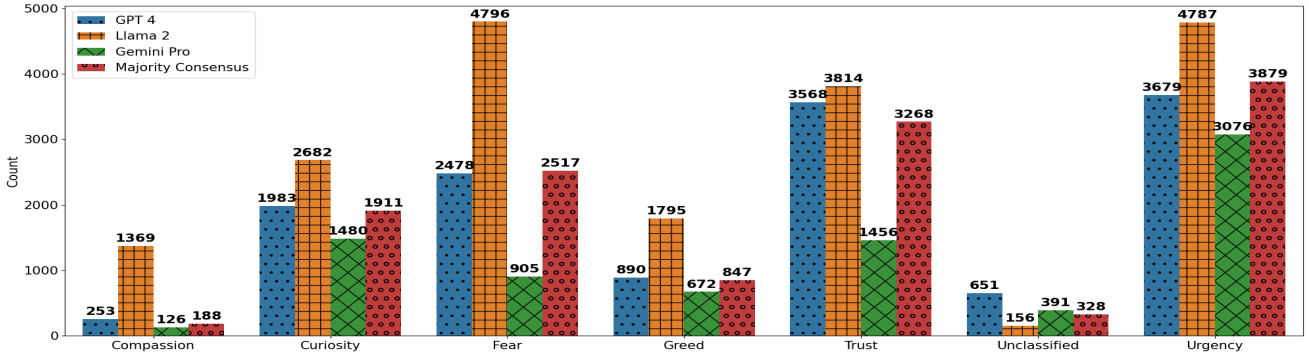
(a) Affect distribution of email subjects for various models considering only primary labels



(b) Affective distribution of email bodies for various models considering only primary labels

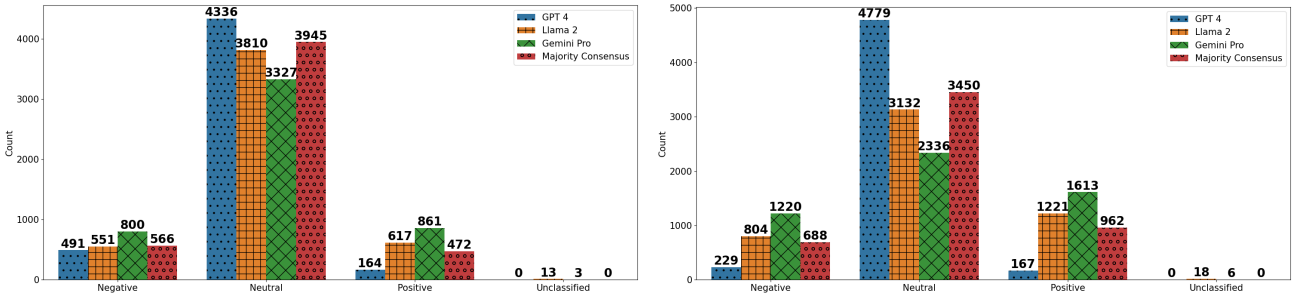


(c) Affect distribution of email subjects for various models considering all multi-labels



(d) Affect distribution of email bodies for various models considering all multi-labels

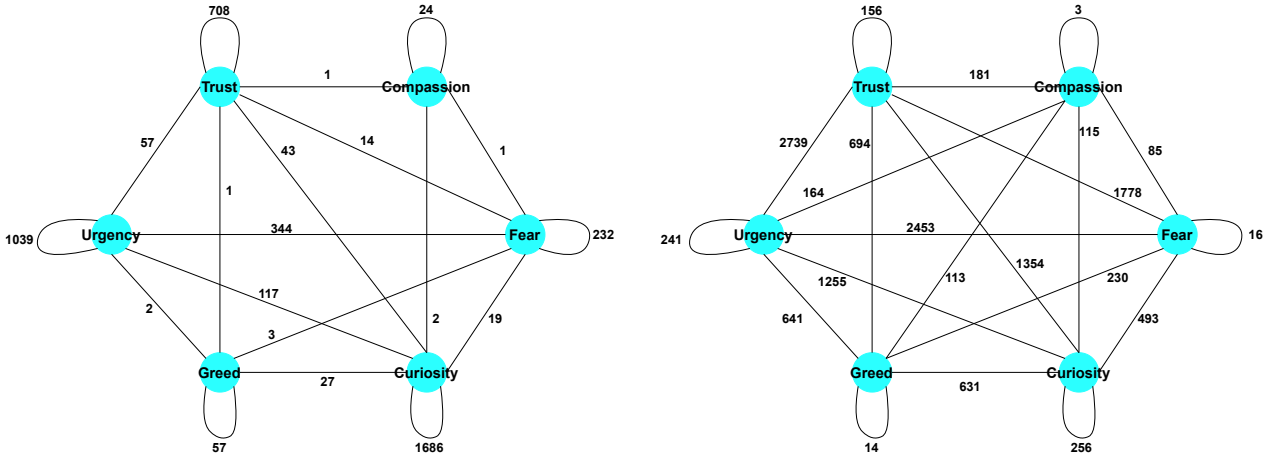
Figure 6: Distribution of affective states in email subjects and bodies among different LLMs. *Majority Consensus* refers to labels where the LLMs reached a consensus. *Unclassified* refers to the case where the LLMs failed to identify any affective states. The y-axis refers to the LLM classification count while the x-axis refers to the affective state.



(a) Valence distribution of email subjects for various LLMs

(b) Valence distribution of email bodies for various LLMs

Figure 7: Valence distribution of email subjects and bodies for the LLMs and the majority LLM vote. Attackers often adopt a neutral tone when crafting subjects and bodies of phishing emails to evade suspicion and appear as regular emails.



(a) Paired accompanied affective states in email subjects

(b) Paired accompanied affective states in email bodies

Figure 8: Paired accompanied affective states in email subjects and bodies for *Majority Consensus* labels based on LLMs. Attackers tend to exploit a single affective state when creating email subjects while leveraging multiple affective states in email bodies. This strategy is influenced by the fact that email subjects serve as the initial point of contact, and also due to character limitations in email subjects.

to email subjects, with the most substantial reduction observed for Curiosity. In the case of subjects, *Fear-Urgency* pairing is the two most commonly exploited affective states simultaneously, followed by *Urgency-Curiosity, Urgency-Trust, and Trust-Curiosity*. The dominance of Urgency in the top four pairings goes to show the motive of attackers in crafting phishing email subjects. Be it to trigger *Fear, Trust* or *Curiosity*, they want individuals to take immediate action upon seeing the phishing email.

Similarly, the *Urgency-Fear* pairing is only the second most frequent combination in the email body, while *Urgency-Trust* is the most common for email body responses. The *Urgency-Trust* combination implies that attackers attempt to convey trustworthiness in the email content while urging the victim to take swift action, a pattern reflected in Figure 6d. However, *Fear-Urgency, Urgency-Trust* and *Trust-Curiosity* pairings are among the top four combinations in both subject and body.

5.2. Reliability analysis

Tables 3 and 4 present the inter-rater reliability of LLMs for sentiment analysis of the email bodies and subjects, respectively. The overall Fleiss's Kappa indicates fair agreement ($\kappa = 0.2662$) for the email body and moderate agreement for the subject ($\kappa = 0.5664$). Cohen's Kappa scores between pairs of LLMs vary in the level of agreement, with the highest agreement between Llama2 and GeminiPro and the least agreement between GPT-4 and GeminiPro for both the email body and subject scores. Therefore, the email subject annotations exhibit higher agreement compared to the email body as the LLMs unanimously agree on 45.04% of phishing email body annotations and 74.99% of email subject annotations.

Similarly, Tables 5 and 6 illustrate that the subject affective state analysis achieves relatively better agreement compared to the body analysis. Krippendorff's α values are higher in the subject compared to body analysis, ranging from 0.0923 to 0.3124, suggesting a slight to a fair level of agreement. Also, Hard comparison percentages indicating

Table 3
Pairwise Inter-rater reliability of the LLMs for phishing email body sentiment analysis

Rater 1	Rater 2	κ	Hard (%)
GPT-4	Llama 2	0.2267	67.01
GPT-4	Gemini Pro	0.1517	52.27
Llama 2	Gemini Pro	0.5029	69.35
Overall		0.2662	45.04

Table 4
Pairwise Inter-rater reliability of the LLMs for phishing email subject sentiment analysis.

Rater 1	Rater 2	κ	Hard (%)
GPT-4	Llama 2	0.6178	87.70
GPT-4	Gemini Pro	0.4813	79.28
Llama 2	Gemini Pro	0.6206	82.85
Overall		0.5664	74.99

Table 5
Pairwise Inter-rater reliability of the LLMs for phishing email body affective state identification.

Rater 1	Rater 2	α	Soft (%)	Hard (%)
GPT-4	Llama2	0.0907	87.18	21.57
GPT-4	GeminiPro	0.0826	85.58	16.73
Llama2	GeminiPro	-0.0267	85.50	5.01
Overall		0.0584	77.39	2.83

Table 6
Pairwise Inter-rater reliability of the LLMs for phishing email subject affective state identification.

Rater 1	Rater 2	α	Soft (%)	Hard (%)
GPT-4	Llama2	0.0923	46.19	24.04
GPT-4	GeminiPro	0.3124	76.21	45.36
Llama2	GeminiPro	0.1405	43.44	32.33
Overall		0.1975	34.68	18.39

complete agreement, are much higher in the subject compared to the body. However, Soft comparison percentages, reflecting partial agreement, are much higher in the body compared to the subject, as an email body can elicit several affective states compared to the subject. This result is consistent with our findings that attackers often rely on a single affective state for email subjects but diversify strategies in email bodies.

To evaluate LLM performance, we compared it with human annotators who assessed 200 randomly sampled phishing emails (8 lacking subjects). Table 7 presents reliability analysis results between human annotators for affective state Identification and Sentiment analyses, subdivided into Subject and Body components.

Agreement between human annotators was higher for the subject than the body in both affective state identification and sentiment analyses. In Sentiment analysis, the Subject component exhibits fair agreement ($\kappa = 0.2621$) with substantial hard agreement (73%), while the Body component shows slight agreement ($\kappa = 0.1855$) with substantial hard agreement (58.5%). In the affective stateidentification, both Subject and Body components demonstrate fair agreement, with Cohen's Kappa values of 0.1884 and 0.1646, respectively. Notably, the Body component shows higher soft (85%) but lower hard (27%) agreement compared to the Subject component.

Human annotators established a consensus based on the *intersection* and *union* of annotations for subject and body affective states respectively. This human consensus was contrasted with the LLM consensus, determined through the Majority Consensus Algorithm (in Appendix A). Inter-rater reliability in affective state identification and Sentiment analyses, presented in Table 8, shows Sentiment analysis displaying higher agreement than affective state analysis. Poor agreement is observed in affective state identification in email bodies for GeminiPro ($\alpha = -0.0055$) and Llama2 ($\alpha = -0.0281$). Overall, GPT-4 outperforms Llama2 and GeminiPro.

5.3. Affective states mapped to valence-arousal space

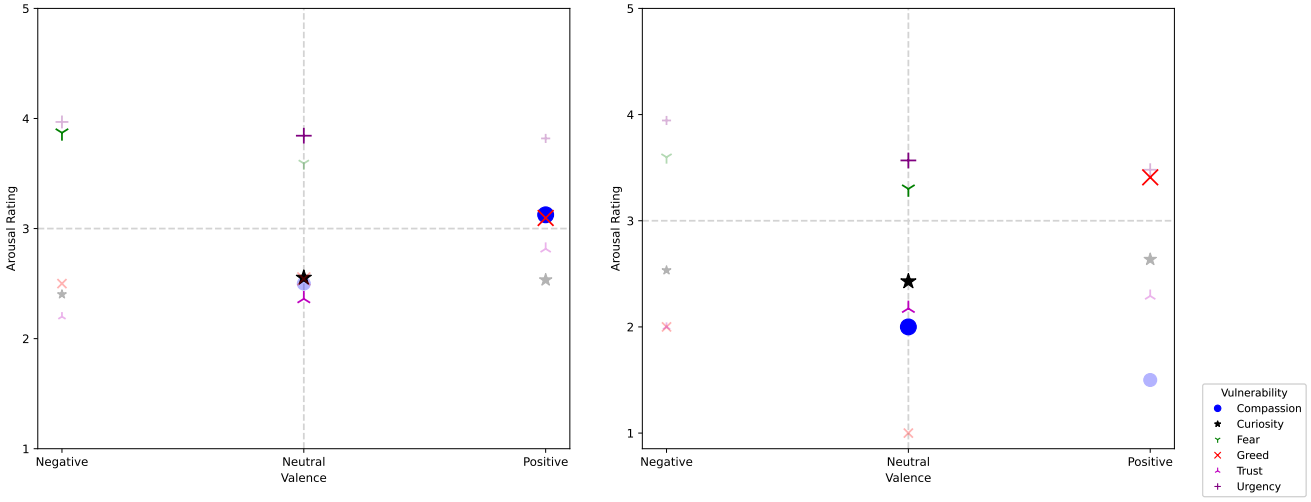
In Figure 9, the mapping of affective states to the valence-arousal Cartesian space is illustrated, based on the *Majority Consensus* results considering multi-label responses for both email bodies and subjects. Different markers in each plot represent distinct affective states. Given that each affective state can be expressed in positive, negative, and neutral sentiments, each plot has at most three similar markers. The size scale of each marker reflects the number of emails in each sentiment category (lowest, medium, and highest). The most opaque marker signifies the most dominant valence for the respective affective state. For individual model-wise comparisons, refer to Appendix B.

There are small similarities in valence-arousal between email subjects and bodies, despite the distinctions in their objectives and character lengths. Notably, Urgency and Fear emerge as the two affective states with the highest arousal levels in both subjects and bodies. In instances where attackers seek to instill urgency or fear in the victim, they employ overt and potent language. Trust and Curiosity exhibit lower, subtler arousal ratings in both cases, while Greed and Compassion predominantly fall within the mid-range of arousal. Additionally, our observations indicate an absence of cases where email subjects/bodies labeled as fear had a positive valence, or those labeled as compassion had a negative valence. This aligns with the general association of negative connotations with fear and positive connotations with compassion.

Our observation indicates that email subjects rarely exhibit affective states with an arousal rating lower than 2. At the same time, there are numerous affective states with arousal ratings equal to or lower than 2 in email bodies. This discrepancy is primarily attributed to the necessity for the arousal level of the email subject to be sufficiently high to persuade recipients to open the email in the first

Table 7
Inter-rater reliability analysis for human annotators.

Analysis	Email	N	κ	α	Soft (%)	Hard (%)
Psy. Vul.	Subject	192	-	0.19	79.5	35.5
	Body	200	-	0.16	85	27
Sentiment	Subject	192	0.26	-	-	73
	Body	200	0.19	-	-	58.5



(a) Valence-arousal cartesian for email subjects

(b) Valence-arousal cartesian for email bodies

Figure 9: Valence-arousal cartesian for email subjects and bodies. The scale of each affective state reflects the frequency of the affect for different valence states, that is, the scale represents frequency count. The frequency is also demonstrated with transparency shed with the most opaque being most frequent. The plots only consider emails that have consensus on single labels.

Table 8
Inter-rater reliability analysis for LLMs and humans.

Analysis	Email	N	κ	α	Soft (%)	Hard (%)
Reliability Between Humans and LLMs						
Psy. Vul.	Subject	133	-	0.25	56.39	43.61
	Body	200	-	0.02	88.38	17.17
Sentiment	Subject	137	0.41	-	-	90.51
	Body	113	0.40	-	-	78.76
Reliability Between Humans and GPT-4						
Psy. Vul.	Subject	133	-	0.23	64.66	42.11
	Body	200	-	0.02	84.85	19.7
Sentiment	Subject	137	0.4	-	-	91.97
	Body	113	0.44	-	-	88.5
Reliability Between Humans and GeminiPro						
Psy. Vul.	Subject	133	-	0.33	67.67	54.14
	Body	200	-	-0.01	77.78	5.05
Sentiment	Subject	137	0.26	-	-	82.48
	Body	113	0.22	-	-	60.18
Reliability Between Humans and Llama2						
Psy. Vul.	Subject	133	-	-0.08	18.8	14.29
	Body	200	-	-0.03	92.93	10.61
Sentiment	Subject	137	0.34	-	-	87.59
	Body	113	0.34	-	-	73.45

place. Examining the most frequent samples, represented by opaque markers, we find that the subject has only three neutral affective states, whereas the body encompasses five out of six affective states.

Table 9
Utilization of special characters in phishing emails.

Type	Subject	Body
Special Character	4	3
Leetspeak-like writing	0	8
Punctuations	0	15

Traditionally, greed is associated with negative characteristics Seuntjens, Zeelenberg, Breugelmans and Van de Ven (2015); however, we observe that positive valence for greed is the most frequent occurrence in both subjects and bodies. This phenomenon can be attributed to attackers employing more positive language when attempting to evoke greed in email subjects, resulting in a higher frequency of greed with positive valence. Fear, recognized as a negative emotion by Demszky, Movshovitz-Attias, Ko, Cowen, Nemade and Ravi (2020), predominantly accompanies negative sentiment in subjects and neutral sentiment in bodies. Both subjects and bodies with negative fear exhibit high arousal.

5.4. Special character exploitation and LLM detection

Table 9 shows the distribution of different types of special characters used in the subject and body of phishing

Table 10
Inter-rater reliability analysis for consensus of LLMs and humans in special character detection. Cohen's Kappa Statistic is not used since the sample size is too small.

Analysis	Email	N	Soft (%)	Hard (%)
Psy. Vul.	Subject	4	75	0
	Body	26	96.15	42.31
Sentiment	Subject	4	-	100
	Body	26	-	84.62

emails. They tend to underutilize special characters in subject lines, presumably to reduce suspicion, as the recipients encounter them first in an email. Conversely, a variety of special character combinations are found within email bodies.

The overall occurrence of emails featuring special characters is low (30 out of 5187). These characters predominantly evoke feelings of urgency, followed by fear and curiosity in recipients. The average arousal rating for urgency within subject lines and bodies is 4 and 4.05, respectively, compared to the overall averages of 3.84 and 3.61. Concerning fear, the ratings for subject lines and bodies are 3.17 and 3.69, respectively, against the overall averages of 3.31 and 3.3. Additionally, for curiosity, the ratings for subject lines and bodies are 2.7 and 4.05, respectively, contrasting with the overall ratings of 2.6 and 2.3.

Table 10 presents reliability findings for the consensus between LLMs and human evaluators regarding the identification of special characters, employing both soft and hard comparisons. The results indicate strong agreement between LLMs and humans in discerning sentiments in emails featuring special characters. While the identification of affective states had agreement in soft comparisons, the agreement was lower in hard comparisons.

5.5. LLM hallucination for affective states

Table 11 indicates a prevalence of self-contradiction hallucination in existing LLMs when identifying affective states. Both soft and hard comparisons reveal a lower occurrence of hallucinations in the email body in comparison to the subject. This discrepancy can be attributed to the increased content in the body, enhancing LLMs' ability to comprehend affective states in the text. The minimal errors in soft comparisons suggest that, although the models exhibit hallucinations, these instances remain low. Compared to the subject, the relatively lower soft comparison error for the body is influenced by LLMs providing more labels for bodies, thereby increasing the likelihood of some overlap.

GPT-4 exhibits poor performance for subjects, as it fails to provide a response for 81.6% of email subjects, citing insufficient information when prompted for an explanation. Llama 2 demonstrates a low hallucination rate for soft comparisons but a notably high hallucination rate for hard comparisons. This discrepancy is largely attributed to 6.39% of the responses for subjects by Llama 2 having more than 3 labels, compared to 82.39% for the body. This substantial difference allows for a higher overlap in soft comparisons but results in a notably poor score for hard comparisons. In the case of short-text subjects, Gemini Pro appears to be more robust, while for long-text bodies, GPT-4 exhibits greater robustness.

5.6. Threats to validity

The findings of this study draw from a dataset of phishing emails targeting universities. Further analysis with more diverse phishing email datasets should be conducted to support our results. Variability in human judgment assessing affective states in phishing emails may impact the robustness of LLM-human comparisons. To mitigate this, two human annotators (co-authors) evaluated a randomly selected sample of 200 emails, using the label *union* for email bodies since the body of emails elicits multiple affective states and the *intersection* of their annotations for subjects since subjects mostly elicit a single affective state. We utilized GPT-4, Gemini Pro, and Llama 2 to ensure LLM reliability during the research. The rapidly evolving nature of cyber threats suggests that attackers may develop new strategies not addressed in the current analysis.

5.7. Implications for phishing detection

The findings of this study have several important implications for improving current phishing email detectors. They extend beyond the conventional reliance on senders' domain names, email addresses, and URLs in detection systems, and suggest avenues for sending more effective warnings to users:

- **Incorporating affective states:** According to Carroll, Adejobi and Montasari (2022) the success of a phishing attack can be measured by the attacker's ability to persuade and manipulate victims into complying with their objectives. Our study identified comprehensive affective states that attackers exploit in phishing emails, such as curiosity, urgency, fear, trust, compassion and greed. Incorporating the identification of these affective states into phishing detection models can enhance their accuracy in identifying malicious emails that target human emotions and cognitive biases.
- **Leveraging LLMs for detection:** The study suggested that state-of-the-art LLMs have some agreement with human annotators in detecting affective states in phishing emails.
- **Analyzing email content and structure:** The study examined how attackers exploit affective states differently in email subjects versus bodies. Detectors should analyze both the subject line and body content for signs of emotional manipulation, as well as the overall structure and flow of the email.
- **Mapping affective states to arousal and valence:** Mapping the identified affective states to the valence-arousal Cartesian space revealed that fear and urgency evoke the highest arousal levels in phishing emails.

Table 11
Self-contradiction hallucination observed in email responses generated by LLMs.

		ModelType	GPT-4	GeminiPro	Llama2
Subject	Soft comparison	#Self-contradiction	4353	2157	1289
		Hallucination %	83.92	41.58	**24.85**
	Hard comparison	#Self-contradiction	4758	2679	4741
		Hallucination %	91.73	**51.65**	91.4
	Hallucination % Change		9.31	24.22	267.81
Body	Soft comparison	#Self-contradiction	623	1246	284
		Hallucination %	12.01	24.02	**5.48**
	Hard comparison	#Self-contradiction	3322	3319	5047
		Hallucination %	64.04	**63.99**	97.3
	Hallucination % Change		433.22	166.41	1675.55

Detectors can use this insight to prioritize and weight these emotional triggers more heavily in their analysis.

- **Detecting paired affective states:** The study found that certain affective state pairings, such as fear-urgency, urgency-trust, and trust-curiosity, are commonly exploited together in phishing emails. Detectors should look for these combinations as they may be more indicative of a phishing attempt than individual affective state alone, mostly with respect to phishing email bodies.
- **Tailoring user warnings:** By understanding how attackers leverage specific emotional triggers and affective state pairings, phishing warnings can be tailored to the email content. For example, if a suspicious email is detected as exploiting fear and urgency, the warning can emphasize these emotional manipulation tactics to help users recognize the phishing attempt. By understanding how different emotional states are exploited by attackers, developers can design more targeted warning messages to alert users about potential phishing threats. For example, emails flagged as exploiting fear or urgency could trigger warnings emphasizing caution and verification steps, while those exploiting trust could prompt reminders about verifying sender authenticity.

6. Conclusion

This paper explored how attackers exploit affective states in phishing emails using LLMs. We identified six comprehensive phishing-specific psychological susceptibility factors (Fear, Urgency, Greed, Curiosity, Trust, and Compassion) based on the emotional states targeted by cybercriminals in phishing emails. These affective states were inspired by five previous theories on human susceptibility to influence, scams, and fraud. We then assessed how three LLMs (GPT-4, Llama 2, and Gemini Pro) could detect these affective states automatically using a corpus of phishing emails targeting six universities. Our results show that phishing emails predominantly exhibit neutral sentiments, with curiosity being the most exploited affective state in subjects, followed by urgency. In the email body, fear is the primary affective state in single-label cases, while urgency takes precedence in multi-label instances, followed by trust and fear. Paired affective states often differ between subject and body, with fear-urgency, urgency-trust, and trust-curiosity being among the top four pairings. Also, urgency and fear evoke the highest arousal levels in both subjects and bodies. Language models LLMs show fair agreement with human annotators, with GPT-4 slightly outperforming Gemini Pro and Llama 2 in performance. Thus, employing email labels that specifically highlight affective states and their valence and arousal levels, especially from new email senders, users can receive alerts prompting them to exercise extra caution when engaging with such emails. This heightened awareness can improve protection against phishing emails.

In future work, we plan to expand the scope of this study to include a more comprehensive phishing email dataset targeting various organizations beyond the university context. This will enable us to better understand how affective states manifest in phishing attacks across different sectors. Additionally, we will assess whether identifying affective state exploits in phishing emails can enhance the performance of automated phishing detection methods.

CRediT authorship contribution statement

Faithful Chiagoziem Onwuegbuche: Conceptualization, Data curation, Funding acquisition, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Validation, Writing – original draft, Writing – review & editing, Project administration. **Rajesh Titung:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft. **Esa Rantanen:** Funding acquisition, Supervision, Writing – review & editing. **Anca Delia Jurcut:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Cecilia O. Alm:** Funding acquisition, Supervision, Writing – review & editing. **Liliana Pasquale:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183).

This material is based upon work supported by the U.S. National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

AlEroud, A., Karabatis, G., 2020. Bypassing detection of url-based phishing attacks using generative adversarial deep neural networks, in: Proceedings of the sixth international workshop on security and privacy analytics, pp. 53–60.

Almomani, A., Gupta, B.B., Atawneh, S., Meulenberg, A., Almomani, E., 2013. A Survey of Phishing Email Filtering Techniques. IEEE communications surveys & tutorials 15, 2070–2090.

APWG, 2024. Phishing activity trends report, 2023. https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf.

Arduin, P.E., 2023. A cognitive approach to the decision to trust or distrust phishing emails. International Transactions in Operational Research 30, 1263–1298.

Atari, M., Al-Mousa, A., 2022. A machine-learning based approach for detecting phishing urls, in: 2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), IEEE. pp. 82–88.

Baki, S., Verma, R.M., 2022. Sixteen years of phishing user studies: What have we learned? IEEE Transactions on Dependable and Secure Computing 20, 1200–1212.

Beesdo, K., Knappe, S., Pine, D.S., 2009. Anxiety and anxiety disorders in children and adolescents: developmental issues and implications for dsm-v. Psychiatric Clinics 32, 483–524.

Benenson, Z., Gassmann, F., Landwirth, R., 2017. Unpacking spear phishing susceptibility, in: Financial Cryptography and Data Security: FC 2017 International Workshops, WAHC, BITCOIN, VOTING, WTSC, and TA, Sliema, Malta, April 7, 2017, Revised Selected Papers 21, Springer. pp. 610–627.

von Bieberstein, F., Essl, A., Friedrich, K., 2021. Empathy: A clue for prosocialty and driver of indirect reciprocity. Plos one 16, e0255071.

Bitaab, M., Cho, H., Oest, A., Zhang, P., Sun, Z., Pourmohamad, R., Kim, D., Bao, T., Wang, R., Shoshitaishvili, Y., et al., 2020. Scam pandemic: How attackers exploit public fear through phishing, in: 2020 APWG Symposium on Electronic Crime Research (eCrime), IEEE. pp. 1–10.

Braca, A., Dondio, P., 2023. Persuasive communication systems: a machine learning approach to predict the effect of linguistic styles and persuasion techniques. Journal of Systems and Information Technology .

Bradley, M.M., Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry 25, 49–59.

Bright, C., Wziatka, M., Ngaruko, W., 2022. An examination of the role of big five personality traits, cognitive processes and heuristics on individuals' phishing attack susceptibility levels .

Burda, P., Chotza, T., Allodi, L., Zannone, N., 2020. Testing the effectiveness of tailored phishing techniques in industry and academia: a field experiment, in: Proceedings of the 15th International Conference on Availability, Reliability and Security, pp. 1–10.

Butavicius, M., Parsons, K., Pattinson, M., McCormac, A., 2016. Breaching the human firewall: Social engineering in phishing and spear-phishing emails. arXiv preprint arXiv:1606.00887 .

Button, M., Nicholls, C.M., Kerr, J., Owen, R., 2014. Online Frauds: Learning from Victims Why They Fall for These Scams. Australian & New Zealand journal of criminology 47, 391–408.

Canfield, C.I., Fischhoff, B., Davis, A., 2019. Better beware: comparing metacognition for phishing and legitimate emails. Metacognition and learning 14, 343–362.

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.

Chae, Y., Davidson, T., 2023. Large language models for text classification: From zero-shot learning to fine-tuning. Open Science Foundation .

Chatterjee, A., Basu, S., 2021. How vulnerable are you? a novel computational psycholinguistic analysis for phishing influence detection, in: Proceedings of the 18th International Conference on Natural Language Processing (ICON), pp. 499–507.

Cialdini, R.B., 1993. Influence: The Psychology of Persuasion. Rev. ed., Morrow, New York.

Ciambrone, G., Wilson, S., 2023. Creation and analysis of a corpus of scam emails targeting universities, in: Companion Proceedings of the ACM Web Conference 2023, Association for Computing Machinery, New York, NY, USA. p. 24–27.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20, 37–46.

Danet, D., 2021. Punish and perish: The human factor in cybersecurity., in: ITASEC, pp. 1–8.

Davies, V., 2023. 82% of all cyberattacks involve the human element. https://cybermagazine.com/articles/82-of-all-cyberattacks-involve-the-human-element.

Deloitte, 2020. 91% of all cyber attacks begin with a phishing email to an unexpected victim. https://www2.deloitte.com/my/en/pages/risk/articles/91-percent-of-all-cyber-attacks-begin-with-a-phishing-email-to-an-unexpected-victim.html.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S., 2020. Goemotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547 .

Digman, J.M., 2004. Personality structure: Emergence of the five-factor model. The psychology of individual differences , 71–93.

Eftimie, S., Moinescu, R., Răcuciu, C., 2022. Spear-phishing susceptibility stemming from personality traits. IEEE Access 10, 73548–73561.

Egress, 2023. Phishing threat trends report. https://www.egress.com/media/mq4kwitu/egress_phishing_threat_trends_report.pdf.

Fan, Z., Li, W., Laskey, K.B., Chang, K.C., 2024. Investigation of phishing susceptibility with explainable artificial intelligence. Future Internet 16, 31.

Ferreira, A., Coventry, L., Lenzini, G., 2015. Principles of persuasion in social engineering and their use in phishing, in: Human Aspects of Information Security, Privacy, and Trust: Third International Conference, HAS 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015. Proceedings 3, Springer. pp. 36–47.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychological bulletin 76, 378.

Frauenstein, E.D., Flowerday, S., 2020. Susceptibility to phishing on social network sites: A personality information processing model. Computers & security 94, 101862.

Ge, Y., Lu, L., Cui, X., Chen, Z., Qu, W., 2021. How personal characteristics impact phishing susceptibility: The mediating role of mail processing. Applied Ergonomics 97, 103526.

Gemini, 2023. Gemini: A family of highly capable multimodal models.

Ghazi-Tehrani, A.K., Pontell, H.N., 2022. Phishing evolves: Analyzing the enduring cybercrime, in: The New Technology of Financial Crime. Routledge, pp. 35–61.

Giboney, J.S., Schuetzler, R.M., Grimes, G.M., 2023. Know your enemy: Conversational agents for security, education, training, and awareness at scale. Computers & Security 129, 103207.

Gopavaram, S., Dev, J., Grobler, M., Kim, D., Das, S., Camp, L.J., 2021. Cross-national study on phishing resilience, in: Proceedings of the Workshop on Usable Security and Privacy (USEC).

Gragg, D., 2003. A multi-level defense against social engineering. SANS Reading Room 13, 1–21.

Greitzer, F.L., Li, W., Laskey, K.B., Lee, J., Purl, J., 2021. Experimental investigation of technical and human factors related to phishing susceptibility. ACM Transactions on Social Computing 4, 1–48.

Gross, C.T., Canteras, N.S., 2012. The many paths to fear. Nature Reviews Neuroscience 13, 651–658.

Gruber, M.J., Gelman, B.D., Ranganath, C., 2014. States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. Neuron 84, 486–496.

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 .

Jampen, D., Gür, G., Sutter, T., Tellenbach, B., 2020. Don't click: towards an effective anti-phishing training. a comparative literature review. Human-centric Computing and Information Sciences 10, 1–41.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P., 2023. Survey of hallucination in natural language generation. ACM Computing Surveys 55, 1–38.

Kang, M.J., Hsu, M., Krajbich, I.M., Loewenstein, G., McClure, S.M., Wang, J.T.y., Camerer, C.F., 2009. The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. Psychological science 20, 963–973.

Kashapov, A., Wu, T., Abuadbba, S., Rudolph, C., 2022. Email summarization to assist users in phishing identification, in: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security, pp. 1234–1236.

Kidd, C., Hayden, B.Y., 2015. The psychology and neuroscience of curiosity. Neuron 88, 449–460.

Kleitman, S., Law, M.K., Kay, J., 2018. It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. PloS one 13, e0205089.

Krippendorff, K., 2011. Computing krippendorff's alpha-reliability .

Lambie, G.W., Haugen, J.S., 2019. Understanding greed as a unified construct. Personality and Individual Differences 141, 31–39.

Lawson, P., Pearson, C.J., Crowson, A., Mayhorn, C.B., 2020. Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy. Applied ergonomics 86, 103084.

Li, F., Betts, S.C., et al., 2003. Trust: What it is and what it is not. International Business & Economics Research Journal (IBER) 2.

Li, W., Lee, J., Purl, J., Greitzer, F., Yousefi, B., Laskey, K., 2020. Experimental investigation of demographic factors related to phishing susceptibility .

Lin, T., Capecci, D.E., Ellis, D.M., Rocha, H.A., Dommaraju, S., Oliveira, D.S., Ebner, N.C., 2019. Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. ACM Transactions on Computer-Human Interaction (TOCHI) 26, 1–28.

Mansfield-Devine, S., 2017. Bad behaviour: exploiting human weaknesses. Computer Fraud & Security 2017, 17–20.

Marchal, M., Scholman, M., Yung, F., Demberg, V., 2022. Establishing annotation quality in multi-label annotations, in: Proceedings of the 29th International Conference on Computational Linguistics, pp. 3659–3668.

Mitnick, K.D., Simon, W.L., 2003. The art of deception: Controlling the human element of security. John Wiley & Sons.

Mohebzada, J.G., El Zarka, A., BHojani, A.H., Darwish, A., 2012. Phishing in a university community: Two large scale phishing experiments, in: 2012 international conference on innovations in information technology (IIT), IEEE. pp. 249–254.

Moody, G.D., Galletta, D.F., Dunn, B.K., 2017. Which phish get caught? an exploratory study of individuals' susceptibility to phishing. European Journal of Information Systems 26, 564–584.

Muncaster, P., 2021. What is phishing and how do you prevent phishing attacks? https://www.verizon.com/business/resources/articles/s/what-is-phishing-and-how-do-you-prevent-phishing-attacks/.

Muralidharan, T., Nissim, N., 2023. Improving malicious email detection through novel designated deep-learning architectures utilizing entire email. Neural Networks 157, 257–279.

Musuva, P.M., Getao, K.W., Chepken, C.K., 2019. A new approach to modelling the effects of cognitive processing and threat detection on phishing susceptibility. Computers in Human Behavior 94, 154–175.

Neupane, A., Rahman, M.L., Saxena, N., Hirshfield, L., 2015. A multimodal neuro-physiological study of phishing detection and malware warnings, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 479–491.

Okey, O.D., Udo, E.U., Rosa, R.L., Rodríguez, D.Z., Kleinschmidt, J.H., 2023. Investigating chatgpt and cybersecurity: A perspective on topic modeling and sentiment analysis. Computers & Security 135, 103476.

OpenAI, 2023. Gpt-4 technical report.

Packetlabs, 2023. WormGPT and PoisonGPT: Beware Malicious Generative AI Models. https://www.packetlabs.net/posts/wormgpt-and-poisongpt/.

Parrish Jr, J.L., Bailey, J.L., Courtney, J.F., 2009. A personality based model for determining susceptibility to phishing attacks. Little Rock: University of Arkansas , 285–296.

Perez-Bret, E., Altisent, R., Rocafort, J., 2016. Definition of compassion in healthcare: a systematic literature review. International journal of palliative nursing 22, 599–606.

Picard, R.W., 2000. Affective computing. MIT press.

Proofpoint, 2021. 2022 State of Phish. https://go.proofpoint.com/en-2022-state-of-the-phish.html.

Ribeiro, L., Guedes, I.S., Cardoso, C.S., 2024. Which factors predict susceptibility to phishing? an empirical study. Computers & Security 136, 103558.

Rocha Flores, W., Holm, H., Nohlberg, M., Ekstedt, M., 2015. Investigating personal determinants of phishing and the effect of national culture. Information & Computer Security 23, 178–199.

Salloum, S., Gaber, T., Vadera, S., Shaalan, K., 2021. Phishing email detection using natural language processing techniques: a literature survey. Procedia Computer Science 189, 19–28.

Sarno, D.M., Harris, M.W., Black, J., 2023. Which phish is captured in the net? understanding phishing susceptibility and individual differences. Applied Cognitive Psychology .

Schneier, B., 2000. Inside risks: semantic network attacks. Communications of the ACM 43, 168.

Schneier, B., 2023. A Hacker's Mind: How the Powerful Bend Society's Rules, and how to Bend Them Back. WW Norton & Company.

Seuntjens, T.G., Zeelenberg, M., Breugelmans, S.M., Van de Ven, N., 2015. Defining greed. British Journal of Psychology 106, 505–525.

Shahriar, S., Mukherjee, A., Gnawali, O., 2022. Improving phishing detection via psychological trait scoring, in: Proceedings of the IADIS International Conference Web Based Communities 2022 (part of MCCSIS 2022), pp. 131–139.

Sharma, M., Kumar, M., Gonzalez, C., Dutt, V., 2022. How the presence of cognitive biases in phishing emails affects human decision-making?, in: International Conference on Neural Information Processing, Springer. pp. 550–560.

Sharma, T., Bashir, M., 2020. An analysis of phishing emails and how the human vulnerabilities are exploited, in: Advances in Human Factors in Cybersecurity: AHFE 2020 Virtual Conference on Human Factors in Cybersecurity, July 16–20, 2020, USA, Springer. pp. 49–55.

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L.F., Downs, J., 2010. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions, in: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 373–382.

Simoiu, C., Zand, A., Thomas, K., Bursztein, E., 2020. Who is targeted by email-based phishing and malware? measuring factors that differentiate risk, in: Proceedings of the ACM Internet Measurement Conference, pp. 567–576.

Stajano, F., Wilson, P., 2011. Understanding scam victims: seven principles for systems security. Communications of the ACM 54, 70–75.

Strauss, C., Taylor, B.L., Gu, J., Kuyken, W., Baer, R., Jones, F., Cavanagh, K., 2016. What is compassion and how can we measure it? a review of

definitions and measures. Clinical psychology review 47, 15–27.
Taib, R., Yu, K., Berkovsky, S., Wiggins, M., Bayl-Smith, P., 2019. Social engineering and organisational dependencies in phishing attacks, in: IFIP Conference on Human-Computer Interaction, Springer. pp. 564–584.
Tekumalla, R., Banda, J.M., 2023. Leveraging large language models and weak supervision for social media data annotation: An evaluation using covid-19 self-reported vaccination tweets, in: Mori, H., Asahi, Y., Coman, A., Vasilache, S., Rauterberg, M. (Eds.), HCI International 2023 – Late Breaking Papers, Springer Nature Switzerland, Cham. pp. 356–366.
Thomas, J., 2018. Individual cyber security: Empowering employees to resist spear phishing to prevent identity theft and ransomware attacks. Thomas, JE (2018). Individual cyber security: Empowering employees to resist spear phishing to prevent identity theft and ransomware attacks. International Journal of Business Management 12, 1–23.
Tornblad, M.K., Jones, K.S., Namin, A.S., Choi, J., 2021. Characteristics that predict phishing susceptibility: A review, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications Sage CA: Los Angeles, CA. pp. 938–942.
Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Scialom, T., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.
Turnage, A.K., 2007. Email flaming behaviors and organizational conflict. Journal of Computer-Mediated Communication 13, 43–59.
Valimail, 2019. More than 3 billion fake emails are sent worldwide every day, valimail report finds post date. https://www.valimail.com/newsroom/more-than-3-billion-fake-emails-are-sent-worldwide-every-day-valimail-report-finds/.
Van Der Heijden, A., Allodi, L., 2019. Cognitive triaging of phishing attacks, in: 28th USENIX Security Symposium (USENIX Security 19), pp. 1309–1326.
Vishwanath, A., 2015. Examining the distinct antecedents of e-mail habits and its influence on the outcomes of a phishing attack. Journal of Computer-Mediated Communication 20, 570–584.
Vishwanath, A., Harrison, B., Ng, Y., 2015. Suspicion, cognition, automaticity model (scam) of phishing susceptibility, in: Proceedings of the Annual Meeting of 65th International Communication Association Conference, San Juan.
Williams, E.J., Hinds, J., Joinson, A.N., 2018. Exploring susceptibility to phishing in the workplace. International Journal of Human-Computer Studies 120, 1–13.
Wright, R.T., Jensen, M.L., Thatcher, J.B., Dinger, M., Marett, K., 2014. Research note—influence techniques in phishing attacks: an examination of vulnerability and resistance. Information systems research 25, 385–400.
Yang, R., Zheng, K., Wu, B., Li, D., Wang, Z., Wang, X., et al., 2022. Predicting user susceptibility to phishing based on multidimensional features. Computational Intelligence and Neuroscience 2022.
Zeng, V., Baki, S., Aassal, A.E., Verma, R., De Moraes, L.F.T., Das, A., 2020. Diverse datasets and a customizable benchmarking framework for phishing, in: Proceedings of the Sixth International Workshop on Security and Privacy Analytics, pp. 35–41.
Zhang, X., Talukdar, N., Vemulapalli, S., Ahn, S., Wang, J., Meng, H., Murtaza, S.M.B., Leshchiner, D., Dave, A.A., Joseph, D.F., et al., 2024. Comparison of prompt engineering and fine-tuning strategies in large language models in the classification of clinical notes. medRxiv , 2024–02.
Zhu, M., Yang, Y., Hsee, C.K., 2018. The mere urgency effect. Journal of Consumer Research 45, 673–690.
Zhuo, S., Biddle, R., Koh, Y.S., Lottridge, D., Russello, G., 2023. Sok: Human-centered phishing susceptibility. ACM Transactions on Privacy and Security 26, 1–27.

A. Majority consensus algorithm

Algorithm 1 is the majority consensus algorithm as used in this paper.

Data: row: a data row with columns GPT-4, Llama2, GeminiPro
Result: List of labels representing the majority decision

```
begin
    labels ← empty list;
    foreach column in [GPT-4, Llama2, GeminiPro] do
        for label in column.split(',') do
            if label is not empty then
                Add label to labels list;
            end
        end
    end
    label_counts ← Counter(labels);
    if label_counts is empty then
        return ['NA'];
    end
    max_count ← max(label_counts.values());
    majority_labels ← [label for label, count in
     label_counts.items() if count >= 2];
    if majority_labels is empty then
        return "Disagreement";
    end
    return ",".join(a for a in majority_labels);
end
```

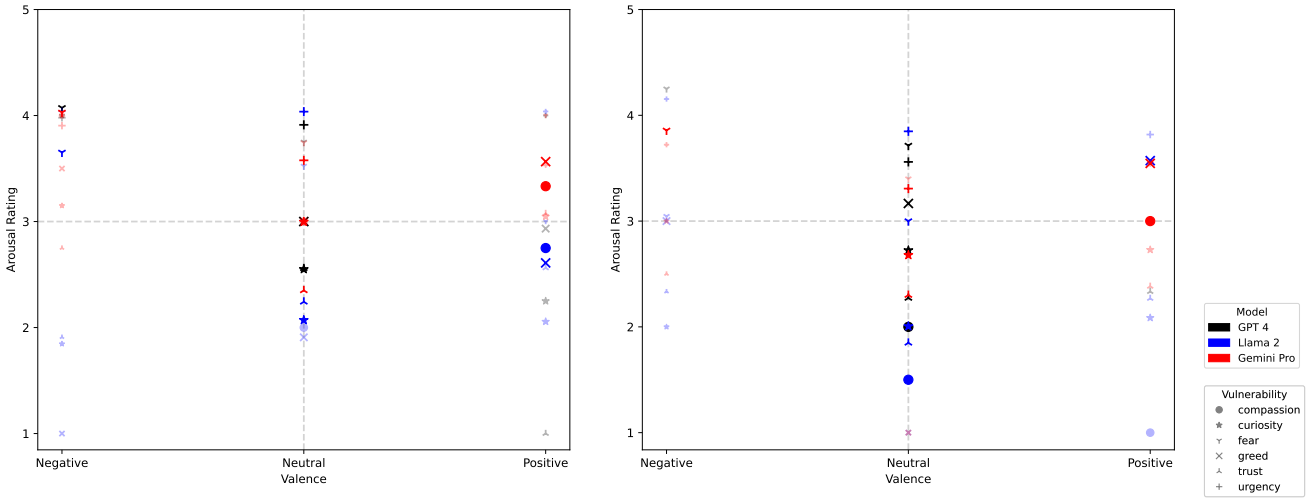
Algorithm 1: Majority decision

B. Valence-arousal space

Figure 10(a) displays the Valence-Arousal Cartesian chart for Email Subjects, while Figure 10(b) illustrates the same for Email Bodies across different LLMs. Table 12 provides the observations of valence-arousal Cartesian coordinates for email subjects across various LLMs and their consensus. Likewise, Table 13 outlines the valence-arousal Cartesian coordinates for email bodies.

C. Literature summary

Table 14 shows a summary of the literature reviewed in this study.



(a) Valence-arousal cartesian for email subjects

(b) Valence-arousal cartesian for email bodies

Figure 10: Valence-arousal cartesian for email (a) subjects and (b) bodies for various LLMs

Model	Affective State	Frequency Order	Valence	Arousal Rating
Majority Consensus	Compassion	2	0	2.50
		3	1	3.13
	Curiosity	1	-1	2.40
		3	0	2.55
		2	1	2.53
	Fear	3	-1	3.87
		2	0	3.59
	Greed	1	-1	2.50
		2	0	2.54
		3	1	3.10
	Trust	1	-1	2.20
		3	0	2.36
		2	1	2.82
	Urgency	2	-1	3.97
		3	0	3.84
		1	1	3.82
GPT-4	Curiosity	1	-1	4.00
		3	0	2.55
		2	1	2.25
	Fear	3	-1	4.07
		2	0	3.75
	Greed	3	0	3.00
		2	1	2.93
	Trust	3	0	2.56
		2	1	1.00
	Urgency	2	-1	3.98
		3	0	3.91
		1	1	4.00
Llama 2	Compassion	2	0	2.00
		3	1	2.75
	Curiosity	1	-1	1.85
		3	0	2.07
		2	1	2.06
	Fear	3	-1	3.65
		2	0	3.52
		1	1	3.00
	Greed	1	-1	1.00
		2	0	1.91
		3	1	2.61
	Trust	1	-1	1.91
		3	0	2.24
		2	1	2.56
	Urgency	2	-1	4.04
		3	0	4.04
		1	1	4.04
Gemini Pro	Compassion	1	-1	4.00
		2	0	3.00
		3	1	3.33
	Curiosity	1	-1	3.15
		3	0	3.00
		2	1	3.05
	Fear	3	-1	4.03
		2	0	3.74
		1	1	4.00
	Greed	1	-1	3.50
		2	0	3.00
		3	1	3.57
	Trust	1	-1	2.75
		3	0	2.35
		2	1	3.08
	Urgency	2	-1	3.90
		3	0	3.58
		1	1	3.53

Table 12
Average valence-arousal rating for various LLMs in email subject

Model	Affective State	Frequency Order	Valence	Arousal Rating
Majority Consensus	Compassion	3	0	2.00
		2	1	1.50
	Curiosity	1	-1	2.53
		3	0	2.43
		2	1	2.64
	Fear	2	-1	3.60
		3	0	3.30
	Greed	1	-1	2.00
		1	0	1.00
		3	1	3.41
	Trust	1	-1	2.00
		3	0	2.17
		2	1	2.29
	Urgency	1	-1	3.94
		3	0	3.57
		2	1	3.48
GPT-4	Compassion	3	0	2.00
	Curiosity	3	0	2.72
		2	1	3.00
	Fear	2	-1	4.25
		3	0	3.71
	Greed	3	0	3.17
	Trust	3	0	2.27
		2	1	2.33
	Urgency	3	0	3.56
Llama 2	Compassion	3	0	1.50
		2	1	1.00
	Curiosity	1	-1	2.00
		3	0	2.01
		2	1	2.09
	Fear	2	-1	3.05
		3	0	3.00
	Greed	2	-1	3.00
		1	0	1.00
		3	1	3.57
	Trust	1	-1	2.33
		3	0	1.85
		2	1	2.27
	Urgency	1	-1	4.15
		3	0	3.85
		2	1	3.82
Gemini Pro	Compassion	3	1	3.00
	Curiosity	1	-1	3.00
		3	0	2.68
		2	1	2.73
	Fear	3	-1	3.86
		2	0	3.40
	Greed	1	0	1.00
		3	1	3.55
	Trust	1	-1	2.50
		3	0	2.30
		2	1	2.38
	Urgency	1	-1	3.72
		3	0	3.31
		2	1	3.55

Table 13
Average valence-arousal rating for various LLMs in the email body

Author	Year	affective states						Theories of Scams and affective states					Method				Detection			
		Fear	Curiosity	Trust	Empathy	Urgency	Greed	Cialdini	Stajano	Gragg	Ferreira	Schneier	NLP	LLM	ML	Others	Binary	Multi-class	Sentiment	Others
Shahriar et al. (2022)	2022	✓	×	×	×	✓	✓	×	×	×	×	×	✓	✓	✓	×	✓	×	×	×
Zhuo et al. (2023)	2023	✓	×	×	×	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×	×	×	×
Chatterjee and Basu (2021)	2021	×	×	×	×	×	×	✓	×	×	×	×	✓	✓	✓	×	×	✓	×	×
Sharma et al. (2022)	2023	×	×	✓	×	×	✓	×	×	×	×	×	✓	×	✓	×	×	✓	×	×
Sharma and Bashir (2020)	2020	✓	✓	✓	×	×	×	×	×	×	×	×	✓	×	×	×	×	✓	✓	×
Kashapov et al. (2022)	2022	×	×	×	×	×	×	✓	×	×	×	×	✓	✓	×	×	✓	✓	×	×
Shahriar et al. (2022)	2022	✓	×	×	×	✓	✓	×	×	×	×	×	✓	✓	✓	×	✓	×	×	×
Muralidharan and Nissim (2023)	2023	×	×	×	×	×	×	×	×	×	×	×	✓	×	✓	✓	✓	×	×	✓
Carroll et al. (2022)	2022	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Braca and Dondio (2023)	2023	×	×	✓	×	×	✓	✓	×	×	×	×	×	×	✓	×	×	✓	×	×
Bright et al. (2022)	2022	×	×	✓	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Salloum et al. (2021)	2021	×	×	×	×	×	×	×	×	×	×	×	✓	×	✓	×	✓	×	×	×
Arduin (2023)	2023	×	×	✓	×	✓	×	✓	×	×	×	×	×	×	×	×	×	×	×	×
AlEroud and Karabatis (2020)	2020	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	✓	×	×	×
Van Der Heijden and Allodi (2019)	2019	×	×	×	×	×	×	✓	×	×	×	×	✓	×	✓	✓	✓	✓	×	×
Musuva et al. (2019)	2019	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	✓
Kleitman, Law and Kay (2018)	2018	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	✓	×	×	×
Thomas (2018)	2018	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Ferreira et al. (2015)	2015	×	×	×	×	×	×	✓	✓	✓	-	×	×	×	×	×	×	×	×	×
Butavicius, Parsons, Pattinson and McCormac (2016)	2016	×	×	×	×	×	×	✓	×	×	×	×	✓	×	×	✓	✓	✓	×	×
Neupane et al. (2015)	2015	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	✓
Wright et al. (2014)	2014	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	✓	×	×	×	✓
Giboney, Schuetzler and Grimes (2023)	2023	✓	×	✓	×	✓	×	✓	×	×	×	×	×	×	×	✓	×	×	×	✓
Canfield et al. (2019)	2019	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
Zeng, Baki, Aassal, Verma, De Moraes and Das (2020)	2020	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	✓	×	×	×
Simoiu et al. (2020)	2020	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
Tornblad et al. (2021)	2021	×	✓	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Jampen et al. (2020)	2022	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Yang, Zheng, Wu, Li, Wang, Wang et al. (2022)	2022	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	✓	×	×	×
Eftimie et al. (2022)	2022	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	✓	✓	×	×	✓
This Work	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓

Table 14: Summary of Literature

Faithful Chiagoziem Onwuegbuche is a PhD candidate in Machine Learning at the SFI Center for Research Training in Machine Learning (ML-Labs) at University College Dublin (UCD), Ireland. He is also affiliated with the SFI Research Centre for Software (LERO). His research interests lie at the intersection of artificial intelligence and blockchain technology, with a focus on their applications in finance and cybersecurity. His ultimate goal is to develop AI systems capable of securing critical infrastructure, detecting financial fraud, and promoting financial inclusion. Prior to his PhD, he earned two master's degrees: a Master of Science in Financial Technology (FinTech) with Distinction from the University of Stirling, UK, and a Master of Science in Financial Mathematics from the Pan African University, Kenya, as a Commonwealth Shared Scholar and African Union Scholar, respectively. He has professional experience as a data scientist, machine learning researcher, and lecturer. You can read more about him at: https://faithfulco.github.io/.

Rajesh Titung is a Computing and Information Sciences PhD student in the Computational Linguistics and Speech Processing Lab. His research interests are in natural language processing and multimodality, human-in-the-loop machine learning, and affective computing. His work focuses on interactive machine learning and federated learning for multimodal personalized affective computing. He is an AWARE-AI NSF Research Traineeship trainee. He has worked previously as a machine learning engineer.

Dr. Esa M. Rantanen trained as a commercial pilot. He also has seven years of experience as an air traffic controller and an air traffic control instructor. Dr. Rantanen has a Bachelor of Science and a Master of Aeronautical Science degree from Embry-Riddle Aeronautical University, Daytona Beach, Florida. He also has a Master of Science in Industrial Engineering degree from the Pennsylvania State University, with specialization in human factors/ergonomics engineering. His Ph.D degree is from Penn State as well, in Engineering Psychology. Dr. Rantanen has served as an assistant professor at the Institute of Aviation of the University of Illinois at Urbana-Champaign. Presently Dr. Rantanen is an associate professor of psychology at RIT. He is primarily teaching courses in the MS in Experimental Psychology program and supervising graduate students' thesis research in the Engineering Psychology track of the program. He also has an Extended Faculty Appointment in the Department of Industrial and Systems Engineering and the Engineering Ph.D. program in the Kate Gleason College of Engineering, and he is an Affiliate of the Global Cybersecurity Institute and Center for Human-Aware AI at RIT. Dr. Rantanen's research interests lie in the areas of human factors in complex systems, human performance measurement and modeling, mental workload, decision making, and human error and reliability.

Dr. Anca Jurcut is an Assistant Professor in the School of Computer Science, University College Dublin (UCD), Ireland, since 2015. She received a BSc in Computer Science and Mathematics from West University of Timisoara, Romania in2007 and a PhD in Security Engineering from the University of Limerick (UL), Ireland in 2013 funded by the Irish Research Council for Science Engineering and Technology. She worked as a postdoctoral researcher at UL as a member of the Data Communication Security Laboratory and as a Software Engineer in IBM in Dublin, Ireland in the area of data security and formal verification. Dr. Jurcut research interests include Security Protocols Design and Analysis, Automated Techniques for Formal Verification, Network Security, Attack Detection and Prevention Techniques, Security for the Internet of Things, and Applications of Blockchain for Security and Privacy. Dr. Jurcut has several key contributions in research focusing on detection and prevention techniques of attacks over networks, the design and analysis of security protocols, automated techniques for formal verification, and security for mobile edge computing (MEC). More Info: https://people.ucd.ie/anca.jurcut.

Cecilia O. Alm (Ph.D., UIUC) is a professor at the Rochester Institute of Technology (RIT). In addition to the Department of Psychology and the School of Information, she is affiliated with RIT's Cognitive Science and Computing and Information Sciences PhD programs. At RIT, she is the joint program director for the MS program in Artificial Intelligence, directs the Computational Linguistics and Speech Processing Lab, and serves as an associate director for the Center for Human-aware AI. In addition, she directs an NSF-funded research traineeship program for graduate students and has led two iterations of an NSF REU site. Her research interests center on responsibly developing human-inspired and human-centered AI systems, AI systems offering interaction insights, and also on preparing the AI research workforce. She teaches natural language processing, speech processing, and artificial intelligence coursework.

Liliana Pasquale received her PhD from Politecnico di Milano (Italy) in 2011. She is an Associate Professor at University College Dublin (Ireland) and a funded investigator at Lero - the SFI Research Centre for Software. Her research interests include requirements engineering and adaptive systems, focusing on security, privacy, and digital forensics. She has served in the Program and Organizing Committee of prestigious software engineering conferences, such as ICSE, FSE, ASE, RE. She is an associate editor of the IEEE TSE journal, department editor of the IEEE Security & Privacy Magazine and a member of the review board of the ACM TOSEM journal.