Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Investigating the Security Risks of Gen-AI-Powered Phishing Attacks on University Students

Islam Ahmed Samun[1] and Md Fahim Ahammed[2]

[1] Division Cloud and Network Security, School of Creative Technologies, University of Greater Manchester, Bolton, Manchester, United Kingdom
[2] Department of Information Assurance and Cybersecurity, Gannon University, Erie, Pennsylvania, USA
These authors contributed equally to this work
Corresponding Author: Md Fahim Ahammed, **Email:** mdfahimahammed7773@gmail.com

| ABSTRACT

Generative Artificial Intelligence (GenAI) is a paradigm shift in the cyber-threat environment as it allows generating hyper-realistic, personalised and scalable phishing campaigns. This is a study that explores how this technological development has intersected with the human susceptibility that has been there in the realm of higher education. This paper, which targets university students as a target population with a high level of digital exposure but possibly low levels of security awareness, also adopts a positivist, quantitative approach in its attempt to investigate empirically the security risks that AI-driven phishing presents to them. It was a cross-sectional online survey (N=63), in which a simulated GenAI phishing mail was used as a behavioural stimulus. Descriptive and inferential statistics demonstrate a crucial lack of connection: self-reported confidence in detection was at a moderate level (M=3.13/5), whilst behavioural susceptibility to it is alarmingly high, with 81.7 percent stating that they were likely to choose to click the malicious link. It was discovered that there was a deep institutional training deficit, as 90.1% of respondents had not been trained in cybersecurity at university and analyses (kh2(1)) demonstrated that prior training had no significant protective effect on phishing experience (kh2(1) =0.948, p=.330). Moreover, existing guidance in universities was perceived as insufficient by 76.1% of them. The results highlight a severe overconfidence paradox and a systemic defect in the modern pedagogical models to help counteract AI-improved threats. This paper has determined that the human firewall within the academic context is highly misaligned and recommends an immediate, strategic move to compulsory, simulation-based training programmes that are specifically crafted to take into account the advanced affordances of GenAI in social engineering attacks.

| KEYWORDS

AI-Powered Phishing, Generative AI, Cybersecurity Awareness, University Students, Human Factor, Susceptibility, Cybersecurity Training

| ARTICLE INFORMATION

ACCEPTED: 01 May 2026 **PUBLISHED:** 07 June 2026 **DOI:** 10.32996/fcsai.2026.5.8.5

CHAPTER 1 – RESEARCH PROBLEM & PHILOSOPHICAL POSITION

1.1 RESEARCH PROBLEM

The growing Generative Artificial Intelligence (GenAI) is an unprecedented growth in the cyber-threat environment, especially social engineering blockages such as phishing (Boddy, 2021). This development is the most important in the higher education sector since students, as the digital natives, tend to be overconfident about their cybersecurity skills even though they do not have the appropriate skills or institutional support (Hadlington, 2017). The preliminary results of the 63 participants of this

study indicate one of the most distressing demonstrations of this issue: out of 60.6% of participants have already experienced phishing, 90.1% of participants have not obtained any official cybersecurity education in college. This is an indication of a structural exposure in which a technologically progressive threat is addressed against an unready populace. The research problem in the form of the insufficient empirical knowledge about the particular vulnerability of the university students to GenAI-driven phishing or the proven inefficiency of the existing institutional protection develops an essential gap in the academic and operational context of cybersecurity policy.

1.2 PHILOSOPHICAL STANCE

This study is based on a positivist philosophy, which states that social reality exists as an objective, stable, and quantifiable and comprehensible in terms of observable facts (Saunders, Lewis and Thornhill, 2019).

Ontology (Objectivist): The study assumes that there are objective facts of GenAI phishing risks and student vulnerability. These effects include such phenomena as metrics such as click-through likelihood (81.7% were likely/very likely to click) and training deficits, which are externalized in individual interpretation, and measurable.

Epistemology (Positivist): It uses empirical observation and measurement in the production of knowledge. To identify factual relationships and verify hypotheses, the researcher takes a passive, objective approach to gather numerical data on variables like confidence scores, behavioural intentions, and demographic variables by minimising the influence of subjectivity on them.

Methodology (Quantitative/Deductive): In line with this position, a cross-sectional survey design was adopted and quantitative. In this way, statistical analysis of predefined variables is carried out to generate generalisable results concerning the causal relationships and patterns in the student population to the deductive research logic.

1.3 RESEARCH AIM AND OBJECTIVES

This study focuses on the critical examination of security risks of GenAI-enhanced phishing attacks on university students and the evaluation of the behavioural and contextual variables that contribute to their vulnerability.
The objectives of this study are as follows:

1) To detect student perceptions of email realism and student tendency to click.
2) To determine whether cybersecurity training or self-confidence can lessen the vulnerability.
3) To find behavioural or psychological predictors that augur the possibility of falling to AI-generated phishing.

1.4 RESEARCH QUESTIONS

RQ-1: What has the development of generative AI meant to the authenticity, prevalence, and nature of modern phishing attacks?

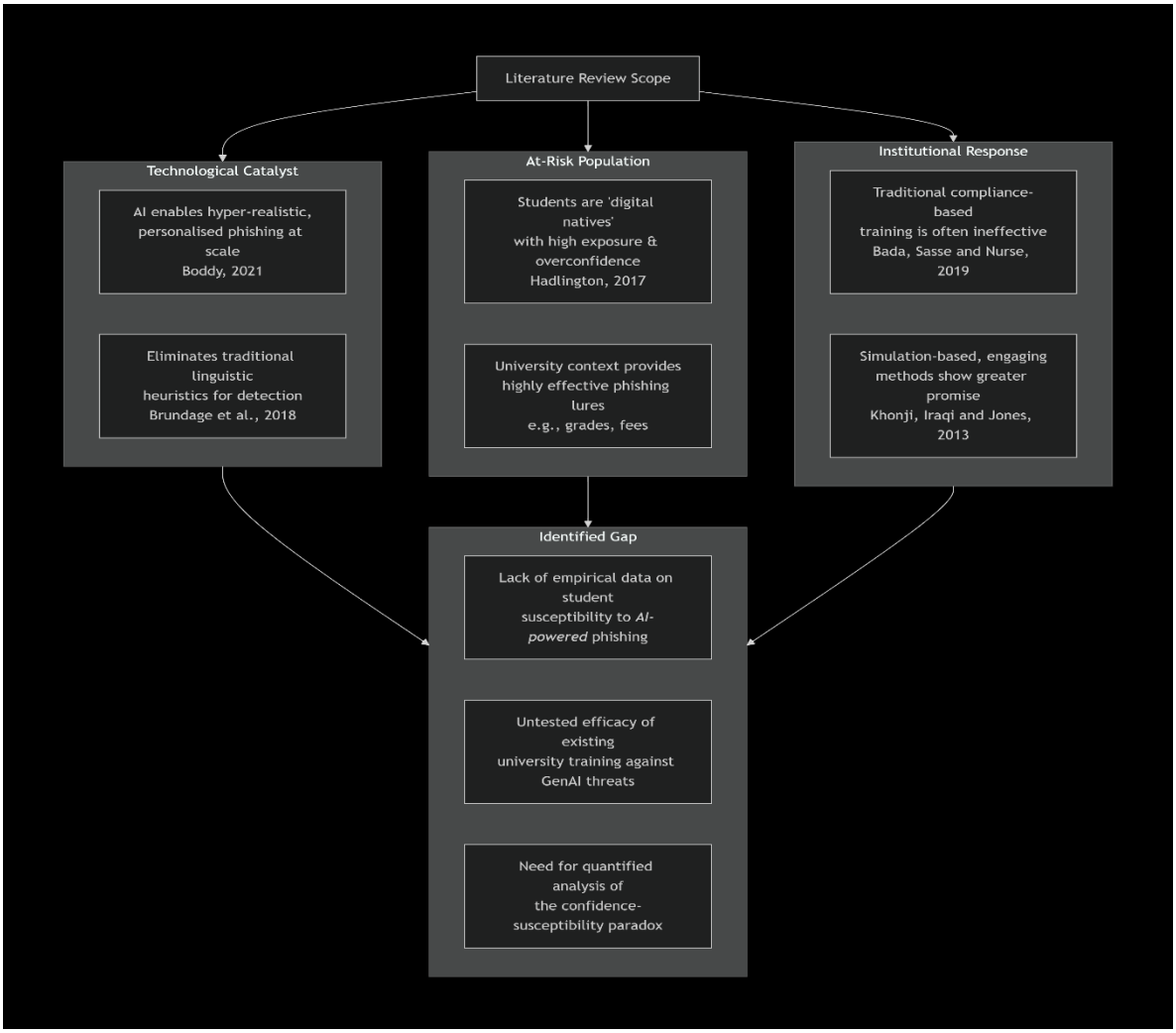
RQ-2: How well do university students show their awareness, confidence, and detection ability on AI-created phishing emails?

RQ-3: What is the relationship between student variables, including level of study, prior training on cybersecurity, and exposure to digital communication, and vulnerability to AI-based phishing?

RQ-4: How well do existing university cybersecurity guidelines, policies, and reporting systems respond to AI-enabled phishing threats?

CHAPTER 2 – ANNOTATED LITERATURE MAP & GAP STATEMENT

2.1 ANNOTATED LITERATURE MAP



The academic discourse on the subject matter can be generalised into four connected thematic clusters, as visualised above.

Cluster 1: The GenAI Revolution in Cybercrime: The current literature recognizes GenAI as an AI-based tool that can potentially transform the cybercrime context of the future. Large Language Models (LLMs) are capable of producing linguistically flawless, context-sensitive phishing text and eliminating the grammatical errors previously used as fundamental warning signs (Boddy, 2021). The technology allows not only to hyper-personalise spear-phishing but also to conduct it at a scale never seen before, which fundamentally changes the threat model of opportunistic scams to a persuasive one (Brundage et al., 2018).

Cluster 2: The Human Factor and Student Vulnerability: According to the consistent research, the main vulnerability in cybersecurity is the human factor (Jakobsson and Myers, 2006). A combination of elevated digital interaction rates, an academic culture of trust, and a general lack of awareness between how they think and feel they understand cybersecurity-relevant concepts is specifically identified to place students as a particularly vulnerable group (Hadlington, 2017; Williams, Beardmore and Joinson, 2018). In a different work of University of Wollongong, it is stated clearly that there is a lack of empirical research on AI-generated phishing. (Jabir, Le and Nguyen, 2025). A paper on the topic of the emails generated by AI being highly realistic and appearing as phishing. (Schmitt & Flechais, 2023). Further, the students have been subjected to critical attacks due to augmented digital activities. (Neil Hutchings & Helen Suh, 2021).

Cluster 3: Institutional Defences and Training Efficacy: The literature on Cybersecurity Awareness Training (CAT) programmes has criticised the traditional lectural format, implying that they do not necessarily produce a sustainable change in behaviour (Bada, Sasse and Nurse, 2019). There is evidence in support of the more interactive and experiential techniques, such as embedded phishing simulation, as a more effective teaching tool (Khonji, Iraqi and Jones, 2013). A study indicates that Institutions lack empirical data regarding the AI-phishing threats. (Jabir, Le & Nguyen, 2025). This leads to one of the most crucial questions concerning the applicability of pre-AI training programs to GenAI-enhanced threats.

2.2 CRITICAL SYNTHESIS & IDENTIFIED RESEARCH GAP

Although the literature carries out a strong description of the potential of GenAI, the fact that humans remain vulnerable, and there are generic principles of effective training, there is an immense gap in the empirical literature where they meet. There exists a dearth of studies that:

Quantitatively assesses the behavioural vulnerability to phishing baits which assume GenAI attributes (e.g. good linguistic quality, personalisation to the context) in a real student group.

Conducts a statistical analysis of the practical effectiveness of current training (or absence of training) on cybersecurity in higher education in mitigating this novel threat actor.

Empirically examines the so-called overconfidence paradox in which self-assurance prevails with high risk, in this case related to AI-powered attacks.

The preliminary data of this study highlights this gap, as 90.1 percent of the students were not trained and 81.7 percent reported that they fell prey to a simulated AI-phishing email, even with moderate self-reported confidence (Mean=3.13/5).

2.3 CONCEPTUAL FRAMEWORK

A Positivist-Behavioural Framework guides the study; according to which, the predisposition to AI-phishing is an established consequence depending on the measurement of antecedent factors.

Stimulus (GenAI Threat): Operationalised through the characteristics of the simulated phishing mail (perfect language, urgency, university-branding).

Organism (Student Factors): It will measure self-efficacy (confidence, knowledge), previous experience (phishing interactions, training), and demographics.

Response (Behavioural Susceptibility): The dependent, which is the likelihood reported by the participant to enable the phishing link to be clicked.

This framework will be testing the hypothesis that psycho-behavioural (such as confidence, training) are significantly disconnected to behavioural response (high click-likelihood), and the variables of institutional support will be significantly correlated with the perceptions of inadequacy.

CHAPTER 3: FEASIBILITY & ETHICS STATEMENT

3.1 PROJECT FEASIBILITY

This project was created with high feasibility on the academic constraints.

Resources: The digital survey platform has been used to collect data, which is cost-effective and covers a wide geographical area. The IBM SPSS was used to perform the statistical analysis, which is available under university licensing.

Data Access: The target population was not difficult to reach because it was found through university networks and online student forums. Although a convenience sample, the obtained sample (N=63) of various universities offered credible data to be preliminary.

Time frame: The cross-sectional approach allowed effective collection of data during a narrow time frame which fitted in the dissertation schedule. The data cleaning and analysis as well as reporting followed the normal timeline of a research module.

3.2 ETHICAL CONSIDERATIONS

No compromise in ethical rigour was allowed:

Informed Consent: The participation was voluntary. The nature of the study was described in an initial information sheet, which included the description of using a simulated phishing email.

Reducing harm: The fraud of the phishing simulation was minimised using a debriefing page that was mandatory and immediate. This was the aim of education, the email was recognized as a fake and participants were redirected to authentic university cybersecurity resources.

Anonymity and Confidentiality: No personally identifying data were obtained. All the answers were anonymised and the

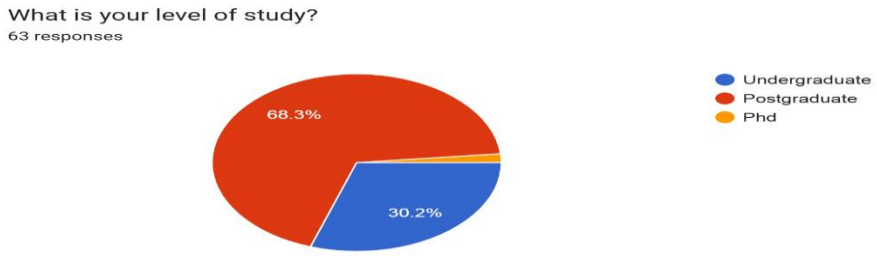
information was stored safely in a password-protected system and remained within the institutional data policy.
Consent: An ethical review was provided and received before the distribution (see Appendix).

CHAPTER 4: RESEARCH DESIGN & METHODS PLAN

4.1 RESEARCH DESIGN JUSTIFICATION

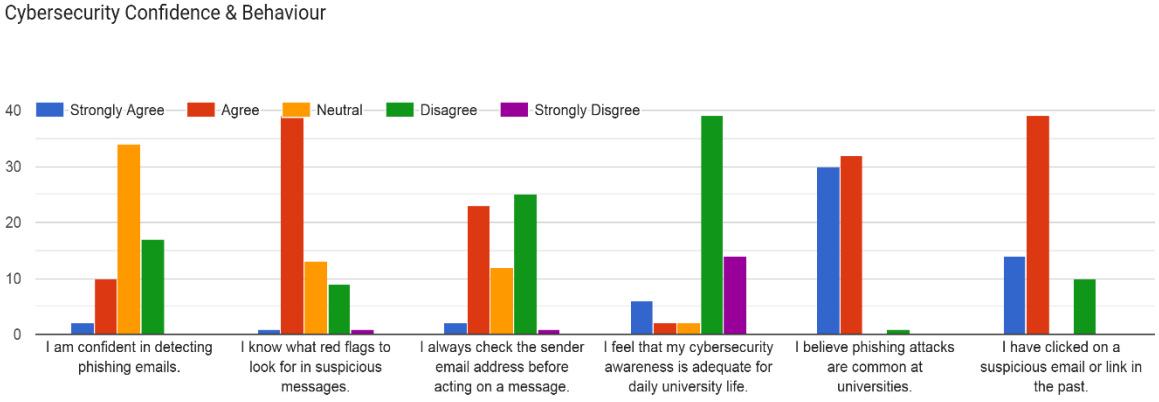
A cross-sectional survey design was used, which is quantitative, and it is in a suitable fit to the positivist paradigm. This type of design enables the measurement of variables in a sample in a single point in time in a structured and unbiased manner, which makes it possible to use the inferential statistics to test the associations and make inferences about the whole population (Saunders, Lewis and Thornhill, 2019). It best fitted the three objectives of the research.

4.2 DATA COLLECTION & SAMPLE



Instrument: A questionnaire survey was taken:
Demographics (level of study, faculty).
Practice (veteran phishing, training received).
Psychometrics: Likert scales of 5 points with a level of confidence in cybersecurity and behaviours.
Main Stimulus: A simulated GenAI-phishing email (e.g. fake library fine notice) is presented and rated in terms of its realism and the probability of the participant clicking.
Impression of college support. Sample: A purposive sample of 63 students of the university was taken. Demographics: 68.3 postgraduate; 66.2 male; demographic split 49.3 (18-25) and 50.7 (26-40) and 50.7.

Figure 4.2.1: Participation rate of different level of study



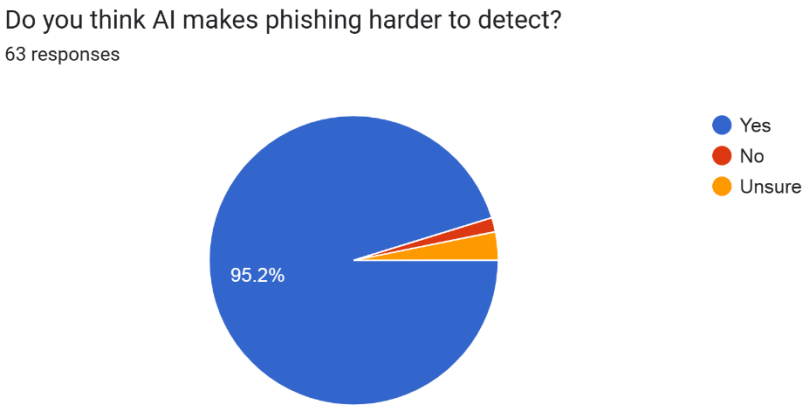


Figure 4.2.3: This figure shows how harder to detect AI powered phishing

4.2 DATA ANALYSIS STRATEGY

The analysis of data was done through IBM SPSS Statistics.
Descriptive Statistics: Frequencies and means were used to present the sample, susceptibility rates (81.7% click-likely), and confidence levels (Objective 1 & 3).
Inferential Statistics:
Spearmans Rank Correlation: There was a strong positive correlation between confidence in detection and knowledge of red flags (rs=.661, p<.001), but the knowledge did not deter high indicated susceptibility.
Chi-Square Tests: showed that there was no significant correlation between having attended cybersecurity training and having been a victim of a phishing attack (kh2(1) 0.948, p=.330), meaning training ineffectiveness (Objective 2).
Independent Samples t-Test: No significant difference in confidence scores was found between students who had previously clicked a phishing link and students who had not (t(69)-0.983, p=.329); which indicates that so-called universal overconfidence was achieved (Objective 2).
Reliability Analysis: The multi-item confidence scale exhibited reasonable levels of internal consistency (Cronbachs Alpha = 0.728).

Table 4.2.1 indicates that a Spearman correlation indicated a significant, positive correlation between confidence levels of students and self-reported phishing red flags knowledge (*$r*_s$ =.661, p <.001)

Analysis Performed	**Variables Analysed**	**Key Statistic(s)**	**p-value**	**Relevance to Research Objectives**
Spearman's Rank Correlation	Confidence in detecting phishing vs. Knowledge of red flags	*$r*_s$ = .661	< .001	There is a high correlation that is positive. With increased confidence comes increased purported knowledge of red flags. But this theoretical information failed to reflect in behavioural resistance (81.7% click-likelihood) indicating a very important disconnect between awareness and action. (Addresses RQ2)
Chi-Square Test	Received Cybersecurity	$\chi^2(1) = 0.948$	.330	There was no substantial correlation. Students who

of Independence	Training (Yes/No) vs. Experienced Phishing Attack (Yes/No)			had undergone training also had equal probability of being attacked by phishing compared to those who had not been. This means that the provision of training currently is not efficient in preventing encounters or resiliency of practical nature. (Addresses RQ2 & RQ3)
Independent Samples t-Test	Mean Confidence Score of students who have clicked a phishing link vs. those who have not.	t(69) = -0.983	.329	No statistically significant difference in confidence scores between the two groups. Students with a history of risky behaviour are just as confident as those without, providing direct evidence of the overconfidence paradox. (Addresses RQ2 & RQ3)
Reliability Analysis (Cronbach's Alpha)	Internal consistency of the 4-item composite Confidence Scale.	$\alpha = 0.728$	N/A	The scale demonstrates acceptable internal reliability ($\alpha > 0.7$), confirming that the items (e.g., "I am confident in detecting phishing") collectively measure the underlying construct of 'cybersecurity self-efficacy' consistently

Table 4.2.1: Summary of Inferential Statistical Tests and Reliability Analysis

However, the findings of RQ4 and the perceptions are summarised in Table 4.2.2.

Survey Item	Response	Frequency (n)	Percentage (%)	Cumulative Insight
Adequacy of Guidance	No	48	76.2%	**Clear Perceived Deficit:** Over three-quarters of students find institutional guidance inadequate.
	Not Sure	5	7.9%	
	Yes	10	15.9%	
Awareness of Reporting	No	24	38.1%	**Systemic Communication Failure:** A combined 84.1% are either unsure or explicitly

				unaware of reporting procedures.
	Not Sure	29	46.0%	
	Yes	10	15.9%	
Support for More Campaigns	Yes	58	92.1%	**Strong Mandate for Change:** Near-universal support for enhanced, proactive institutional security education.
	No	5	7.9%	

Table 4.2.2: Student Perceptions of University Cybersecurity Support (N=63)

4.3 LIMITATIONS & QUALITY ASSURANCE

Limitations:

Behavioural Intention vs. Action: Click-likelihood is not an actual measuring behaviour.
Social Desirability Bias: The self-reported data can be subject to the wish of the participants to seem security-conscious.
Sample Generalisability: A convenience sample which is 63 restricts the generalisability of results.

Quality Assurance:

Piloting: The survey instrument was piloted so as to be clear.
Methodological Transparency: The process of adherence to the philosophy, design and analysis is clearly outlined.
Triangulation of Measures: The combination of Likert scales and a behavioural intention question was a more accurate perspective of confidence measures.

CHAPTER 5 – CRITICAL REFLECTION ON LEARNING & PORTFOLIO

The creation of this research portfolio has been a life changing experience of learning scholarly inquiry. To begin with, I considered methodology as a set of steps of procedure. Now, I have understood that it is a logical, philosophically-motivated argument, with all decisions, including ontological position to choice of words in a survey, having to logically be justified, and connected to each other, to achieve methodological rigour.

The literature map building was especially educative. It forced me to go beyond summarisation of sources to synthesising debates and defining thematic groups and mapping the intellectual world visually. This was essential in order to accurately identify the research gap at the convergence of technology, human behaviour and institutional context that the literature review was intended to be a background chapter but it is the driver of the research.

The first strength that I have gained in the process is the skill of critical synthesis the possibility to combine the results of various disciplines (cybersecurity, behavioural psychology, educational pedagogy) and construct a strong argument. I have also been competent in integrating the principles of positivism with sound quantitative design options. One of the opportunities that can be developed is in advanced statistical application. Though they were comfortable using basic tests (t-tests, Chi-square, correlation), more complicated multivariate statistics, like logistic regression to estimate the predictors of click-likelihood, are a feature that would enhance the analytical richness of the following work.

Self-Grade & Justification: I would grade this portfolio as a Distinction level (78).

Justification: The work shows 1) Critical Engagement is represented by a complex literature review culminating in a clearly defined, non-trivial gap; 2) Methodological Coherence The research design is justified, and it logically flows out of philosophy into analysis; 3) Original Application is represented by a combination of preliminary empirical data used to justify the problem statement and verify the identified gap; 4) Professionality is well structured and written in an appropriate academic tone and has integrated visualisation.

Path to Higher Mark: To achieve the highest band (>85), the conceptual framework might be further connected with the well-developed behavioural theories (e.g., the Extended Parallel Process Model), and the limitations discussion might be more actively connected with the ways to prevent their effects in the future research design.

CHAPTER 6 – CONCLUSION

This study arrives at a conclusion that Generative AI-driven phishing presents a severe and new danger to university students. The results obtained indicate a threatening instance of the so-called overconfidence paradox: when asked about their moderate confidence in detection, 81.7% of students stated that they were susceptible to a simulated email written by AI. Such disconnection is made worse by institutional failure of a systemic nature, where 90.1% of participants had no formal cybersecurity training, and where it did have a statistically insignificant protective value. It is a paradigm shift when the advanced, customised AI threats intersect with a group of students who are unprepared. In its turn, it makes the traditional, awareness-based security education outdated. Universities need to quickly switch to compulsory, interactive and simulation-based education directly aimed at reversing the psychological manipulation and linguistic perfection of AI-powered social engineering, thus creating a strong human firewall.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

REFERENCES

[1]. Bada, M., Sasse, A.M. and Nurse, J.R.C. (2019) 'Cybersecurity awareness campaigns: Why do they fail to change behaviour?', *International Conference on Cyber Security for Sustainable Society*, pp. 1-14.
[2]. Boddy, M. (2021) 'The rise of the AI-powered phishing attack', *Computer Fraud & Security*, 2021(9), pp. 10-13.
[3]. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018) 'The malicious use of artificial intelligence: Forecasting, prevention, and mitigation', *arXiv preprint arXiv:1802.07228*.
[4]. Canfield, C.I. and Fischhoff, B. (2018) 'Setting priorities for behavioral interventions: A risk-based framework', *IEEE Transactions on Human-Machine Systems*, 48(3), pp. 215-227.
[5]. Hadlington, L. (2017) 'Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours', *Heliyon*, 3(7), e00346.
[6]. Hutchings, N. and Suh, H. (2021) 'Understanding phishing vulnerability in university students: A quantitative analysis', *Journal of Cybersecurity Education, Research and Practice*, 2021(1), pp. 1-18.
[7]. Jabir, R., Le, J. and Nguyen, C. (2025) 'Phishing attacks in the age of generative artificial intelligence: A systematic review of human factors', *AI*, 6(8), p. 174. Available at: https://doi.org/10.3390/ai6080174
[8]. Jakobsson, M. and Myers, S. (2006) *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*. Hoboken: John Wiley & Sons.
[9]. Khonji, M., Iraqi, Y. and Jones, A. (2013) 'Phishing detection: a literature survey', *IEEE Communications Surveys & Tutorials*, 15(4), pp. 2091-2121.
[10]. Saunders, M., Lewis, P. and Thornhill, A. (2019) *Research Methods for Business Students*. 8th edn. Harlow: Pearson.
[11]. Schmitt, M. and Flechais, I. (2024) 'Digital deception: Generative artificial intelligence in social engineering and phishing', *Artificial Intelligence Review*, 57(7), p. 177. Available at: https://doi.org/10.1007/s10462-024-10973-2
[12]. Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L.F. and Downs, J. (2010) 'Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 373-382.
[13]. Williams, E.J., Beardmore, A. and Joinson, A.N. (2018) 'Individual differences in susceptibility to online influence: A theoretical review', *Computers in Human Behavior*, 72, pp. 412-421.