TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Farkas Pongrácz 245181IVCM

Evaluating AI-Based Defences Against AI-Generated Social Engineering: A Technical and Human-Centred Analysis

Master's Thesis

Supervisor: Dr. Adrian Nicholas Venables

PhD

Tallinn 2026

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Farkas Pongrácz 245181IVCM

Tehisintellektil põhinevate kaitsemeetmete hindamine tehisintellekti abil loodud sotsiaalse manipuleerimise vastu: tehniline ja inimkeskne analüüs

Magistritöö

Juhendaja: Dr. Adrian Nicholas Venables

PhD

Tallinn 2026

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Farkas Pongrácz

18.05.2026

Abstract

This thesis evaluates the effectiveness of AI-based defences against AI-generated social engineering attacks through a combined technical and human-centred approach. The rise of generative AI has significantly increased the realism and scalability of phishing attacks, requiring evaluation methods that go beyond technical performance alone.

The research follows a two-phase design. Phase 1 assessed the robustness of five detection models using AI-generated phishing emails and paraphrased variants. Although paraphrasing was designed as an adversarial manipulation, its effect was not uniformly degradative. Traditional and fine-tuned models showed only limited performance changes, while the zero-shot LLM classifier improved under Condition B. This suggests that paraphrasing did not function as successful evasion for all models, but instead exposed model-dependent differences in what each classifier treated as suspicious.

Phase 2 examines user interaction through a controlled study comparing four explanation styles: opaque, rule-based, example-based, and contextual. Results show that explanations did not improve decision accuracy; the opaque condition achieved the highest accuracy. Example-based explanations increased user compliance but also led to higher over-reliance on incorrect AI outputs. Additionally, users with cybersecurity expertise demonstrated higher accuracy and more calibrated trust than non-experts.

Overall, the findings reveal a gap between technical performance and human use. Even highly accurate systems require appropriate trust calibration to be effective. The study contributes to the design of human-centred AI security systems by demonstrating that effective defence depends on both robust models and well-designed user interaction.

This thesis is written in English and is 70 pages long, including 7 chapters, 26 figures and 3 tables.

Lühikokkuvõte

Käesolev magistritöö uurib tehisintellektil põhinevate kaitsemehhanismide tõhusust tehisintellekti abil loodud sotsiaalse manipuleerimise rünnete vastu, kasutades tehnilist ja inimkeskset lähenemist. Generatiivse tehisintellekti areng on suurendanud õngitsusrünnete realismi ja ulatust, mistõttu on vajalik hinnata kaitsesüsteeme lisaks tehnilisele täpsusele ka kasutajate vaatenurgast.

Uuring koosneb kahest etapist. Esimese etapi käigus hinnati viie tuvastusmudeli töökindlust, kasutades tehisintellekti loodud õngitsuskirju ja nende parafraseeritud variante. Kuigi parafraseerimine oli kavandatud adversariaalse manipulatsioonina, ei olnud selle mõju kõigi mudelite puhul ühtlaselt jõudlust vähendav. Traditsioonilised ja peenhäälestatud mudelid näitasid vaid piiratud muutusi tulemustes, samas kui nullnäitega LLM-klassifikaatori jõudlus paranes tingimuses B. See viitab sellele, et parafraseerimine ei toiminud kõigi mudelite puhul eduka vältimisstrateegiana, vaid tõi esile mudelispetsiifilised erinevused selles, milliseid tunnuseid iga klassifikaator pidas kahtlaseks.

Teises etapis uuritakse kasutajate käitumist nelja selgitusstiili puhul: läbipaistmatu, reeglipõhine, näitepõhine ja kontekstuaalne. Tulemused näitavad, et selgitused ei paranda otsuste täpsust; parima tulemuse andis läbipaistmatu tingimus. Näitepõhised selgitused suurendasid kasutajate usaldust, kuid ka ülemäärast sõltuvust tehisintellektist. Samuti saavutasid küberjulgeoleku taustaga osalejad paremaid tulemusi kui mitteeksperdid.

Tulemused näitavad, et tehniline täpsus ei taga tõhusat kasutamist. Tehisintellekti süsteemide edukus sõltub kasutajate usalduse õigest kalibreerimisest. Käesolev töö rõhutab, et tõhusad turvalahendused peavad ühendama tehnilise töökindluse ja inimkeskse disaini.

Lõputöö on kirjutatud Inglise keeles ning sisaldab teksti 70 leheküljel, 7 peatükki, 26 joonist, 3 tabelit.

List of abbreviations and terms

AI	Artificial Intelligence
ANOVA	Analysis of Variance
BAE	BERT-based Adversarial Examples
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
CLT	Cognitive Load Theory
CNN	Convolutional Neural Network
DIPPER	Document-level Interpretable Paraphrasing for Privacy and Evasion Resistance
DL	Deep Learning
F1	F1-score
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-Trained Transformer
GRU	Gated Recurrent Unit
HCI	Human-Computer Interaction
HITL	Human-in-the-Loop
ID	Identifier
LLM	Large Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning
OSINT	Open Source Intelligence
POS	Parts-of-Speech
RQ	Research Question
SD	Standard Deviation
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator
XAI	Explainable Artificial Intelligence

Table of contents

List of figures

List of tables

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Adrian Nicholas Venables for his guidance, support, and valuable feedback throughout the development of this thesis. His insights and constructive suggestions were essential in shaping both the direction and the quality of this work.

I am deeply grateful to my family, Lóránt (father), Krisztina (mother), Karolina (sister) and Alex (our dog) for their continuous support, patience, and understanding throughout my studies. They all helped me shape into the form of who I am today, gave me values to respect and they show me relentless support everywhere my journey takes me. But most importantly, they remind me of my goals every single day

I am also thankful to all participants who took part in the user study. Their time and contribution made the empirical part of this research possible.

Finally, I would like to acknowledge my peers and for their support and discussions about this work.

1 Introduction

1.1 Motivation

In the contemporary digital ecosystem, the architecture of cybersecurity has evolved into a fortress of encryption, firewalls, and anomaly detection systems designed to render technical infrastructure impervious to unauthorized access. Yet, despite these sophisticated fortifications, organizational security perimeters are breached with alarming frequency. The persistence of these breaches highlights a fundamental asymmetry in cyber defence: while hardware and software have been hardened, the human operator remains the most accessible attack surface. Social engineering – the psychological manipulation of individuals into divulging confidential information or performing actions that compromise security – has long exploited this human element. However, the emergence of GenAI has fundamentally altered the threat landscape, transforming social engineering from a manual, high-effort endeavor into an automated, scalable, and hyper-realistic threat. This thesis investigates the efficacy of AI-based defences against this new wave of attacks, arguing that a purely technical evaluation of detection systems is insufficient. To be truly effective, defences must be evaluated through a combination of technical performance and human-centred interaction.

Social engineering remains a potent threat not because of technological failure, but because it bypasses technology to exploit the cognitive heuristics and behavioural patterns of human users. Unlike traditional cyberattacks that probe for software vulnerabilities, social engineering attacks probe for psychological triggers such as urgency, authority, fear, and trust. As Gunning and Aha (2019) note, the opacity of modern machine learning systems used in defence often mirrors the complexity of the attacks, leaving users unable to distinguish genuine threats from false alarms [1]. Without a transparent rationale for why a communication is flagged as malicious, users are left to rely on their own, often flawed, judgement. Furthermore, human factors engineering suggests that security protocols often fail to align with human cognitive capabilities, leading to cognitive overload or desensitization to warnings [2]. Consequently, the human

user remains a vulnerable endpoint, susceptible to manipulation regardless of the sophistication of the underlying network security infrastructure.

The advent of generative AI has exacerbated these vulnerabilities by introducing a qualitative shift in the nature of social engineering. Historically, phishing and pretexting attacks were often identifiable by linguistic errors, cultural incongruities, or a lack of personalization due to the resource constraints of the attacker. GenAI has effectively removed these constraints. As discussed by Goldstein et al. (2023), language models can now generate persuasive, context-aware text at scale, driving down the cost of producing propaganda and targeted deception [3]. This allows malicious actors to craft highly personalised spear-phishing campaigns that mimic the tone and style of trusted colleagues or institutions with uncanny accuracy. Beyond text, GenAI audio and video deepfakes enable attackers to impersonate executives or family in real-time, eroding the foundational trust required for digital communication [4]. This transition from manual to automated, high-fidelity deception signifies an "enlarging and enriching" of the threat landscape, where attacks are not only more frequent but for humans to detect [5].

1.2 Research Purpose

In response to this escalated threat, the cybersecurity community has increasingly turned to AI-based defence mechanisms. These systems utilize machine learning to detect patterns indicative of AI-generated content or malicious intent. However, deploying AI to catch AI introduces a "black box" problem where the decision-making process of the defence system is opaque to the human operator. While a detection algorithm may achieve high statistical accuracy in a controlled environment, its practical utility is contingent upon the user's ability to interpret and act upon its output. Liao and Varshney (2022) argue that explainability is not merely a technical feature but a human-centric requirement [6]. Without meaningful explanations, users cannot form an appropriate level of trust in the system, leading to either over-reliance or skepticism [6]. If a defence system flags a sophisticated AI-generated email as malicious but fails to explain why, the user may dismiss the warning as a false positive when the message appears perfectly written.

Therefore, the evaluation of AI-based defences cannot be restricted to technical metrics such as precision, recall, or F1 scores. A system that technically detects a threat but fails to communicate that threat effectively to the human user is, in practice, a failed defence.

This necessitates a shift toward a human-centred evaluation paradigm. As highlighted by Kim et al. (2024), there is often a disconnect between the algorithmic performance of XAI and its actual impact on human decision-making performance [7]. An effective evaluation must assess not just whether the system detects an attack, but whether the explanation provided is "meaningful" to the user – contextualized, actionable, and cognitively accessible. Integrating human factors engineering into the evaluation process helps ensure that defences support human cognitive processes rather than overwhelm them, bridging the gap between algorithmic complexity and user understanding. [2].

Despite the recognized need for this human-centred approach, a critical research gap remains at the intersection of AI-generated threats and AI-driven defences. On one side, attackers are leveraging Large Language Models (LLMs) to produce highly convincing, adversarially paraphrased phishing messages that evade traditional detection. On the other, existing defence tools and user-facing explanations have not been systematically evaluated for their resilience against these modern, generative threats. **There is limited evidence on which AI-based detection models remain robust against AI-generated phishing attacks. There is also little understanding of how explanation design shapes user trust calibration in AI-driven phishing warnings.**

This thesis contributes a dual-phase evaluation framework that combines adversarial benchmarking of AI-based phishing detection models with a controlled user study assessing how explanation design shapes human trust calibration in AI-assisted cybersecurity.

1.3 Research Questions

This thesis posits that the defence against AI-generated social engineering requires a symbiotic relationship between the technical robustness of detection models and the cognitive resilience of human operators. This thesis addresses 3 research questions.

RQ1: Do different AI-based phishing detection models show significant differences in robustness when evaluating LLM-generated phishing messages, including paraphrased or contextually adapted variants?

RQ2: Does a rule-based explanation style improve user trust calibration and decision accuracy compared to opaque or example-based explanations when interacting with AI phishing warnings?

RQ3: Do example-based explanations increase user compliance with AI warnings, and to what extent do they contribute to over-reliance or blind trust when the AI is incorrect?

1.4 Scope and Goal

Consequently, this research aims to develop a comprehensive evaluation framework that rigorously tests AI-based defences against AI-generated social engineering through a dual lens. The scope of this study is deliberately limited to the detection of phishing-style social engineering messages in written digital communication, such as emails. While social engineering attacks occur across various modalities – including audio and video deepfakes – restricting the analysis to text-based interactions allows for controlled experimentation and a systematic comparison of defensive models.

The primary goal is to benchmark the technical robustness of four distinct categories of detection models: traditional machine learning, deep learning, prompted Large Language Models (LLMs), and fine-tuned LLMs. By evaluating these models against AI-generated phishing content, the study seeks to measure their resilience against adversarial paraphrasing and linguistic variation.

Furthermore, the research aims to assess the human-centric quality of these defences by conducting a controlled user study. This phase investigates how four specific explanation styles – opaque, rule-based, example-based, and contextual – influence user perception. The objective is to determine how these different forms of algorithmic transparency affect user trust calibration and decision accuracy. These will ultimately contribute to the development of security systems that are both technically effective and human-compatible.

1.5 Novelty

This thesis introduces a unified evaluation of AI-generated threats and AI-based defences within a single experimental framework. Specifically, it:

- benchmarks multiple detection models against adversarially paraphrased AI-generated phishing messages rather than static datasets,
- evaluates explanation styles (opaque, rule-based, example-based, contextual) in a controlled phishing detection task,
- and analyzes the interaction between model output, explanation design, and user trust calibration.

By combining adversarial robustness testing with a human-centred user study, this work moves beyond isolated technical or usability evaluations and instead examines the full sociotechnical system of AI-assisted cybersecurity.

1.6 Chapter Overview

This thesis is structured into seven chapters, each addressing a specific component of the research.

Chapter 1 introduces the research problem, motivation, and objectives, and outlines the scope and research questions of the study. It also highlights the novelty of the work and its relevance within the evolving landscape of AI-generated social engineering.

Chapter 2 describes the methodology employed in this research. It presents the overall two-phase study design, consisting of a technical evaluation of AI-based detection models and a controlled user study. The chapter details the dataset construction, experimental procedures, and evaluation metrics used in both phases.

Chapter 3 reviews the relevant literature on AI-generated social engineering, phishing detection models, user behaviour in response to security warnings, and explainable AI. It identifies key research gaps that this thesis aims to address.

Chapter 4 presents the results of the study. It first reports findings from the technical evaluation of detection models (Phase 1), followed by results from the controlled user study examining explanation styles and their impact on user decision-making (Phase 2).

Chapter 5 provides a discussion of the results, interpreting the findings in the context of existing research and theoretical frameworks. It examines the implications for trust calibration, explanation design, and human-AI interaction in cybersecurity.

Chapter 6 outlines areas for future research, highlighting potential extensions of the work, including adaptive explanation systems, improved model robustness, and evaluation in more realistic operational environments.

Chapter 7 concludes the thesis by summarising the key findings, answering the research questions, and reflecting on the broader implications of integrating AI-based defences with human decision-making processes.

This thesis argues that AI-based defence against AI-generated social engineering must be evaluated as a sociotechnical problem rather than a purely technical one. The central issue is not only whether a model can detect phishing content, but also whether its outputs remain robust under linguistic transformation and whether users interpret and act on them appropriately. Accordingly, the thesis examines both technical robustness and human trust calibration within a single framework.

2 Methodology

2.1 Research Design and Approach

This thesis employs a two-phase mixed-methods research design to comprehensively evaluate the efficacy of AI in the context of social engineering. The study is situated within an emerging "triadic" adversarial environment, characterized by the interaction between AI-generated threats, AI-based defensive systems, and human decision-makers. This methodological approach addresses the research problem by integrating a system-level technical benchmarking analysis with a human-centred user study.

The primary objective is to bridge the gap between technical detection capabilities and human usability in cybersecurity. While traditional methodology often isolates these elements – evaluating algorithms on static datasets or studying user behaviour with simulated, non-adversarial warnings – this research design unifies them. It assesses how detection models withstand the specific linguistic capabilities of LLMs and subsequently investigates how the outputs of these detection models can be explained to users to optimize trust calibration.

The research is structured into two distinct but interconnected phases:

1. **Phase 1** focuses on the technical robustness of detection systems. It utilizes experimental benchmarking to quantify the resilience of diverse AI detection architectures against LLM-generated and adversarially paraphrased phishing content.
2. **Phase 2** focuses on the human-computer interaction aspect. It uses a controlled user study to determine how different Explainable AI styles influence user accuracy, confidence, and reliance when interacting with warnings generated by the systems evaluated in Phase 1.

2.2 Overview of the Two-Phase Study

The study follows a sequential exploratory design. The insights gained from the technical evaluation in Phase 1 regarding model performance and failure modes inform the context of the user study in Phase 2.

In **Phase 1 (Technical Evaluation)**, the study constructs a specialized dataset comprising legitimate communication and AI-generated phishing attacks. These attacks are subjected to adversarial paraphrasing to simulate sophisticated evasion techniques. Five distinct categories of detection models – ranging from traditional machine learning to fine-tuned LLMs – are evaluated against this dataset. The goal is to isolate the impact of AI-driven linguistic variations on detection performance (RQ1).

In **Phase 2 (Controlled User Study)**, the focus shifts to the "last line of defence": the user. Using a within-subjects experimental design, participants evaluate email messages accompanied by AI-generated warnings. The study manipulates the explanation style of these warnings (Opaque, Rule-based, Example-based, and Contextual) to measure their effect on trust calibration. This phase seeks to understand whether specific explanation designs can mitigate over-reliance on incorrect AI predictions or under-reliance on correct ones (RQ2 and RQ3).

2.3 Phase 1: Technical Evaluation of AI-Based Detection Models

The first phase establishes the technical baseline of the research. It rigorously tests the hypothesis that AI-based phishing detectors exhibit varying levels of robustness when faced with LLM-generated content, particularly when that content is paraphrased or contextually adapted.

In this thesis, robustness refers to the extent to which detection performance remains stable when phishing messages are linguistically transformed. An adversarial manipulation refers to a transformation designed to challenge detection while preserving malicious intent. Evasion refers specifically to a successful adversarial outcome in which the transformation reduces detection performance. This distinction is important in the present study because Condition B was adversarial by design, but not uniformly evasive in outcome.

2.3.1 Dataset Construction

To reflect the evolving threat landscape, a custom dataset was constructed comprising both malicious and legitimate workplace messages. The dataset design controls for

confounding variables such as message length and complexity while introducing controlled adversarial variation.

The full dataset generation pipeline, prompts, and configuration parameters are version-controlled and publicly available to ensure reproducibility and transparency. All generated phishing content was created solely for academic research and is not deployed or disseminated beyond controlled evaluation environments.

AI-Generated Phishing Messages

The malicious portion of the dataset consists of 120 phishing emails generated using Gemini 3. These emails represent four prevalent social engineering scenarios:

1. **Credential Theft:** Attempts to solicit login details through deceptive landing pages or forms.
2. **Impersonation:** Pretexting scenarios where the attacker mimics a senior executive or trusted colleague.
3. **Invoice/Payment Fraud:** Requests for urgent payment or changes to banking details.
4. **Account Recovery:** False alerts regarding security breaches or account lockouts requiring immediate action.

Each scenario contains 30 original messages.

Adversarial Paraphrasing

To simulate the evasion tactics of a modern attacker, the study employs an adversarial paraphrasing process. Each of the 120 original phishing messages serves as a seed for five distinct variants. These variants are created through targeted prompting of the LLM, instructing it to alter the linguistic style without changing the underlying malicious intent. Prompt strategies include instructions such as:

- Professional tone reformulation
- Friendly tone rewriting
- Removal of suspicious wording
- Structural reordering

- Reduction of urgency cues

This process yielded 600 paraphrased variants. The total phishing pool therefore comprises 720 messages (120 original + 600 paraphrased).

Legitimate Messages

To provide a control condition, 120 legitimate workplace emails were syntethically generated to resemble standard professional communication. These messages were screened to ensure they did not contain phishing-like urgency cues or structural anomalies that could bias classification.

Standardization and Quality Control

All messages underwent a screening process to ensure comparability in terms of length, tone, and syntactic complexity. Messages were constrained to a controlled word-length range and screened for incomplete or duplicate outputs. Duplicate detection was performed via hashing, and truncated outputs were discarded and regenerated. This standardization ensures that classifier performance is driven by linguistic indicators of deception rather than superficial structural differences.

It is important to note that this study benchmarks detection models against AI-generated phishing messages rather than a fully representative real-world distribution of attacks. While recent evidence suggests that AI-generated phishing is becoming increasingly prevalent, real-world phishing campaigns still include a mix of human-authored, templated, and hybrid messages. Therefore, the dataset used in this study should be interpreted as a controlled approximation of emerging threats rather than a complete representation of the current threat landscape.

2.3.2 Detection Models

Five representative detection approaches were selected to span the spectrum of AI-based defence paradigms, ranging from traditional statistical classifiers to large language models. The selected models represent progressively increasing representational capacity and contextual awareness, enabling comparative analysis of how model complexity relates to adversarial robustness.

1. **Logistic Regression (Baseline ML)**: This model serves as the baseline, utilizing simple text-based features. It employs Term Frequency-Inverse Document

Frequency (TF-IDF) vectorization to identify keyword-based patterns associated with phishing.

2. **Random Forest (Traditional ML)**: This ensemble method is utilized to capture non-linear relationships. It is trained on engineered features, including both lexical attributes (word choice) and structural features (e.g., URL presence, capitalization patterns).
3. **LSTM Neural Network (Deep Learning)**: A Long Short-Term Memory (LSTM) network is employed to evaluate the effectiveness of deep learning in capturing sequential dependencies in text. This model learns patterns automatically from the raw text sequence without manual feature engineering.
4. **Prompted GPT-4 Classifier (LLM Zero-Shot):** This model represents the use of general-purpose LLMs for security tasks. It uses a zero-shot prompting approach in which the model receives a structured instruction to classify a message as phishing or legitimate using its internal knowledge, without task-specific training.
5. **Fine-Tuned BERT Model (LLM Fine-Tuning)**: This represents the current state-of-the-art in specialized text classification. A Bidirectional Encoder Representations from Transformers (BERT) model is fine-tuned specifically on the study's phishing dataset to adapt its contextual understanding to the domain of social engineering detection.

2.3.3 Evaluation Procedure and Metrics

The evaluation procedure is designed to measure both raw performance and adversarial robustness.

Training and Testing Protocol

The dataset was divided using an 80/20 split between training and testing sets. To prevent data leakage, partitioning was performed at the level of base phishing messages. All paraphrases derived from a given original email were assigned to the same split. This ensured that stylistically similar variants did not appear across partitions.

Five-fold cross-validation was applied within the training partition to ensure model stability and mitigate variance in performance estimation.

All data splits and model initializations were controlled using fixed random seeds to ensure reproducibility.

Supervised models were trained exclusively on original phishing emails and legitimate messages. Paraphrased variants were reserved for evaluation under adversarial conditions.

This design ensures that the robustness evaluation reflects generalization to unseen linguistic transformations rather than memorization of paraphrased patterns.

Experimental Conditions

The evaluation is conducted under two distinct conditions to isolate the effect of AI-generated evasion:

1. **Condition A (Original Messages):** The models are tested against the original set of 120 AI-generated phishing emails. This establishes a baseline for detection performance.
2. **Condition B (Paraphrased Messages):** The models are tested against the 600 adversarially paraphrased variants. This tests the models' resilience to linguistic variation and stylistic mimicry.

Performance Metrics

Quantitative performance is assessed using standard classification metrics:

- **Accuracy:** The overall percentage of correct classifications.
- **Precision:** The ratio of correctly identified phishing emails to all emails flagged as phishing.
- **Recall (Sensitivity):** The ratio of correctly identified phishing emails to the total number of actual phishing emails.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced view of performance.
- **Confusion Matrices:** To visualize true positives, false positives, true negatives, and false negatives.

Defining Robustness

Robustness is operationalized as the relative degradation in performance between the original and paraphrased conditions:

$$\Delta F1 = F1_{\{Original\}} - F1_{\{Paraphrased\}}$$

$$\Delta Recall = Recall_{\{Original\}} - Recall_{\{Paraphrased\}}$$

A smaller Δ indicates higher robustness against adversarial linguistic variation.

2.3.4 Statistical Analysis

Performance metrics were computed across five cross-validation folds. Statistical tests were conducted on fold-wise performance scores to determine whether differences between conditions were statistically significant. Depending on normality assumptions, paired t-tests or repeated-measures ANOVA were applied.

Effect sizes were calculated to quantify the magnitude of performance degradation under adversarial paraphrasing.

2.3.5 Error Analysis

A qualitative error analysis was conducted on misclassified instances to identify recurring linguistic patterns associated with detection failures. Misclassified paraphrased messages were examined alongside their corresponding originals to determine which transformations most frequently degraded model performance. Particular attention was given to tone softening, urgency reduction, and structural reordering.

2.4 Phase 2: Controlled User Study on Explanation Styles

Phase 2 addresses the human-centric research questions (RQ2 and RQ3). It investigates how the presentation of AI-generated security warnings influences the decision-making process of users**.**

2.4.1 Study Design

The study utilizes a within-subjects experimental design. This ensures that individual differences in user capability or skepticism are controlled for, as every participant interacts with all experimental conditions. The order of conditions is counterbalanced to prevent order effects or learning effects.

Procedure

Participants are presented with a sequence of 20 distinct email messages, consisting of 10 phishing attempts and 10 legitimate emails. The interface simulates a standard email client. For every message, the participant is provided with:

1. The full text of the email.
2. An AI-generated classification (labeling the email as either "Phishing" or "Legitimate").
3. An explanation for that classification, corresponding to one of the four experimental styles.

User Tasks

For each trial, the participant must perform two actions:

1. **Decision:** Determine whether the message is phishing or legitimate. This decision effectively measures whether the user chooses to trust the AI's assessment or override it.
2. **Confidence Rating:** Rate their confidence in that decision on a 7-point Likert scale.

The experimental interface automatically logs decision accuracy, the specific choice made (Phishing/Legitimate), the confidence rating, and the response time for each trial.

2.4.2 Explanation Conditions

Four distinct explanation styles are evaluated, reflecting the most common current and emerging paradigms in XAI and security warnings:

1. **Opaque:** This condition provides a classification without any supporting rationale. It serves as a control condition representing standard "black box" warnings.
   - *Example text:* "Warning: This message may be phishing."
2. **Rule-based:** This condition justifies the classification by citing specific, deterministic rules or feature mismatches.
   - *Example text:* "Flagged due to mismatched sender domain and urgent request."
3. **Example-based:** This condition leverages a case-based reasoning approach, comparing the current message to known historical threats.

- *Example text:* "This message resembles known phishing examples targeting employees."

4. **Contextual (LLM-style):** This condition simulates a natural language reasoning process, offering a concise summary of the anomaly in context.
    - *Example text:* "The language and tone differ from previous communication patterns associated with this sender."

2.4.3 Measures

To evaluate the impact of these explanation styles, the following metrics are collected and analyzed:

- **Decision Accuracy:** The percentage of correctly identified messages (distinguishing phishing from legitimate).
- **Trust Calibration Metrics:**
    - **Over-reliance:** The frequency with which a participant accepts the AI's classification when the AI is incorrect (false positive or false negative).
    - **Under-reliance:** The frequency with which a participant rejects the AI's classification when the AI is actually correct.
- **Confidence Ratings:** Self-reported confidence (1-7) assessed in relation to decision accuracy.
- **Response Time:** The duration taken to reach a decision, utilized as a proxy for cognitive load and engagement depth.
- **Perceived Trust:** Evaluated via a short post-study questionnaire measuring the user's subjective trust in the system.

2.4.4 Data Analysis

The quantitative data from the user study is analyzed to identify significant differences between explanation styles.

- **Repeated-Measures** ANOVA: This test is used to examine differences in accuracy, confidence scores, and reliance rates across the four explanation conditions.

- **Post-hoc Comparisons:** Pairwise comparisons (with appropriate corrections) are conducted to isolate specific differences between styles (e.g., comparing Rule-based vs. Contextual).
- **Correlation Analysis:** Correlations are calculated between subjective trust ratings and behavioural reliance to understand the relationship between user sentiment and actual compliance.
- **Qualitative Coding:** Open-ended comments regarding the clarity and usefulness of the explanations are analyzed using thematic coding to provide context to the statistical results.

2.5 Validity, Reliability, and Data Quality Considerations

To ensure the scientific rigor of the findings, several validity and quality control measures are integrated into the research design.

Technical Validity (Phase 1)

Reliability in the technical benchmarking is ensured through the use of cross-validation techniques and a strict separation of training and testing data. The use of adversarial testing (paraphrasing) ensures that the findings reflect robustness against active evasion rather than just performance on static data.

Experimental Validity (Phase 2)

- **Attention Checks:** To filter out low-quality data, attention checks are embedded within the user study to identify and exclude disengaged participants.
- **Standardization:** All stimuli (messages) are standardized in formatting to prevent confounding variables related to visual presentation (e.g., font size or layout) from influencing user decisions.
- **Interface Consistency:** A consistent experimental interface is maintained across all trials to ensure that any observed differences in user behaviour can be attributed solely to the manipulation of the explanation style.

By combining a rigorous technical benchmark with a controlled psychological experiment, this methodology ensures a robust analysis of the interplay between AI attack generation, AI defence mechanisms, and human decision-making.

While the synthetic dataset enables controlled evaluation of adversarial paraphrasing, it may not fully reflect the linguistic diversity and distribution of real-world phishing campaigns. Similarly, the controlled experimental interface does not replicate the complexity of real email environments, such as inbox context, multitasking, or notification fatigue. Additionally, the sample size and demographic composition (student population) may limit the generalizability of the findings. These factors should be considered when interpreting the results.

2.6 Ethical Considerations

This study adheres to standard ethical principles for human-subject research. Participation in Phase 2 was voluntary, and all participants provided informed consent prior to taking part in the study. Participants were informed about the nature of the task, the type of data collected, and their right to withdraw at any time without penalty.

No personally identifiable information was collected. All data were anonymized at the point of collection, and participant identifiers were replaced with pseudonymous IDs. The dataset used for analysis contains no direct or indirect identifiers that could be used to trace responses back to individuals.

The experimental task involved simulated phishing emails; however, all messages were synthetically generated and presented within a controlled research environment. Participants were not exposed to real malicious content, and no real credentials or sensitive actions were requested.

Several limitations related to ethics and generalizability must also be acknowledged. The participant sample consisted primarily of students (psychology and cybersecurity), which may not fully represent real-world populations. Additionally, the experimental interface simplified real email environments, potentially affecting ecological validity.

Finally, the dataset used in Phase 1 was synthetically generated using large language models. While this allows controlled experimentation, it may not fully capture the diversity and distribution of real-world phishing attacks.

3 Literature Review

3.1 The Niche

Existing research has made important progress on three related but often separate problems: AI-enabled phishing generation, the technical detection of machine-generated malicious text, and the effect of explanations on user trust in AI-supported decisions. However, these strands are rarely brought together within a single evaluation framework. In particular, prior work often assumes that adversarial paraphrasing reduces detection performance and examines explanation styles separately, without considering contexts where both the attack and the defence are AI-mediated. This thesis addresses that gap by combining model robustness testing with a controlled study of user trust calibration.

3.1.1 Human-in-the-Loop Decision Making in Adversarial AI Environments

The rapid spread of generative artificial intelligence has created a triadic adversarial environment in which a human decision-maker works with a defensive AI system to identify threats generated by an adversarial AI. Recent work describes this setting as one requiring robust Human-in-the-Loop (HITL) collaboration that balances AI autonomy, human oversight, and trust calibration [8], [9]. In such contexts, the human is not merely a recipient of AI output, but an active part of the defensive process.

As threat actors increasingly use AI to generate convincing social engineering attacks, evaluations that focus only on algorithmic accuracy become insufficient. Research on human factors in adversarial AI and algorithm-in-the-loop decision-making shows that attackers may influence not only model performance but also how users interpret warnings and explanations, while users often struggle to assess AI advice appropriately under uncertainty [10], [11]. Related work further suggests that human-AI teams are vulnerable to miscalibrated trust, automation bias, algorithm aversion, and weak mental models, particularly when explanations appear fluent but are not necessarily reliable [10], [12], [13], [14].

Taken together, these studies suggest that effective defence against AI-generated social engineering requires both technical robustness and human-centred interface design. Work on adversarial robustness and adaptive human-AI interaction indicates that defensive systems must not only withstand manipulation, but also present outputs in ways that support effective oversight and calibrated trust [15], [16]. For this thesis, this matters because AI-based phishing defence is treated as a sociotechnical problem in which classifier robustness and user judgement are interdependent.

3.1.2 User Perception of AI-Generated Warnings When the Threat is Also AI-Generated

When both the threat and the warning are AI-mediated, users must judge not only the authenticity of the content but also the credibility of the warning itself. Because people generally struggle to identify synthetic media reliably on their own, they are often forced to rely on automated detection systems [17]. This makes purely technical solutions insufficient: effective defence also depends on whether warnings enhance situational awareness and help users interpret system outputs appropriately [18], [19].

The literature suggests that warning effectiveness is shaped by presentation, source attribution, and prior user beliefs. Studies of warning labels show that users respond differently depending on whether the warning is attributed to AI, framed as a process or impact label, and presented as diagnostic or authoritative [20], [21], [22]. At the same time, warning labels do not always translate into better behavioural outcomes, and may even create unintended effects such as misplaced trust in unlabelled content [21], [22]. This indicates that warning design influences perception, but does not automatically produce better judgement.

User responses are also filtered through pre-existing attitudes, cognitive styles, and familiarity with AI. Transparent explanations can improve perceived diagnosticity, but trust remains strongly shaped by baseline beliefs, perceived AI identity threat, and users' sense of self-efficacy when facing deceptive content [23], [24], [25], [26]. Taken together, these studies suggest that explanations and warnings do not simply transfer trust from system to user; instead, they interact with user expectations in ways that can support or distort judgement. For this thesis, this matters because Phase 2 examines whether

different explanation styles promote calibrated reliance rather than blind acceptance of AI warnings.

3.1.3 Defending against AI-generated social engineering: A survey of technical and human-centric approaches

The rise of generative AI has transformed social engineering from a primarily human-to-human threat into a triadic human-AI-AI environment, where offensive AI generates persuasive attacks and defensive AI tries to detect and explain them in real time [27]. This shift has created an AI-on-AI arms race, prompting researchers to argue that static defence mechanisms are no longer sufficient against adaptive generative models [27], [28], [29]. Proposed responses include multimodal detection systems and retrieval-augmented pipelines that combine semantic, conversational, and behavioural analysis to identify AI-driven social engineering [27], [28], [29].

However, the human remains the final decision point and the most vulnerable component in this system. As generative models become more fluent and realistic, they reduce users' natural skepticism and increase the cognitive burden of evaluating both suspicious content and the warnings attached to it [28], [30]. In practice, this means that even technically strong detectors may fail if false positives, alert fatigue, or poorly designed warnings erode trust in the defensive system [27]. Explainability is therefore important not simply as a transparency feature, but as a mechanism for helping users decide when to rely on the system and when to question it.

Taken together, this literature points to a research gap at the intersection of adversarial AI, phishing detection, and human-centred security design. For this thesis, that gap is addressed by examining both how AI-based detectors respond to paraphrased phishing and how users respond to the warnings those detectors produce. The contribution is therefore not only technical or behavioural, but sociotechnical.

3.2 The Threat Landscape (AI Generated Phishing)

3.2.1 Linguistic Characteristics Distinguishing AI-Generated and Human-Written Phishing Emails

Large Language Models (LLMs) have altered the phishing landscape by producing messages that differ linguistically from many traditional human-written attacks. Whereas human-authored phishing often contains typographical errors, awkward phrasing, and inconsistent grammar, AI-generated phishing is more likely to display polished grammar, syntactic uniformity, and a formal tone [31], [32]. Related work also suggests that machine-generated phishing tends to rely on more structured wording and fewer colloquialisms or spontaneous emotional cues than human communication [32].

Beyond surface fluency, AI-generated phishing also exhibits distinct structural and lexical patterns. Studies report longer average word lengths, higher verb and pronoun frequencies, more consistent readability structures, and lower lexical diversity, with repeated vocabulary and entity references appearing more often than in human-written phishing [33], [34]. These features can produce text that appears coherent and professional, but also stylistically repetitive and unusually uniform [31], [34].

The persuasive framing of AI-generated phishing may also differ from traditional attacks. Rather than relying mainly on overt urgency or obvious spam markers, LLM-generated messages often use imperative verbs, professional tone, and blended persuasion strategies to reduce suspicion while maintaining pressure to act [32], [35], [36]. They may also adopt a more collective or institutional voice and avoid informal stylistic cues such as emojis or casual formatting [34], [37]. Taken together, this literature suggests that AI-generated phishing is not simply more grammatically correct than traditional phishing, but may alter the linguistic cues used by both people and detection systems.

3.2.2 The Effectiveness of LLMs in Adversarial Paraphrasing for Evasion

Adversarial paraphrasing has emerged as a major vulnerability in text-based detection systems because it allows attackers to preserve malicious intent while altering the surface patterns on which detectors rely. Early work showed that even single-pass paraphrasing can severely reduce detector performance: Krishna et al. (2023) demonstrated that paraphrased AI-generated text caused DetectGPT accuracy to collapse while preserving

semantic content [38]. Subsequent studies found that recursive paraphrasing can further increase evasion success by introducing enough linguistic variation to remove detectable machine-generated signatures without substantially degrading text quality [39], [40].

More recent work has made these attacks increasingly adaptive. Instead of paraphrasing blindly, some methods use detector feedback to guide token selection or word substitution toward outputs that minimize detection scores, pushing detector performance toward near-random levels [41], [42]. Others use reinforcement learning to optimise paraphrasing policies against multiple detectors simultaneously, achieving high attack success rates while maintaining strong semantic similarity to the source text [43], [44]. This suggests that paraphrasing is not merely a cosmetic rewriting strategy, but an increasingly effective and systematic evasion technique.

Taken together, this literature shows that paraphrasing can meaningfully disrupt machine-generated text detection by altering stylistic cues while preserving underlying content. For this thesis, this matters because Phase 1 evaluates whether phishing detectors remain robust when suspicious messages are rewritten in ways intended to obscure detection-relevant features without changing malicious purpose.

3.2.3 Comparative Success Rates of AI-Generated Versus Traditional Phishing Attacks

Recent literature suggests that AI-generated phishing has rapidly become competitive with, and in some cases superior to, traditional phishing methods. Early studies found that large language models outperformed generic phishing but still trailed expert-crafted human campaigns; later work showed that automated AI spear phishing could match human experts' click-through rates and substantially outperform traditional control emails. [45], [46]. Longer-term and field-based studies reinforce this trend, indicating that AI-generated campaigns can achieve higher click-through and credential submission rates than conventional phishing in real organizational settings [47], [48].

This advantage does not always appear as higher engagement alone, but often as greater persuasive effectiveness once a target engages with the message. Some studies report that AI-generated phishing can achieve higher compromise rates even when traditional emails attract more initial clicks, suggesting that LLM-generated content may be more persuasive at the point of decision. [49]. Similar patterns have been observed in targeted

campaigns against vulnerable populations, where AI-generated phishing has outperformed traditional simulated scams [50]. Together, these findings suggest that generative AI improves not only the scale of phishing but also its conversion potential.

A likely explanation for this growing effectiveness is that AI-generated phishing is often perceived as more persuasive and is harder for users to distinguish from legitimate communication. Comparative studies show that people rate AI-generated phishing as more convincing than both traditional phishing and some human-crafted alternatives, while human detection of AI-generated text remains near chance in some settings. [51], [52]. Taken together, this literature suggests that AI-generated phishing is not simply a faster way to produce familiar attacks, but a qualitatively more convincing form of social engineering. For this thesis, this matters because it justifies evaluating defensive systems against AI-generated and paraphrased threats rather than relying only on traditional phishing examples.

3.2.4 Recent Literature on the Use of Generative AI for Personalised Social Engineering at Scale

Recent literature suggests that generative AI has transformed spear phishing from a labour-intensive attack into a scalable and increasingly automated form of personalised deception. Large language models reduce the time, skill, and cost of gathering victim information and generating tailored messages, making targeted campaigns feasible on a far broader scale than before. [46], [53], [54]. Studies of automated phishing pipelines show that AI systems can scrape digital traces, infer personal characteristics, and generate persuasive spear-phishing content with far less manual effort than human operators [46].

A key development is the ability of LLMs to convert Open Source Intelligence into psychologically tailored lures. Research on personalised persuasion at scale shows that generative models can adapt messages to individual traits, while frameworks such as SpearBot show how victim-specific data can generate convincing, context-aware phishing emails. [55], [56], [57]. This suggests that generative AI does not merely automate message production, but also increases the precision with which attackers can exploit cognitive biases and personal context.

This capability also extends beyond email into broader, multi-channel attack ecosystems. Studies of AI-authored smishing show that users often struggle to distinguish human-

written from AI-generated messages, while other work shows that even novice attackers can run phishing campaigns using jailbroken models or tools such as WormGPT and FraudGPT. [37], [58], [59], [60]. Taken together, this literature suggests that generative AI has changed phishing from a low-quality, manually intensive activity into a scalable and increasingly persuasive form of social engineering. For this thesis, this matters because it justifies using AI-generated and paraphrased phishing messages in Phase 1 and explains why older assumptions about obviously suspicious phishing language are no longer sufficient.

3.3 Technical Detection

3.3.1 Traditional Machine Learning vs Deep Learning in Detecting LLM-Generated Malicious Text

Traditional machine learning models such as Logistic Regression and Random Forest remain relevant for detecting malicious text because they can perform well on baseline classification tasks with modest computational resources. Studies of LLM-generated phishing and related text classification problems show that feature-based models can achieve high accuracy, and in some constrained settings may even outperform more complex transformer architectures [61], [62], [63]. Their main strength lies in efficiently capturing shallow statistical patterns such as term frequencies and n-gram distributions, which makes them attractive for real-time deployment.

However, this efficiency comes with limitations. Because traditional ML models rely heavily on surface-level lexical patterns, they are generally less effective at capturing deeper semantic context and may be more vulnerable to adversarial paraphrasing or complex linguistic restructuring. [63], [64], [65]. This has led many researchers to favour deep learning approaches such as LSTM and transformer-based models, which are better suited to modelling sequential dependencies and contextual relationships in text. Studies using BERT, BiLSTM, and related architectures report stronger performance in detecting nuanced or highly fluent AI-generated malicious content, particularly where contextual interpretation matters [65], [66], [67].

Taken together, this literature suggests a trade-off between efficiency, explainability, and contextual sensitivity. Traditional ML models remain attractive because they are fast and

comparatively interpretable, while deep learning models often perform better on complex language patterns but are more opaque and computationally demanding. [61], [63], [64]. For this thesis, this matters because Phase 1 compares model families that differ not only in accuracy, but also in the linguistic cues they rely on when classifying AI-generated and paraphrased phishing messages.

3.3.2 State-of-the-Art in Machine-Generated Text Detection for Cybersecurity and Spam

Research on machine-generated text detection in cybersecurity and spam has produced several major methodological approaches, including supervised classifiers, zero-shot statistical detectors, and watermarking systems. Supervised methods range from deep learning and transformer-based models to stylometric and feature-based classifiers, all of which aim to identify recurring patterns in automated malicious content [35], [68], [69], [70], [71], [72]. In parallel, zero-shot approaches such as DetectGPT, Fast-DetectGPT, Binoculars, and Ghostbuster attempt to detect machine-generated text without task-specific training by exploiting statistical regularities in model outputs [73], [74], [75], [76]. A third line of work uses watermarking to embed detectable signals directly into generated text, making attribution easier when the generating model can be controlled [77], [78], [79].

Despite this methodological variety, the literature consistently shows that current detectors remain fragile under realistic adversarial conditions. Studies report that both supervised and zero-shot detectors can be disrupted by simple modifications, including temperature changes, recursive paraphrasing, targeted word substitutions, and style-tuning methods that more closely mimic human writing. [39], [80], [81], [82]. Large-scale benchmarking and real-world observations further suggest that this fragility is not only theoretical: machine-generated spam is already widespread, and existing detectors often struggle to generalize reliably outside controlled settings [33], [83], [84]. This indicates that state-of-the-art performance on benchmark tasks does not necessarily translate into robust security performance in practice.

Recent work has also begun to focus on comparative evaluation and attribution frameworks, including ensemble approaches, fingerprinting strategies, and broader taxonomies of methods, tools, and datasets [85], [86], [87], [88], [89] Collectively, this

literature suggests that machine-generated text detection is an active arms race rather than a solved problem. For this thesis, this matters because Phase 1 does not assume that strong baseline detectors are automatically robust; instead, it evaluates how different model families respond when phishing messages are rewritten to obscure detection-relevant cues.

3.3.3 Studies Evaluating the Robustness of Fine-Tuned BERT Models Against Adversarial Attacks in Text Classification

Fine-tuned BERT models are widely regarded as strong performers in text classification because they can capture complex contextual relationships and often achieve high baseline accuracy on deceptive-text detection tasks [90]. This makes them attractive for phishing and machine-generated text detection, particularly in settings where shallow lexical features may be insufficient. However, high baseline performance does not necessarily imply robustness under adversarial conditions.

A consistent finding in the literature is that fine-tuned BERT models remain vulnerable to adversarial text attacks. Studies using attacks such as TextFooler and BAE show that small synonym substitutions, insertions, or contextual replacements can substantially degrade performance while preserving the original meaning for human readers. [91], [92]. More recent benchmarking confirms that transformer-based classifiers, including BERT, can suffer severe accuracy collapse under adversarial pressure, even if they are initially harder to attack than simpler architectures [93], [94].

These weaknesses are often attributed to the fine-tuning process itself. Prior work suggests that task-specific fine-tuning can narrow a model's decision boundaries and reduce its reliance on general linguistic knowledge from pre-training, increasing sensitivity to adversarial paraphrasing and distribution shift. [95]. Related research on machine-generated text detection further indicates that performance may deteriorate as adversarial inputs become longer or more naturalistic, increasing susceptibility to evasion strategies that mimic human writing. [96]. Taken together, this literature suggests that fine-tuned BERT models should not be assumed to be robust simply because they perform well in baseline classification. For this thesis, this matters because Phase 1 evaluates whether a fine-tuned BERT detector remains stable when phishing messages are paraphrased to obscure detection-relevant cues.

3.3.4 Performance of Zero-shot LLM Classifiers in Detecting Phishing Intent

Zero-shot large language model classifiers have become an increasingly important approach to phishing detection because they rely on general pre-trained knowledge rather than task-specific fine-tuning. Recent studies suggest that modern LLMs can identify phishing intent with strong baseline performance, often achieving high accuracy and F1-scores using only prompt instructions and the email content itself [97], [98]. This indicates that such models can recognise deceptive linguistic and structural patterns without requiring additional training examples.

Evaluations of state-of-the-art systems further show that zero-shot performance can remain strong even under constrained conditions. Studies of GPT-4, Gemini, Claude, and related models report high accuracy and precision when classifying phishing emails, including cases where header information is removed or messages are adversarially rephrased [45], [99], [100]. These findings suggest that zero-shot LLMs may rely on broader semantic and pragmatic cues, such as urgency, implausible requests, or contextual inconsistency, rather than only on superficial lexical markers.

At the same time, zero-shot performance is not uniform across models and may involve trade-offs. Comparative studies show substantial variation between architectures, with some systems performing strongly and others exhibiting weaker accuracy or more conservative behaviour under the same prompting conditions [101], [102]. In particular, zero-shot classifiers may sometimes favour safety over precision, increasing false positives by flagging legitimate messages as suspicious [97]. Taken together, this literature suggests that zero-shot LLMs are promising phishing detectors, but their effectiveness depends on model architecture and decision tendencies. For this thesis, this matters because Phase 1 includes a zero-shot LLM classifier as a distinct model familywhose robustness may differ from both traditional ML and fine-tuned models.

3.3.5 What are Common 'Misclassification Patterns' When AI Models Attempt to Classify Adversarially Paraphrased Text?

A common source of misclassification in adversarially paraphrased text is the removal of the statistical cues on which detectors rely. Many machine-generated text detectors depend on features such as token likelihood, probability curvature, perplexity, or other stylometric regularities. Adversarial paraphrasing methods such as DIPPER are effective

because they preserve meaning while altering surface-level signals, causing detectors to misclassify machine-generated text as human-written once internal thresholds are crossed. [38], [103].

Misclassification also arises when classifiers rely too heavily on narrow lexical or syntactic cues. Prior research shows that text classifiers can be disproportionately sensitive to particular parts of speech, token substitutions, and local lexical patterns, even when overall meaning remains unchanged [104], [105]. Related work on paraphrase detection suggests that models often struggle when lexical overlap is reduced but semantic equivalence is preserved, especially when adversarial inputs introduce negation, structural shifts, or syntactic distractors. [106], [107]. These findings indicate that some models respond more to altered surface form than to underlying malicious intent.

A further vulnerability concerns embedding-space geometry and confidence calibration. Studies suggest that models with dense or weakly separated embedding spaces are more likely to cross decision boundaries after small textual changes, while adversarial paraphrases can reduce confidence and increase unstable or erroneous outputs. [108], [109], [110]. Taken together, this literature suggests that strong baseline performance does not guarantee robustness under linguistic manipulation. For this thesis, this matters because Phase 1 compares multiple model families under both original and paraphrased conditions rather than assuming that more advanced models are automatically more robust.

3.4 Human-AI Interaction & Explainability

3.4.1 The Influence of Explainable AI on User Trust Calibration in Automated Decision Support Systems

Trust calibration aligns a user's trust in an AI system with its actual capabilities and limitations, so users neither reject correct advice nor accept incorrect recommendations uncritically. [111]. Explainable AI is often proposed as a way to support this calibration by making model reasoning more visible and helping users judge whether the system is likely to be reliable. [112]. In principle, explanations should reduce both under-reliance and over-reliance by giving users a stronger basis for evaluating AI output.

In practice, however, the effect of explanations on trust calibration is inconsistent. Empirical studies show that explanations can sometimes encourage unwarranted reliance by creating false confidence, especially when users treat their mere presence as a sign of trustworthiness rather than evaluating their content critically. [113], [114]. At the same time, highly detailed explanations can reduce agreement with the AI or increase fatigue for some users, especially when they are hard to interpret or poorly matched to user characteristics. [115]. This suggests that explanations do not automatically improve trust calibration and may, under some conditions, distort it.

The literature also shows that trust is difficult to measure and strongly shaped by context. Studies often find a mismatch between self-reported trust and demonstrated reliance, indicating that users may claim skepticism while still following AI advice in practice [116], [117]. Trust is also dynamic rather than stable, increasing with repeated exposure but dropping sharply after visible errors [118]. Together, these findings suggest that the effectiveness of Explainable AI depends not only on explanation content, but also on user expertise, interface design, and how much the system encourages active rather than passive engagement. [112], [119]. For this thesis, this matters because Phase 2 examines whether different explanation styles improve decision quality and calibrated reliance rather than merely increasing user confidence in AI warnings.

3.4.2 The Effect of 'Rule-Based' vs. 'Example-Based' Explanations on User Accuracy in Phishing Detection

As AI systems play a larger role in cybersecurity defence, the way their threat assessments are explained to users becomes increasingly important. In phishing detection, two common explanation styles are rule-based and example-based explanations. Rule-based explanations present explicit criteria or violated cues, such as a mismatched sender domain or an urgent request, while example-based explanations compare the current message to known phishing cases. These two styles reflect different reasoning strategies and may influence user accuracy in different ways [120], [121], [122], [123].

Rule-based explanations are often associated with structured decision-making and clearer boundaries for identifying suspicious contentStudies suggest that users often prefer rule-like feedback in phishing and spam contexts, and that explicit explanations of phishing cues can improve their ability to distinguish malicious from benign messages. [120],

[121]. By contrast, example-based explanations align more closely with case-based human reasoning, and users often find them intuitive and helpful because they connect unfamiliar situations to concrete prior instances [122], [123].

However, the literature does not show a simple winner between the two. Rule-based explanations may perform better when the underlying system is highly accurate, whereas example-based explanations can sometimes help users recognise when the AI may be wrong by offering a more concrete basis for independent judgement. [120], [124]. At the same time, both formats introduce design challenges: example-based explanations may overwhelm users if too dense, while rule-based explanations may fail if overly technical or hard to map onto the message context. [123], [125]. Taken together, this literature suggests that explanation style influences user accuracy, but its effectiveness depends on how the explanation interacts with system accuracy, user cognition, and interface design [126]. For this thesis, this matters because Phase 2 directly compares rule-based and example-based explanations to evaluate how they shape decision accuracy and trust calibration in phishing detection.

3.4.3 Research on Over-Reliance and Blind Trust in AI-Generated Security Warnings

A recurring concern in human-AI interaction is over-reliance, where users follow AI-generated advice even when it is incorrect and would have been better ignored [127], [128]. In high-stakes contexts such as cybersecurity, this tendency is often linked to automation bias, cognitive offloading, and fast, low-effort reasoning under pressure rather than deliberate analytical judgement [129], [130]. While this can reduce workload in the short term, it also increases the risk that users will accept incorrect system outputs without sufficient scrutiny.

Although Explainable AI is often proposed as a remedy, the literature suggests that explanations do not necessarily reduce blind trust. Explanations appear most helpful when they enable users to verify or challenge the AI's recommendation; otherwise, they may simply anchor users more strongly to the system's output [131]. Related work also shows that incorrect or misleading explanations can worsen user reasoning by encouraging flawed decision strategies, while general trustworthiness perceptions do not always match

instance-level reliance behaviour [132], [133], [134]. This suggests that the mere presence of an explanation is not enough to ensure calibrated trust.

These risks are especially acute in cybersecurity, where users and analysts often operate under time pressure, uncertainty, and high cognitive load. Research shows that over-trust in AI can lead to insecure code adoption, acceptance of malicious chatbot advice, or excessive reliance on automated prioritisation in Security Operations Centers [135], [136], [137], [138]. To address this, recent work advocates calibrated skepticism rather than unquestioning trust, using mechanisms such as reliance drills and dynamic autonomy to keep users actively engaged in monitoring AI outputs [8], [127], [132], [139]. For this thesis, this matters because Phase 2 examines not only whether users trust AI warnings, but whether different explanation styles encourage appropriate reliance when the system is correct and resistance when it is wrong. The Impact of Cognitive Load on User Interaction with Security Warnings

Cognitive load plays a central role in how users respond to security warnings because working memory is limited and performance tends to decline when mental demands exceed that capacity [140]. In cybersecurity settings, this means that users may ignore warnings, misunderstand them, or choose the quickest response rather than the safest one. Cognitive Load Theory distinguishes intrinsic load, arising from task complexity, extraneous load, resulting from presentation, and germane load, associated with learning and sense-making [140]. Each of these can shape how effectively a warning is processed.

A major source of cognitive load is interruption. Security warnings often appear while users are focused on another task, creating dual-task interference and forcing them to divide their limited attention [141]. Under these conditions, users may process warnings superficially, dismiss them quickly, or make hasty decisions in order to return to their primary activity [141], [142], [143]. Time pressure further intensifies this effect by encouraging fast, heuristic decision-making rather than slower analytical reasoning, which increases vulnerability to phishing and other risky behaviours [140], [144], [145].

The design of the warning itself can either reduce or amplify this burden. Cluttered, overly technical, poorly timed, or repeatedly presented warnings increase extraneous cognitive load and may cause habituation, leading users to treat them as background noise rather than meaningful security cues [146], [147], [148], [149], [150], [151]. This is especially

problematic for non-experts, who may lack the mental models needed to interpret technical language or evaluate consequences [147], [148]. Taken together, this literature suggests that warning effectiveness depends not only on informational accuracy, but also on whether the interface allows users to process the message under realistic cognitive constraints. For this thesis, this matters because Phase 2 evaluates explanation styles in a phishing task where differences in clarity and processing burden may shape both user accuracy and reliance.

3.4.4 The Asymmetric Impact of AI Errors on User Trust

User trust in AI systems is shaped not only by how often errors occur, but also by what kind of errors they are. In security contexts, AI mistakes are typically experienced as either false positives, where legitimate content is flagged as malicious, or false negatives, where genuine threats go undetected. The literature suggests that these error types affect trust asymmetrically, because they differ in visibility, disruption, and perceived risk [152], [153], [154].

False positives are often more immediately damaging to user trust because they are visible, interruptive, and costly in the moment. They can increase cognitive load, disrupt workflow, and contribute to alert fatigue, making users more likely to ignore future warnings [152], [153], [154]. At the same time, their impact is not always negative: in some domains, users may tolerate visible false alarms if they are manageable or seen as a reasonable cost of caution [155]. This suggests that false positives often erode trust actively, but their acceptability depends on context and recovery cost.

False negatives pose a different kind of problem. Because they are silent failures, users may not realise that an error has occurred, which can foster overconfidence in the system and distort trust calibration [17], [154], [156]. In some cases, the absence of a warning may itself be interpreted as evidence of safety, encouraging misplaced reliance on AI outputs. The literature therefore suggests that false positives and false negatives do not simply reduce trust in the same way. They shape it differently depending on visibility, consequence, and task context [17], [154], [157]. For this thesis, this matters because user reliance on phishing warnings cannot be understood solely through overall accuracy. Trust calibration also depends on how users respond when the system is wrong.

3.4.5 User Studies Comparing 'Opaque' Warnings vs. 'Contextual' Explanations in Digital Security

A longstanding question in cybersecurity usability is whether users respond better to simple opaque warnings or to warnings that explain the nature of the threat. Earlier work found that users often ignore generic, opaque alerts because they do not understand the risk, fail to notice the warning, or become desensitised through repeated exposure and false positives [158], [159]. In such cases, users may fall back on personal heuristics rather than treating the warning as a meaningful signal.

Contextual explanations have been proposed as a way to counter this problem by making the warning more concrete and easier to interpret. Studies show that warnings describing specific risks or consequences tend to increase compliance, improve understanding, and make interruptions feel more justified than vague alerts [160], [161], [162]. Related work also suggests that warnings are more effective when they align with users' mental models and introduce enough friction to prompt deliberate attention rather than habitual dismissal [163], [164], [165].

More recent research has extended this idea to dynamically generated explanations. Studies comparing traditional browser-style warnings with LLM-generated contextual warnings report that tailored explanations are often perceived as more understandable, engaging, and trustworthy than opaque alternatives [166]. However, the broader literature suggests that improved perceived clarity does not necessarily translate into improved decision accuracy. For this thesis, this matters because Phase 2 examines not whether users prefer contextual explanations, but whether different warning styles improve judgement and promote more calibrated reliance in an AI-assisted phishing task.

4 Results

This chapter presents the empirical findings derived from the two-phase mixed-methods research design outlined in Chapter 2. This chapter primarily addresses the thesis's three research questions by evaluating both the technical resilience of AI-based detection models and the human-centred impact of different Explainable AI warning styles. All data collection and experimental procedures followed established ethical guidelines, ensuring participant anonymity in the user study and strict cross-validation in the technical benchmarking to maintain data integrity.

The results are structured sequentially to mirror the study's design, moving from algorithmic performance to human-computer interaction.

Section 4.1 details the outcomes of Phase 1: Technical Evaluation of AI-Based Detection Models. This section presents the accuracy, precision, recall, and F1-scores of five classification paradigms: Logistic Regression, Random Forest, LSTM, prompted GPT-4, and a fine-tuned BERT model. The models are evaluated on a custom dataset of 840 messages, comparing baseline performance on original LLM-generated phishing emails with robustness against adversarially paraphrased variants. The statistical analysis and subsequent error analysis in this section directly address RQ1 by quantifying how linguistic evasion tactics impact different AI defence architectures.

Following the system-level analysis, Section 4.2 presents the behavioural and quantitative findings from Phase 2: The Controlled User Study. This section shifts the focus to the "last line of defence," analyzing how human participants responded to 20 simulated email trials. This section addresses RQ2 and RQ3 by evaluating decision accuracy, reliance rates, response times, and self-reported confidence across four warning conditions: Opaque, Rule-based, Example-based, and Contextual. It determines how different explanation styles influence trust calibration, and whether certain designs successfully mitigate over-reliance or blind trust when the AI makes an incorrect classification.

Together, these two sections provide a holistic evaluation of the defensive landscape, bridging the gap between algorithmic detection capabilities and practical human usability.

4.1 Technical Evaluation of AI-Based Detection Models

This section presents the empirical results from the technical benchmarking of the five selected AI-based detection models. The evaluation compared baseline performance on original AI-generated phishing emails (Condition A) with performance on paraphrased variants (Condition B). Although Condition B was designed as an adversarial manipulation intended to test robustness, the results show that its effect was model-dependent rather than uniformly evasive.

To ensure strict reproducibility and mitigate variance, all supervised models were evaluated using 5-fold cross-validation with a fixed random seed (Seed 9). The zero-shot prompted LLM was evaluated across the identical fold splits to ensure directly comparable testing conditions. Robustness is quantified using the delta metric (Δ) defined as the performance on Condition A minus the performance on Condition B ($\Delta = A - B$). Therefore, a positive Δ indicates performance degradation when faced with paraphrased text, while a negative Δ indicates an improvement.

4.1.1 Overall Baseline Performance (Condition A: Original Messages)

The baseline evaluation establishes how well the models detect original, unmodified LLM-generated phishing messages against legitimate workplace emails. Table 1 summarizes the mean aggregate performance metrics across the 5 folds for Condition A.

Table 1 Baseline Performance (Condition A: Original Messages) across 5-Fold Cross-Validation

Detection Model	**Mean F1-Score (SD)**	**Mean Recall (SD)**
Logistic Regression	0.983 (± 0.010)	0.967 (± 0.019)
Random Forest	0.966 (± 0.019)	0.950 (± 0.035)
Fine-Tuned BERT	0.931 (± 0.056)	0.917 (± 0.098)
BiLSTM	0.928 (± 0.044)	0.917 (± 0.051)
GPT-5 (Zero-Shot Prompted)	0.838 (± 0.040)	1.000 (± 0.000)

(Note: Standard Deviations are provided in parentheses to indicate variance across the 5 validation folds).

The results reveal distinct performance tiers across the modeling paradigms. Surprisingly, the Logistic Regression (Baseline ML) model achieved the highest baseline performance with an exceptional mean F1-score of 0.983. This suggests that the original LLM-generated phishing emails contained highly distinctive keyword patterns that standard lexical analysis could isolate. The Random Forest, BiLSTM, and Fine-Tuned BERT models also demonstrated high proficiency, all achieving F1-scores above 0.92.

Conversely, the Prompted LLM (GPT-5 Zero-Shot) exhibited the lowest baseline F1-score (0.838). However, this metric obscures a critical behavioural trait: the model achieved a perfect mean recall of 1.000. It successfully flagged every single phishing email but suffered from a high false-positive rate, misclassifying legitimate emails as phishing, which dragged down its harmonic mean.

4.1.2 Comparative Analysis & Adversarial Robustness (Condition B)

A notable finding is that paraphrasing did not consistently weaken detection performance across all architectures. While the literature often treats paraphrasing as an evasion tactic, the present results indicate that its practical effect depends on the decision logic of the model being tested. In some cases, paraphrasing may remove overt lexical markers of phishing. In others, it may increase textual coherence or preserve features that certain classifiers treat as suspicious. As a result, adversarial intent should be distinguished from adversarial outcome.

Condition B tested the models' resilience against linguistic variation, simulating an attacker prompting an LLM to rephrase phishing content to evade detection. Table 4.2 presents a comprehensive comparison of all metrics between the original and paraphrased datasets.

Table 2 Comparative Performance Metrics (Condition A vs. Condition B)

Detection Model	Condition	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**A**	0.983	1.000	0.967	0.983
	B	0.968	1.000	0.961	0.980
Random Forest	**A**	0.958	0.979	0.933	0.949

	B	0.969	0.993	0.970	0.981
Fine-Tuned BERT	**A**	0.942	0.951	0.917	0.931
	B	0.938	0.989	0.913	0.949
BiLSTM	**A**	0.929	0.942	0.917	0.928
	B	0.954	0.988	0.938	0.962
GPT-5 Zero-Shot	**A**	0.785	0.723	1.000	0.838
	B	0.890	0.916	0.958	0.924

Contrary to the hypothesis that paraphrasing would degrade performance, four out of five models exhibited improved F1-scores on the paraphrased dataset.

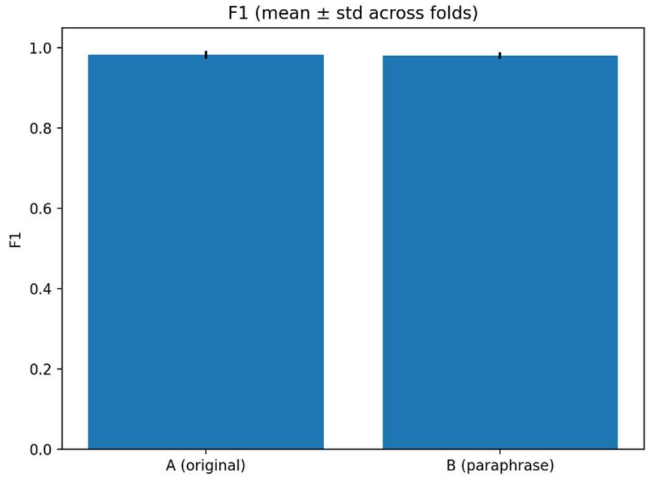

Figure 1 Condition A vs Condition B F1 scores for Logistic Regression model

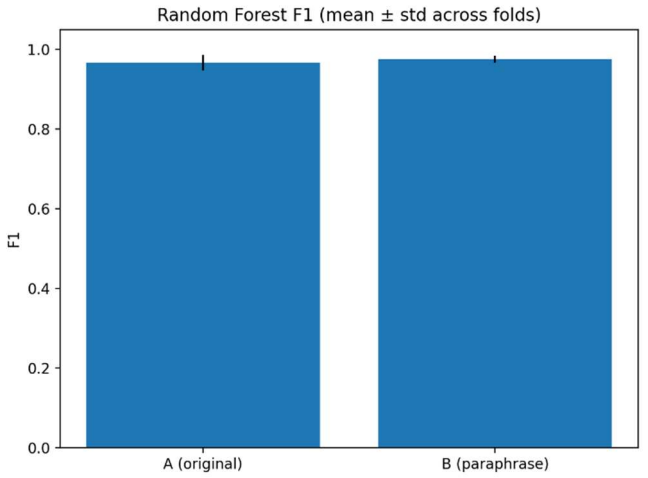

Figure 2 Condition A vs Condition B F1 scores for Random Forest model

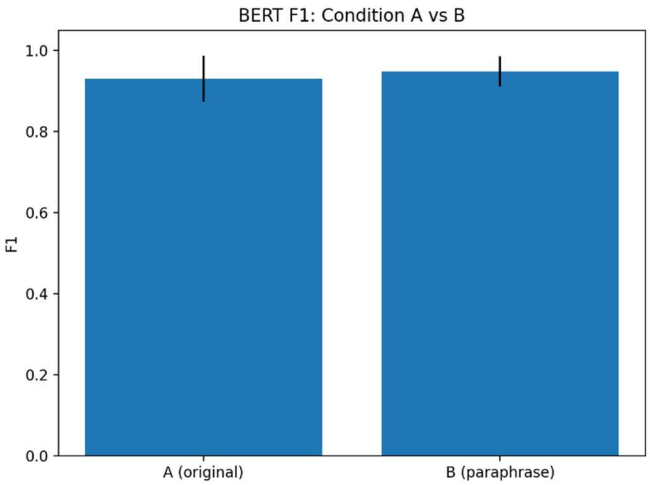

Figure 3 Condition A vs Condition B F1 scores for Fine-Tuned BERT model

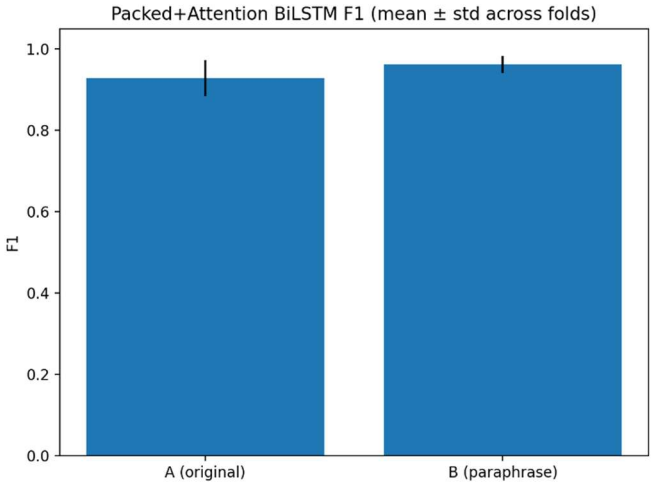

Figure 4 Condition A vs Condition B F1 scores for the LSTM model

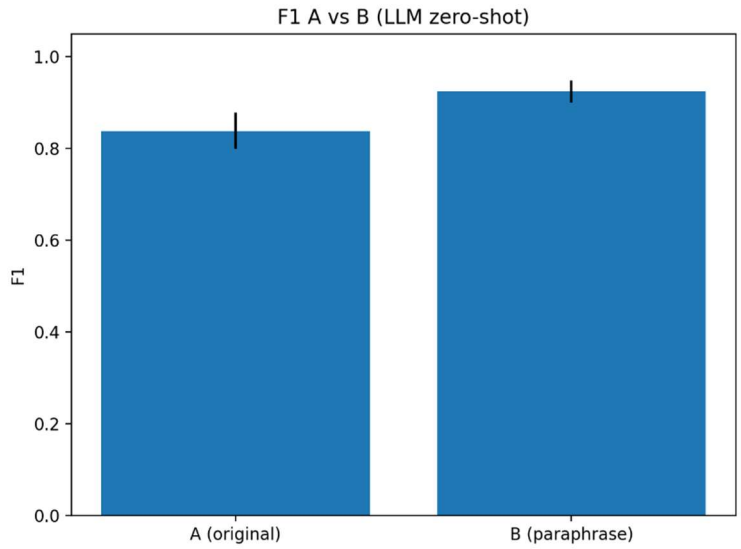

Figure 5 Condition A vs Condition B F1 scores for the LLM zero-shot model

As illustrated in the bar charts and Table 4.2, the prompted zero-shot LLM showed the largest improvement (ΔF1 = −0.086), driven by a sharp rise in precision (from 0.723 to 0.916). The BiLSTM, BERT, and Random Forest models also saw F1 improvements ranging from 1.8% to 3.4%. Only the Logistic Regression model remained statistically static, exhibiting negligible variance between conditions.

4.1.3 Statistical Significance of Robustness

To determine if these performance shifts were statistically significant (addressing RQ1), paired t-tests were conducted on the F1-scores across the five cross-validation folds.

Table 3 Paired T-Test Results for Robustness (ΔF1)

Model	**Mean ΔF1 (A-B)**	***t*-statistic**	***p*-value**	**Result**
GPT-5 Zero-Shot	-0.086	-12.09	< 0.001	Significant Improvement
BiLSTM	-0.034	-1.84	0.139	Not Significant
Fine-Tuned BERT	-0.018	-1.33	0.254	Not Significant

Random Forest	-0.009	-0.84	0.448	Not Significant (robust)
Logistic Regression	+0.003	0.99	0.378	Not Significant (robust)

(Note: A negative t-value here means Condition B (Paraphrased) had a HIGHER score than A.)

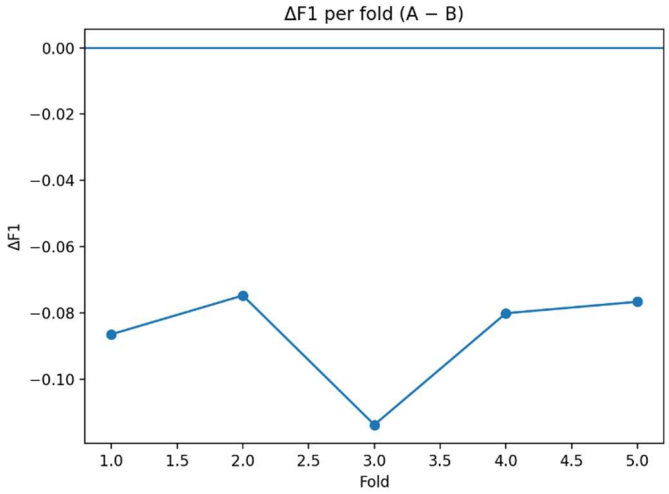

Figure 6 ΔF1 per fold for the LLM zero-shot model

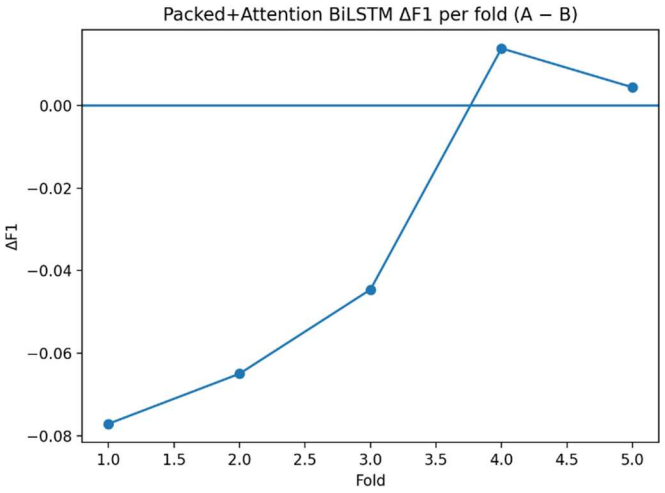

Figure 7 ΔF1 per fold for the LSTM model

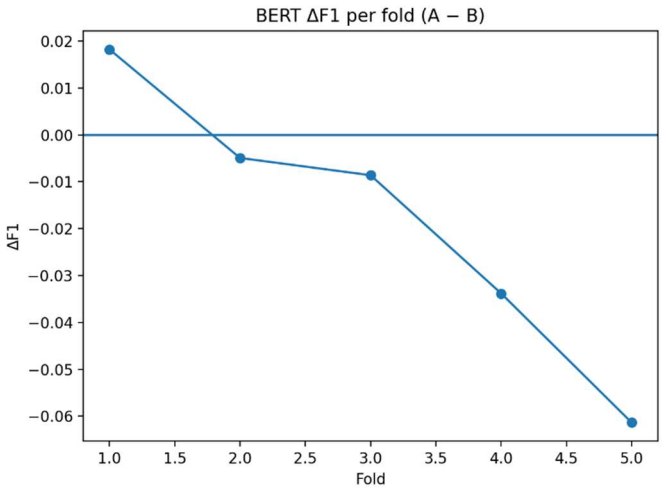

Figure 8 ΔF1 per fold for Fine-Tuned BERT model

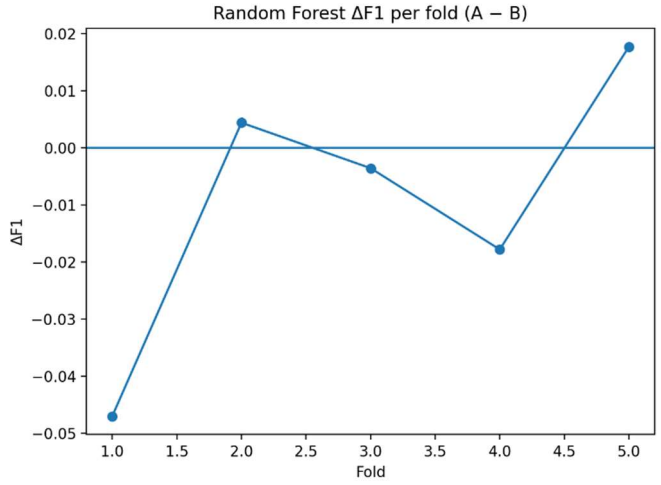

Figure 9 ΔF1 per fold for the Random Forest model

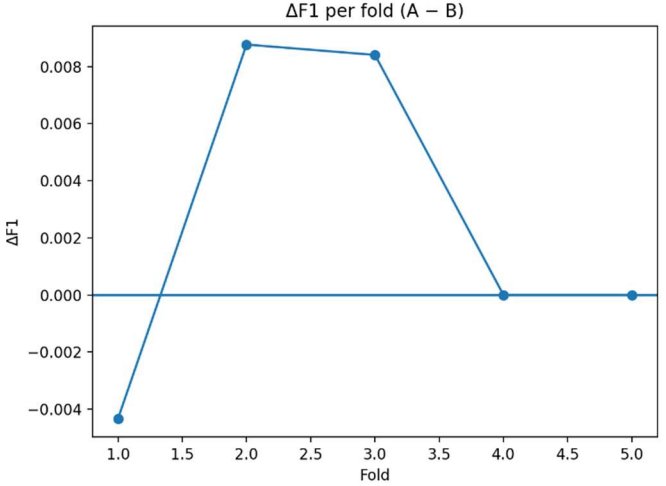

Figure 10 ΔF1 per fold for the Logistic Regression model

The statistical analysis confirms that the performance improvement for GPT-5 Zero-Shot ($p<0.001$) was significant. The "adversarial" paraphrasing inadvertently corrected the Zero-Shot model's tendency to hallucinate threats. The deep learning models (BERT/BiLSTM) showed improvements that were positive but not statistically significant, suggesting high variance across folds.

4.1.4 Error Analysis and Confusion Matrices

To understand why the models performed better on the attacked dataset, an analysis of the confusion matrices was conducted.

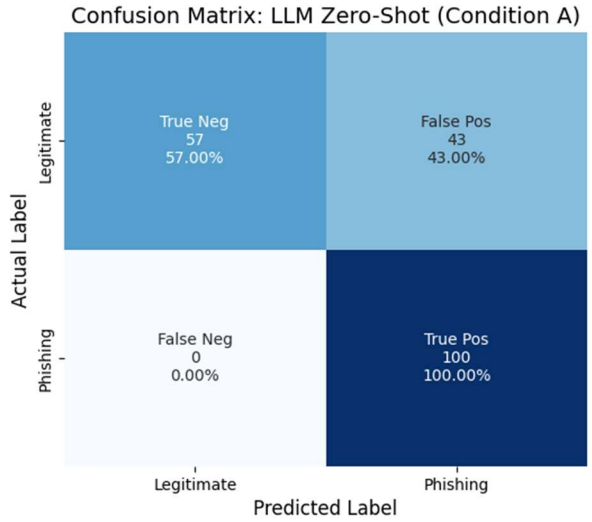

Figure 11 Confusion Matrix: LLM Zero-Shot (Condition A)

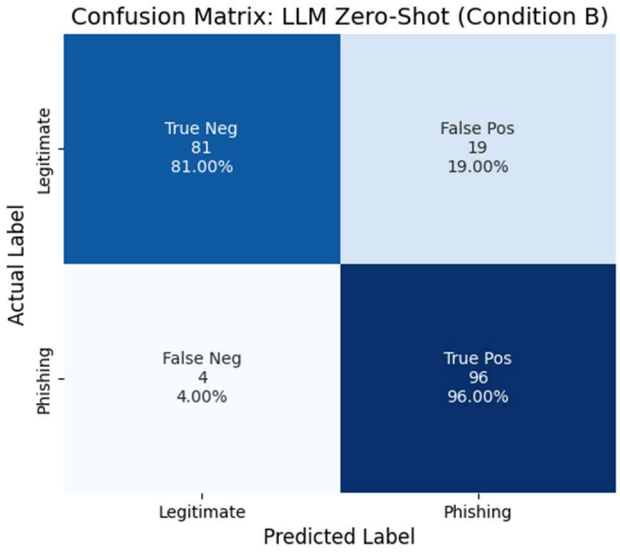

Figure 12 Confusion Matrix: LLM Zero-Shot (Condition B)

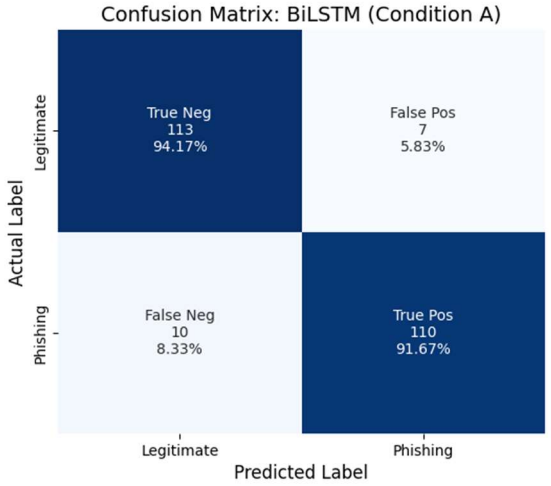

Figure 13 Confusion Matrix: BiLSTM (Condition A)

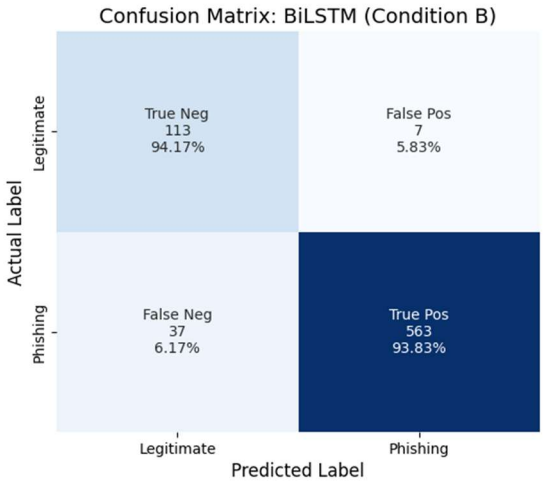

Figure 14 Confusion Matrix: BiLSTM (Condition B)

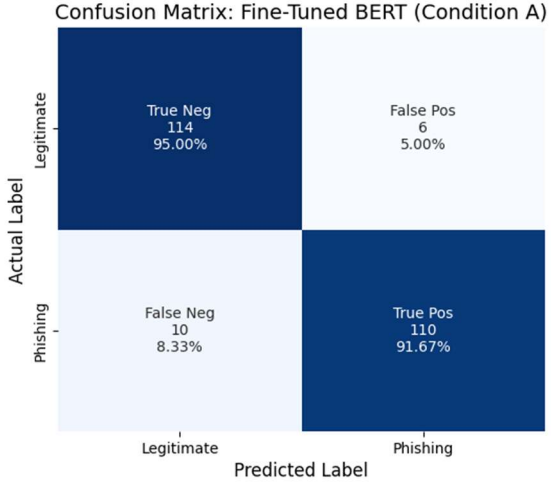

Figure 15 Confusion Matrix: Fine-Tuned BERT (Condition A)

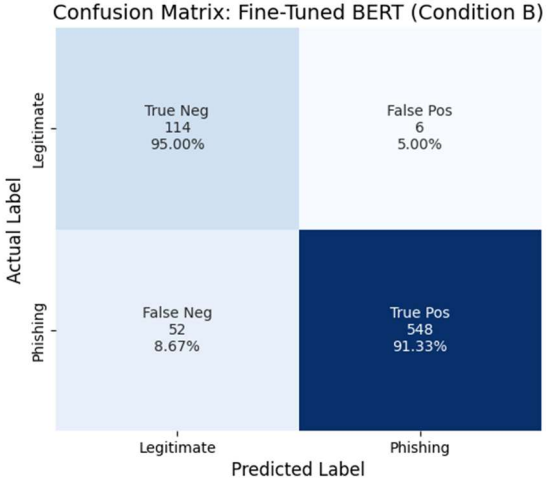

Figure 16 Confusion Matrix: Fine-Tuned BERT (Condition B)

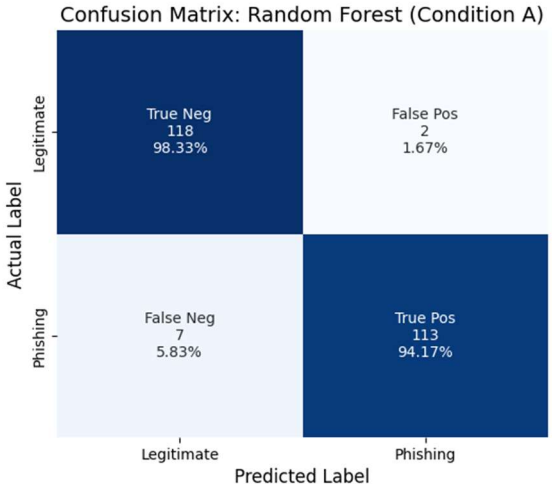

Figure 17 Confusion Matrix: Random Forest (Condition A)

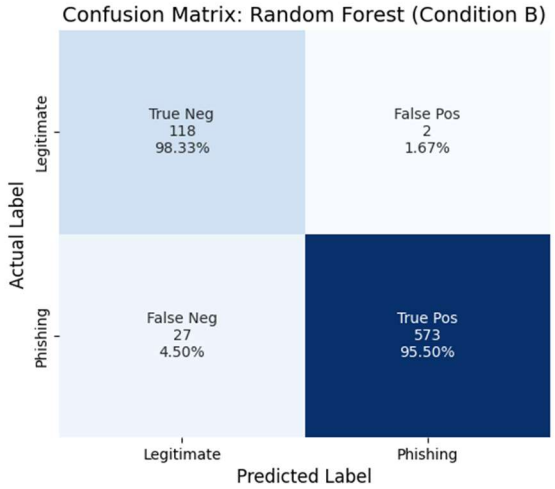

Figure 18 Confusion Matrix: Random Forest (Condition B)

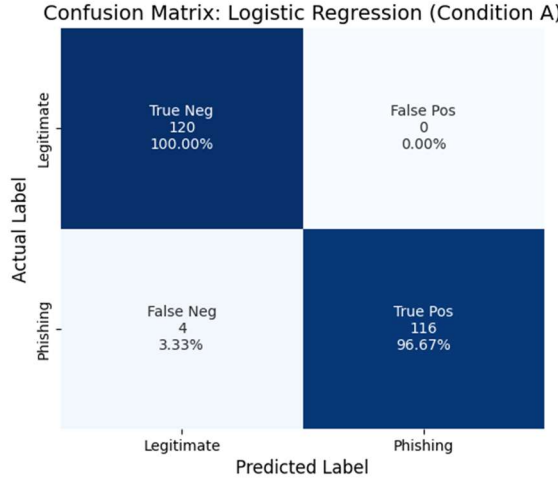

Figure 19 Confusion Matrix: Logistic Regression (Condition A)

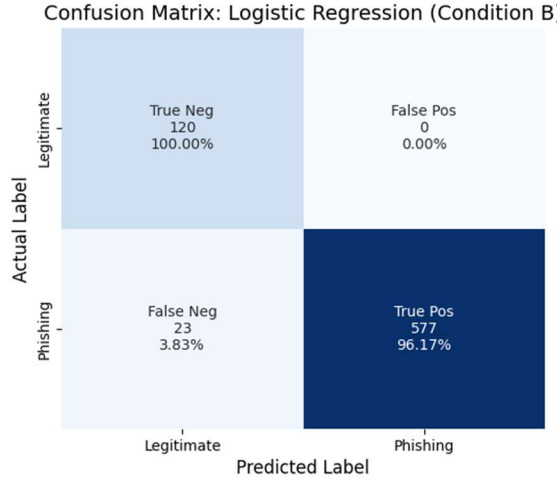

Figure 20 Confusion Matrix: Logistic Regression (Condition B)

For the GPT-5 Zero-Shot model, the primary error mode in Condition A was False Positives (labeling legitimate emails as phishing). In Condition B, the paraphrasing process – which often smoothed the tone and removed "urgent" markers – aligned better with the model's internal heuristics for clear communication, drastically reducing false positives.

However, a trade-off was observed in Recall. While overall F1 scores improved, the Zero-Shot LLM lost its perfect recall (dropping from 1.000 to 0.958), and the Fine-Tuned BERT recall dropped slightly (0.917 to 0.913). This indicates that although the models became more precise, paraphrasing successfully masked certain phishing indicators in a small number of cases, allowing some threats to evade detection.

4.2 Controlled User Study on Explanation Styles

4.2.1 Sample and Data Quality

A total of 55 participants completed the study, of which those who failed attention checks were excluded from the main analysis, in accordance with the study design. The final dataset consisted of participants from two cohorts: psychology students (32) and cybersecurity students (14). This enabled an additional comparison of domain expertise as a moderating factor in user interaction with AI-generated warnings.

4.2.2 Decision Accuracy

Decision accuracy varied across explanation conditions. As shown in Figure 21, the opaque condition yielded the highest mean accuracy ($M \approx 0.87$), followed by the contextual condition ($M \approx 0.84$). In contrast, both rule-based ($M \approx 0.78$) and example-based explanations ($M \approx 0.79$) resulted in lower accuracy.

These findings indicate that providing explanations did not improve user performance. Instead, the presence of explanations may have introduced additional cognitive load, leading to slightly reduced decision accuracy compared to the opaque condition.

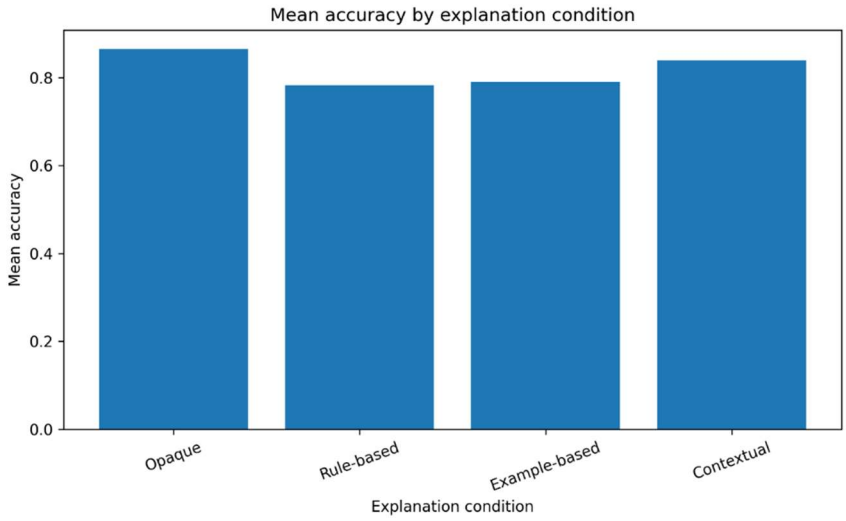

Figure 21 Mean decision accuracy across explanation conditions

4.2.3 Trust Calibration

Trust calibration was assessed through over-reliance and under-reliance metrics. As illustrated in Figure 22, over-reliance was highest in the example-based condition (M ≈ 0.118), followed by the rule-based condition (M ≈ 0.095). The lowest levels of over-reliance were observed in the opaque (M ≈ 0.069) and contextual conditions (M ≈ 0.065).

These results suggest that example-based explanations significantly increase the likelihood of users accepting incorrect AI recommendations. This indicates a tendency toward blind trust when explanations resemble familiar or previously encountered cases.

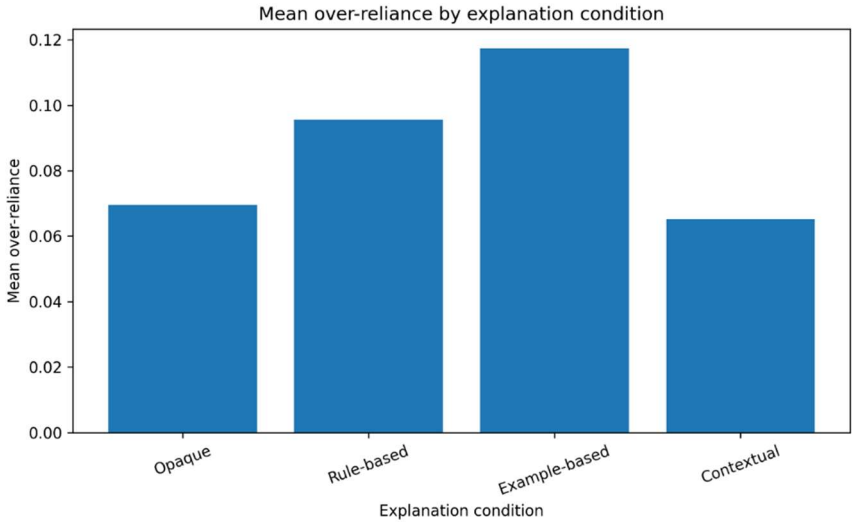

Figure 22 Mean over-reliance across explanation conditions

4.2.4 Confidence

Confidence ratings differed across explanation styles. As shown in Figure 23, contextual explanations produced the highest confidence levels (M ≈ 5.2), followed by opaque (M ≈ 5.1), rule-based (M ≈ 4.95), and example-based explanations (M ≈ 4.85).

However, higher confidence did not correspond to higher accuracy. For instance, although contextual explanations produced the highest confidence, the opaque condition resulted in superior accuracy. This suggests that explanations may influence perceived reliability without improving actual performance, leading to miscalibrated trust.

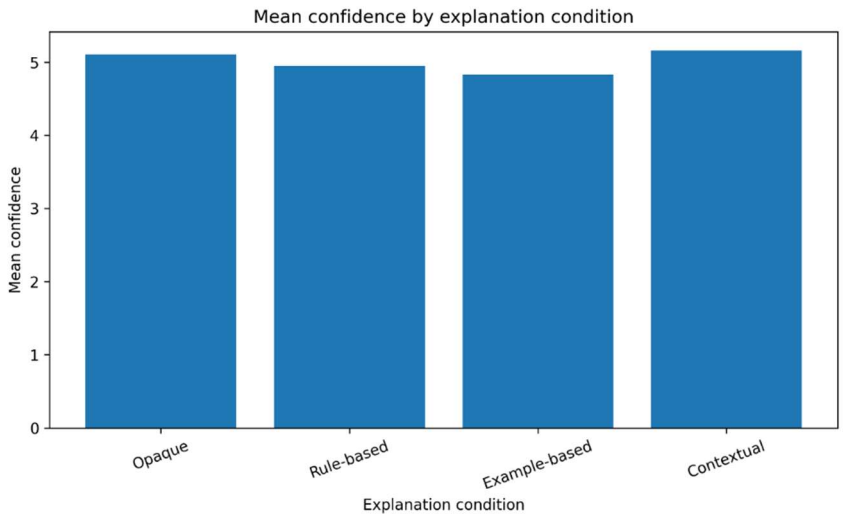

Figure 23 Mean confidence ratings across explanation conditions

4.2.5 Response Time

Response time analysis revealed notable differences between conditions (Figure 24). The opaque condition resulted in the longest mean response time (M ≈ 65,000 ms), whereas example-based explanations produced the fastest responses (M ≈ 36,000 ms). Rule-based and contextual explanations fell between these extremes.

These findings indicate that explanations reduce decision time, likely by providing heuristic cues that simplify decision-making. However, faster decisions did not correspond to improved accuracy or better trust calibration.

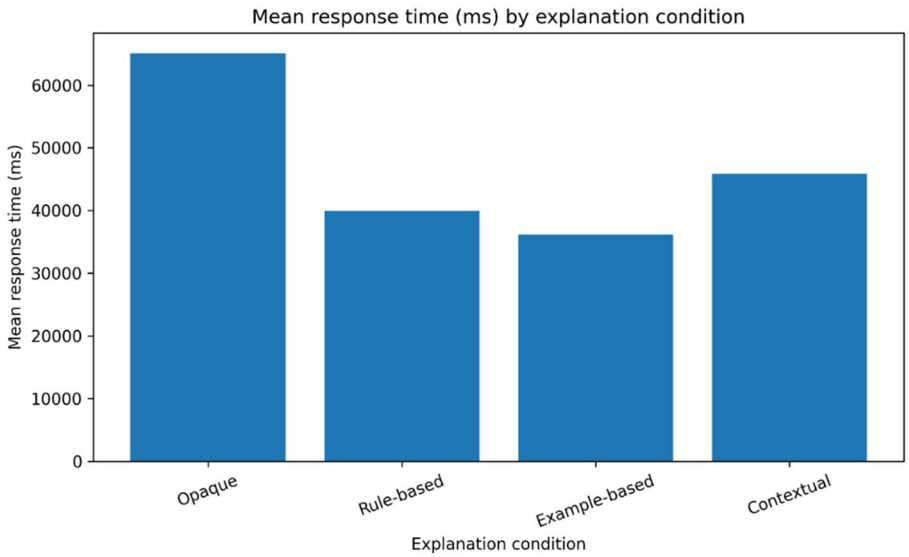

Figure 24 Mean response time across explanation conditions

4.2.6 Group Differences

Significant differences were observed between participant groups. As shown in Figure 25, cybersecurity students consistently achieved higher accuracy across all explanation conditions compared to psychology students.

Furthermore, psychology students exhibited higher levels of over-reliance in all conditions (Figure 26), indicating a greater susceptibility to accepting incorrect AI recommendations. In contrast, cybersecurity students demonstrated more calibrated trust and were less influenced by explanation style.

These results suggest that domain expertise plays a critical role in moderating both decision accuracy and trust calibration when interacting with AI-generated warnings.

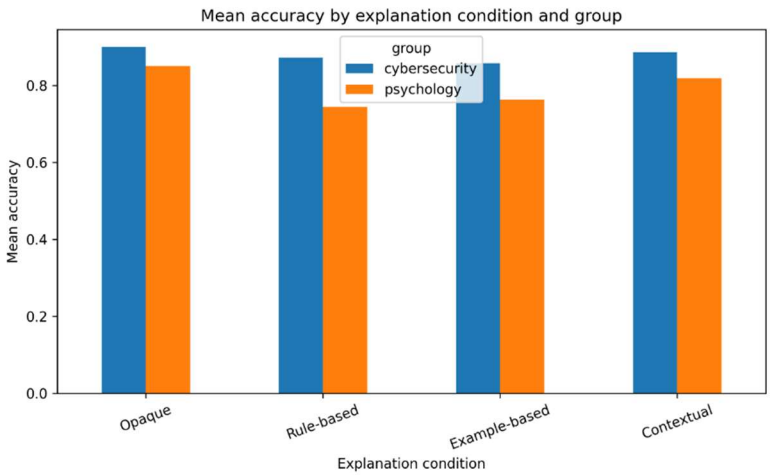

Figure 25 Mean accuracy across explanation conditions by participant group

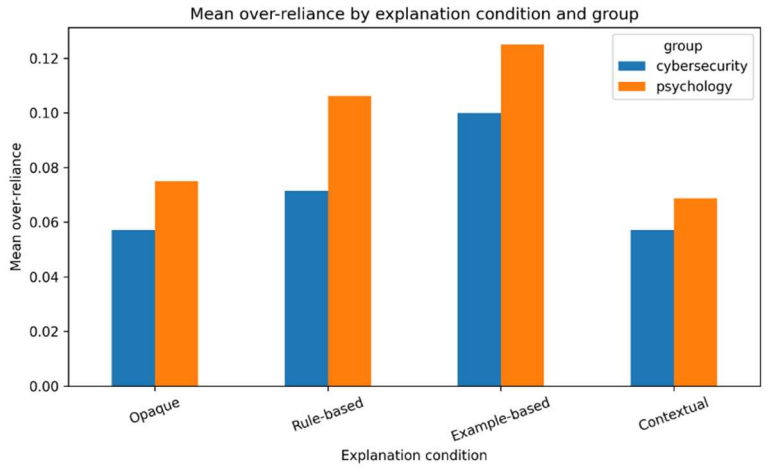

Figure 26 Mean over-reliance across explanation conditions by participant group

4.2.7 Relationship Between Perceived Trust and Behaviour

To further examine the relationship between subjective trust and observed behaviour, correlation analyses were conducted between participants' self-reported trust measures and their actual decision-making performance.

Specifically, correlations were computed between perceived trust variables (e.g., overall trust in the system, perceived clarity and helpfulness of explanations, and self-reported over-reliance) and behavioural metrics (overall accuracy, over-reliance, and under-reliance).

The results revealed a weak relationship between perceived trust and actual performance. Self-reported trust showed only a limited link to decision accuracy, indicating that participants who reported higher trust were not necessarily better at identifying phishing messages.

Similarly, perceived helpfulness and clarity of explanations did not strongly correlate with improved behavioural outcomes. Participants who rated explanations as clear or helpful did not consistently demonstrate lower over-reliance or higher accuracy.

However, a modest positive relationship was observed between self-reported over-reliance and behavioural over-reliance, suggesting that participants had some awareness of their reliance tendencies. Despite this, the strength of the relationship was not sufficient to indicate accurate self-assessment.

Overall, these findings highlight a disconnect between perceived trust and actual behaviour. Users' subjective evaluations of AI explanations do not reliably predict their decision accuracy or their susceptibility to over-reliance. This reinforces the importance of evaluating trust through behavioural metrics rather than relying solely on self-reported measures.

5 Discussion

5.1 Why Paraphrasing Did Not Uniformly Function as Evasion

The Phase 1 findings complicate the assumption, drawn from prior literature, that paraphrasing necessarily operates as a successful evasion strategy. In this study, Condition B was adversarial by design because the paraphrases were intended to obscure suspicious patterns without changing malicious intent. However, the observed outcomes were mixed. This indicates that paraphrasing should be understood as a robustness stressor rather than as guaranteed evasion. Its effectiveness depends on which textual cues a model relies upon, including lexical indicators, semantic consistency, structural regularities, and contextual plausibility.

5.2 Why Opaque Warnings Produced the Highest Accuracy

The finding that the opaque condition achieved the highest decision accuracy suggests that additional explanation is not always beneficial. One possible explanation is cognitive economy: a short warning may have enabled faster judgement without requiring participants to process extra justificatory text. By contrast, more elaborate explanations may have introduced interpretive burden, encouraging participants to engage with the wording of the explanation rather than the message itself. In phishing contexts, more information is therefore not necessarily better information.

5.3 Why Example-Based Explanations Increased Compliance and Over-Reliance

Example-based explanations appear to have increased compliance because they gave the system an impression of concreteness and prior experience. However, this same quality may also have increased over-reliance when the AI output was incorrect. In other words, examples may have made the warning feel more legitimate without necessarily making it more diagnostically useful. This helps explain why example-based explanations can increase trust behaviourally even when they do not improve decision quality.

5.4 Expertise as a Moderating Factor

The stronger performance of cybersecurity participants suggests that domain expertise acted as a moderating factor in trust calibration. Rather than accepting or rejecting the AI output wholesale, more knowledgeable participants may have been better able to integrate their own judgement with system advice. This supports the view that explanation effectiveness is not universal, but depends partly on users' prior knowledge and their ability to judge both the email and the warning.

5.5 Reframing Robustness, Adversariality, and Evasion

The findings of this thesis require a more precise use of the terms robustness, adversarial, and evasion. Robustness should be understood as the stability of classifier performance under meaningful linguistic transformation. Adversarial refers to the intention behind an input transformation, namely to make malicious content harder to detect without changing its underlying purpose. Evasion, however, should be reserved for cases in which that transformation actually reduces detection performance. In the present study, paraphrasing was adversarial by design, but not uniformly evasive in effect. This distinction is important because it prevents adversarial framing from overstating what the data actually show.

5.6 Implications for Human-Centred AI Security Design

Taken together, the results indicate that effective defence against AI-generated social engineering cannot be reduced to either high model accuracy or persuasive explanation design alone. A robust system must perform reliably under linguistic transformation, while a human-centred system must support calibrated rather than blind trust. This points toward AI-assisted security tools evaluated as sociotechnical systems, where technical robustness and user interaction are treated as interdependent rather than separate concerns.

6 Areas for Future Research

While this study provides valuable insights into the interaction between AI-based detection systems and human decision-making, several avenues for future research remain open.

6.1 Adaptive and Personalised Explanation Systems

One of the key findings of this research is that explanation effectiveness varies depending on user expertise. Cybersecurity-trained participants demonstrated more calibrated trust, while non-experts were more susceptible to over-reliance, particularly in the example-based condition.

Future research should explore adaptive explanation systems that dynamically adjust explanation style based on user characteristics, such as expertise, prior performance, or behavioural patterns. Personalised explanations may help mitigate over-reliance in non-expert users while maintaining efficiency for experienced users.

6.2 Dynamic Trust Calibration Mechanisms

This study highlights the importance of trust calibration rather than simply increasing trust in AI systems. However, the current experimental design assumes static explanation styles.

Future work could investigate dynamic trust calibration mechanisms, where the system adapts its level of explanation or warning intensity based on context. For example, explanations could become more detailed or cautionary when the AI system has lower confidence or when the consequences of an error are higher. Such adaptive systems may improve decision-making by encouraging appropriate levels of scepticism.

6.3 Evaluation Under Realistic Operational Conditions

The controlled experimental setup used in this study allowed for systematic comparison of explanation styles but may not fully capture real-world decision-making environments.

Future studies should evaluate user behaviour in more realistic settings, such as simulated email clients integrated into daily workflows or longitudinal studies tracking behaviour over time.

This would allow researchers to examine how factors such as fatigue, habituation, and repeated exposure to warnings influence trust and performance.

6.4 Expansion to Multimodal Social Engineering Attacks

This research focuses on text-based phishing messages. However, modern social engineering attacks increasingly incorporate multiple modalities, including voice (audio deepfakes), video, and hybrid communication channels.

Future research should extend this framework to multimodal attacks, examining how explanation styles influence user trust and decision-making when interacting with more complex and realistic threat scenarios.

6.5 Integration of Human-in-the-Loop Security Systems

Finally, future work could explore the design of human-in-the-loop cybersecurity systems that explicitly combine AI detection with structured human oversight.

Instead of treating users as passive recipients of AI outputs, such systems could encourage active verification, offer decision-support tools, or introduce friction in high-risk scenarios to prevent blind trust. Evaluating how these systems affect both security outcomes and user experience would be a valuable extension of this research.

7 Conclusion

This thesis set out to evaluate the effectiveness of AI-based defences against AI-generated social engineering attacks through a combined technical and human-centred approach. By combining detection-model evaluation with a controlled user study on explanation styles, the research assessed both AI threat detection performance and how effectively human users interpret and act on its outputs.

7.1 Summary of Findings

7.1.1 RQ1: Technical robustness of detection models

RQ1 was addressed by comparing five detection models under original and paraphrased phishing conditions. The results showed that paraphrasing did not uniformly reduce performance. Instead, the effect was model-dependent, indicating that robustness to linguistic transformation varies by architecture and that adversarial intent does not always produce successful evasion.

7.1.2 RQ2: Effect of rule-based explanations

A central contribution of this thesis is the demonstration that AI-based defence against AI-generated social engineering should be evaluated as a sociotechnical system. The technical findings show that linguistic manipulation does not affect all models equally, while the user study shows that more elaborate explanations do not necessarily improve decisions. Together, these findings highlight the importance of aligning model behaviour, explanation design, and user capability.

7.1.3 Impact of example-based explanations

The findings provide strong evidence that example-based explanations increase user compliance with AI warnings. However, this increased compliance is associated with significantly higher levels of over-reliance, particularly when the AI system is incorrect.

Participants exposed to example-based explanations were more likely to accept incorrect classifications, indicating a tendency toward blind trust. This highlights a critical trade-off

between usability and safety in explanation design, where more intuitive explanations may inadvertently increase user vulnerability.

7.2 Bridging Technical Performance and Human Behaviour

A key contribution of this research lies in highlighting the gap between technical system performance and human interaction with AI-based defences.

From a technical perspective, AI detection systems can achieve high levels of accuracy, potentially reaching 90–95% in controlled or real-world deployments. This suggests that such systems can serve as reliable components in cybersecurity infrastructures.

However, the results of the user study demonstrate that high system accuracy does not guarantee effective outcomes in practice. Even when interacting with a generally reliable system, users did not consistently calibrate their trust appropriately. In particular, certain explanation styles encouraged over-reliance, leading users to accept incorrect AI outputs without sufficient scrutiny.

This reveals a fundamental challenge: even highly accurate systems will inevitably produce errors, and inappropriate trust in these systems can lead to critical failures. Therefore, the effectiveness of AI-based defences depends not only on their technical performance but also on how well users understand and appropriately rely on their outputs.

7.3 Implications for AI-Assisted Cybersecurity

The findings of this thesis have several important implications for the design of AI-assisted cybersecurity systems.

First, explanation design must be approached with caution. While explanations are often assumed to improve transparency and trust, this study demonstrates that certain explanation styles can have unintended negative effects. In particular, example-based explanations, while intuitive, can increase the risk of over-reliance.

Second, user expertise plays a significant role in moderating interaction with AI systems. Participants with cybersecurity backgrounds consistently demonstrated higher accuracy and more calibrated trust compared to non-expert users. This suggests that explanation strategies

may need to be tailored to different user groups, rather than applying a one-size-fits-all approach.

Third, the results highlight the importance of trust calibration rather than trust maximisation. The goal of explainable AI should not be to increase user trust indiscriminately, but to ensure that users trust the system appropriately based on its reliability. Users should be able to rely on AI systems in most cases while remaining capable of identifying and correcting errors.

Finally, the observed trade-off between response time and accuracy suggests that reducing cognitive effort may come at the cost of decision quality. Systems that prioritise speed and ease of use must ensure that they do not inadvertently encourage superficial processing or blind acceptance of AI outputs.

7.4 Final Remarks

This thesis demonstrates that effective defence against AI-generated social engineering requires a holistic approach that integrates both technical robustness and human-centred design. While AI systems are becoming increasingly capable of detecting sophisticated attacks, their success ultimately depends on how they are used by human operators.

The results show that explanation design is a critical component of this interaction. Poorly designed explanations can undermine the benefits of advanced detection systems, while well-calibrated interactions can enhance user performance and resilience.

As AI continues to evolve on both the attacker and defender side, future cybersecurity solutions must move beyond purely technical metrics and focus on the broader sociotechnical system. Ensuring that users can appropriately interpret and act on AI-generated warnings will be essential for maintaining security in an increasingly automated threat landscape.

References

[1] D. Gunning and D. W. Aha, 'DARPA's Explainable Artificial Intelligence Program', *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019, doi: 10.1609/aimag.v40i2.2850.

[2] C. Nobles, 'Human Factors Engineering in Explainable AI: Putting People First', *Int. Conf. Cyber Warf. Secur.*, vol. 20, no. 1, pp. 313–322, Mar. 2025, doi: 10.34190/iccws.20.1.3348.

[3] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova, 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations', 2023, *arXiv*. doi: 10.48550/ARXIV.2301.04246.

[4] K. T. Pedersen, L. Pepke, T. Stærmose, M. Papaioannou, G. Choudhary, and N. Dragoni, 'Deepfake-Driven Social Engineering: Threats, Detection Techniques, and Defensive Strategies in Corporate Environments', *J. Cybersecurity Priv.*, vol. 5, no. 2, p. 18, Apr. 2025, doi: 10.3390/jcp5020018.

[5] J. Yu *et al.*, 'The Shadow of Fraud: The Emerging Danger of AI-powered Social Engineering and its Possible Cure', 2024, *arXiv*. doi: 10.48550/ARXIV.2407.15912.

[6] Q. V. Liao and K. R. Varshney, 'Human-Centered Explainable AI (XAI): From Algorithms to User Experiences', 2021, *arXiv*. doi: 10.48550/ARXIV.2110.10790.

[7] J. Kim, H. Maathuis, and D. Sent, 'Human-centered evaluation of explainable AI applications: a systematic review', *Front. Artif. Intell.*, vol. 7, p. 1456486, Oct. 2024, doi: 10.3389/frai.2024.1456486.

[8] A. Mohsin, H. Janicke, A. Ibrahim, I. H. Sarker, and S. Camtepe, 'A Unified Framework for Human AI Collaboration in Security Operations Centers with Trusted Autonomy', Jun. 01, 2025, *arXiv*: arXiv:2505.23397. doi: 10.48550/arXiv.2505.23397.

[9] Z. Aref, S. Wei, and N. B. Mandayam, 'Human-AI Collaboration in Cloud Security: Cognitive Hierarchy-Driven Deep Reinforcement Learning', Apr. 20, 2025, *arXiv*: arXiv:2502.16054. doi: 10.48550/arXiv.2502.16054.

[10] S. Fan, L. Zhang, and X. Yuan, 'Position: Human Factors Reshape Adversarial Analysis in Human-AI Decision-Making Systems', Sep. 25, 2025, *arXiv*: arXiv:2509.21436. doi: 10.48550/arXiv.2509.21436.

[11] B. Green and Y. Chen, 'Algorithm-in-the-Loop Decision Making', *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 09, pp. 13663–13664, Apr. 2020, doi: 10.1609/aaai.v34i09.7115.

[12] C. E. Whyte, 'Machine Expertise in the Loop: Artificial Intelligence Decision-Making Inputs and Cyber Conflict', in *2022 14th International Conference on Cyber Conflict: Keep Moving! (CyCon)*, Tallinn, Estonia: IEEE, May 2022, pp. 135–154. doi: 10.23919/CyCon55549.2022.9811076.

[13] R. J. Tong, 'From Augmentation to Symbiosis: A Review of Human-AI Collaboration Frameworks, Performance, and Perils', 2026, *arXiv*. doi: 10.48550/ARXIV.2601.06030.

[14] J. Holstein and G. Satzger, 'Development of Mental Models in Human-AI Collaboration: A Conceptual Framework', 2025, *arXiv*. doi: 10.48550/ARXIV.2510.08104.

[15] S. Goyal, S. Doddapaneni, M. M. Khapra, and B. Ravindran, 'A Survey of Adversarial Defenses and Robustness in NLP', *ACM Comput. Surv.*, vol. 55, no. 14s, pp. 1–39, Dec. 2023, doi: 10.1145/3593042.

[16] Z. Li, 'Enhancing Human-AI Collaboration through Adaptive Interaction and Explainability', *Proc. AAAIACM Conf. AI Ethics Soc.*, vol. 7, no. 2, pp. 26–27, Jan. 2025, doi: 10.1609/aies.v7i2.31900.

[17] Y. Zhou, E. Chen, M. Pisipati, A. Xiong, and S. Rajtmajer, 'Effect of AI Performance, Risk Perception, and Trust on Human Dependence in Deepfake Detection AI System',

Proc. ACM Hum.-Comput. Interact., vol. 9, no. 7, pp. 1–24, Oct. 2025, doi: 10.1145/3757588.

[18] A. R. Noordeen and M. Bantan, 'How Human Behavior Can Mitigate AI-Generated Cybersecurity Threats', in *Proceedings of the 2025 Computers and People Research Conference*, Waco Texas USA: ACM, May 2025, pp. 1–3. doi: 10.1145/3716489.3728447.

[19] S. Mishler *et al.*, 'Predicting a Malicious Stop Sign: Knowledge, Exposure, Trust in AI', *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 65, no. 1, pp. 347–348, Sep. 2021, doi: 10.1177/1071181321651239.

[20] B. D. Horne, 'Does the Source of a Warning Matter? Examining the Effectiveness of Veracity Warning Labels Across Warners', 2024, *arXiv*. doi: 10.48550/ARXIV.2407.21592.

[21] D. Gamage, D. Sewwandi, M. Zhang, and A. Bandara, 'Labeling Synthetic Content: User Perceptions of Warning Label Designs for AI-generated Content on Social Media', 2025, doi: 10.48550/ARXIV.2503.05711.

[22] C. Wittenberg, Z. Epstein, A. J. Berinsky, and D. G. Rand, 'Labeling AI-Generated Content: Promises, Perils, and Future Directions', *MIT Explor. Gener. AI*, Mar. 2024, doi: 10.21428/e4baedd9.0319e3a6.

[23] D. Shin, A. Koerber, and J. S. Lim, 'Impact of misinformation from generative AI on user information processing: How people understand misinformation from generative AI', *New Media Soc.*, vol. 27, no. 7, pp. 4017–4047, Jul. 2025, doi: 10.1177/14614448241234040.

[24] Y. Shibuya, T. Nakazato, and S. Takagi, 'How do people evaluate the accuracy of video posts when a warning indicates they were generated by AI?', *Int. J. Hum.-Comput. Stud.*, vol. 199, p. 103485, May 2025, doi: 10.1016/j.ijhcs.2025.103485.

[25] C. A. Kuchmaner, 'Mitigating Identity Threats Posed by Artificial Intelligence: The Role of Perceived AI Relatedness', *J. Consum. Behav.*, vol. 25, no. 1, pp. 176–190, Jan. 2026, doi: 10.1002/cb.70066.

[26] V. Autavia and E. D. Madyaymadja, 'User Perception and Awareness of AI-Generated Cyber Threats: A TAM-Based Study', in *2025 2nd Beyond Technology Summit on Informatics International Conference (BTS-I2C)*, Jember, East Java, Indonesia: IEEE, Dec. 2025, pp. 132–137. doi: 10.1109/BTS-I2C67944.2025.11399306.

[27] L. Ai *et al.*, 'Defending Against Social Engineering Attacks in the Age of LLMs', in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 12880–12902. doi: 10.18653/v1/2024.emnlp-main.716.

[28] K. Gonzaga, S. Serra, M. Gomes, and S. Malta, 'AI-Powered Social Engineering: Emerging Attack Vectors, Vulnerabilities, and Multi-Layered Defense Strategies', *Computers*, vol. 15, no. 2, p. 128, Feb. 2026, doi: 10.3390/computers15020128.

[29] H. N. Fakhouri, B. Alhadidi, K. Omar, S. N. Makhadmeh, F. Hamad, and N. Z. Halalsheh, 'AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response', in *2024 2nd International Conference on Cyber Resilience (ICCR)*, Dubai, United Arab Emirates: IEEE, Feb. 2024, pp. 1–8. doi: 10.1109/ICCR61006.2024.10533010.

[30] J. Yu *et al.*, 'The Shadow of Fraud: The Emerging Danger of AI-powered Social Engineering and its Possible Cure', Jul. 22, 2024, *arXiv*: arXiv:2407.15912. doi: 10.48550/arXiv.2407.15912.

[31] V. K. B. Parasaram, 'Phishing Detection 2.0: A Natural Language Processing Approach to Identifying Generative AI-Crafted Social Engineering', *Int. J. Res. Publ. Semin.*, vol. 14, no. 2, pp. 276–284, May 2023, doi: 10.36676/jrps.v14.i2.1680.

[32] S. D. R. Somula and G. Thadi, ‘Detection of AI-Generated Phishing Emails : Comparing The Efficiency Of SVM, Random Forest, CNN And BiLSTM In Detecting AI-Generated Phishing Emails’, Blekinge Institute of Technology, 2025. Accessed: Mar. 02, 2026. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:bth-28339

[33] W. Hao *et al.*, ‘Do Spammers Dream of Electric Sheep? Characterizing the Prevalence of LLM-Generated Malicious Emails’, in *Proceedings of the 2025 ACM Internet Measurement Conference*, Madison WI USA: ACM, Oct. 2025, pp. 192–203. doi: 10.1145/3730567.3732922.

[34] W. Li, Y. Lai, S. Soni, and K. Saha, ‘Emails by LLMs: A Comparison of Language in AI-Generated and Human-Written Emails’, in *Proceedings of the 17th ACM Web Science Conference 2025*, New Brunswick NJ USA: ACM, May 2025, pp. 391–403. doi: 10.1145/3717867.3717872.

[35] C. Opara, P. Modesti, and L. Golightly, ‘Evaluating spam filters and Stylometric Detection of AI-generated phishing emails’, *Expert Syst. Appl.*, vol. 276, p. 127044, Jun. 2025, doi: 10.1016/j.eswa.2025.127044.

[36] F. Chen, T. Wu, V. Nguyen, S. Wang, A. Abuadbba, and C. Rudolph, ‘PEEK: Phishing Evolution Framework for Phishing Generation and Evolving Pattern Analysis using Large Language Models’, 2024, *arXiv*. doi: 10.48550/ARXIV.2411.11389.

[37] J. Francia, D. Hansen, B. Schooley, M. Taylor, S. Murray, and G. Snow, ‘Assessing AI vs Human-Authored Spear Phishing SMS Attacks: An Empirical Study’, 2024, *arXiv*. doi: 10.48550/ARXIV.2406.13049.

[38] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, ‘Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense’, 2023, *arXiv*. doi: 10.48550/ARXIV.2303.13408.

[39] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, ‘Can AI-Generated Text be Reliably Detected?’, 2023, *arXiv*. doi: 10.48550/ARXIV.2303.11156.

[40] Y. Zha, R. Min, and S. Sushmita, ‘PADBen: A Comprehensive Benchmark for Evaluating AI Text Detectors Against Paraphrase Attacks’, 2025, *arXiv*. doi: 10.48550/ARXIV.2511.00416.

[41] Y. Cheng, V. S. Sadasivan, M. Saberi, S. Saha, and S. Feizi, ‘Adversarial Paraphrasing: A Universal Attack for Humanizing AI-Generated Text’, 2025, *arXiv*. doi: 10.48550/ARXIV.2506.07001.

[42] Z. Shi, Y. Wang, F. Yin, X. Chen, K.-W. Chang, and C.-J. Hsieh, ‘Red Teaming Language Model Detectors with Language Models’, *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 174–189, Feb. 2024, doi: 10.1162/tacl_a_00639.

[43] S. Ranganath and A. Ramesh, ‘StealthRL: Reinforcement Learning Paraphrase Attacks for Multi-Detector Evasion of AI-Text Detectors’, 2026, *arXiv*. doi: 10.48550/ARXIV.2602.08934.

[44] I. David and A. Gervais, ‘AuthorMist: Evading AI Text Detectors with Reinforcement Learning’, 2025, *arXiv*. doi: 10.48550/ARXIV.2503.08716.

[45] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, ‘Devising and Detecting Phishing Emails Using Large Language Models’, *IEEE Access*, vol. 12, pp. 42131–42146, 2024, doi: 10.1109/ACCESS.2024.3375882.

[46] F. Heiding, S. Lermen, A. Kao, B. Schneier, and A. Vishwanath, ‘Evaluating Large Language Models’ Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects’, Nov. 30, 2024, *arXiv*: arXiv:2412.00586. doi: 10.48550/arXiv.2412.00586.

[47] P. Åvist, ‘AI-Powered Phishing Outperforms Elite Cybercriminals in 2025 - Hoxhunt’. Accessed: Mar. 02, 2026. [Online]. Available: https://hoxhunt.com/blog/ai-powered-phishing-vs-humans

[48] M. Weinz, N. Zannone, L. Allodi, and G. Apruzzese, ‘The Impact of Emerging Phishing Threats: Assessing Quishing and LLM-generated Phishing Emails against Organizations’, in *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, Hanoi Vietnam: ACM, Aug. 2025, pp. 1550–1566. doi: 10.1145/3708821.3736195.

[49] N. Sniegowski, ‘Evaluating the Effectiveness of LLM-Generated Phishing Campaigns’, Master’s Thesis, Marquette University, 2025. [Online]. Available: https://epublications.marquette.edu/theses_open/857

[50] S. Stecklow and P. McPherson, ‘We wanted to craft a perfect phishing scam. AI bots were happy to help’, *Reuters*, Sep. 15, 2025. Accessed: Mar. 02, 2026. [Online]. Available: https://www.reuters.com/investigates/special-report/ai-chatbots-cyber/

[51] E. Ekekihl, ‘Getting the general public to create phishing emails : A study on the persuasiveness of AI-generated phishing emails versus human methods’, Master’s Thesis, University of Skövde, 2024. Accessed: Mar. 02, 2026. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-24094

[52] M. Madleňák and S. Hubočan, ‘Phishing 2.0: Human Ability to Detect AI-Generated Content’, *Transp. Res. Procedia*, vol. 93, pp. 1125–1132, 2026, doi: 10.1016/j.trpro.2025.12.051.

[53] J. Hazell, ‘Spear Phishing With Large Language Models’, 2023, *arXiv*. doi: 10.48550/ARXIV.2305.06972.

[54] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, ‘From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy’, *IEEE Access*, vol. 11, pp. 80218–80245, 2023, doi: 10.1109/ACCESS.2023.3300381.

[55] S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf, ‘The potential of generative AI for personalized persuasion at scale’, *Sci. Rep.*, vol. 14, no. 1, p. 4692, Feb. 2024, doi: 10.1038/s41598-024-53755-0.

[56] Q. Qi, Y. Luo, Y. Xu, W. Guo, and Y. Fang, ‘SpearBot: Leveraging large language models in a generative-critique framework for spear-phishing email generation’, *Inf. Fusion*, vol. 122, p. 103176, Oct. 2025, doi: 10.1016/j.inffus.2025.103176.

[57] M. Schmitt and I. Flechais, ‘Digital deception: generative artificial intelligence in social engineering and phishing’, *Artif. Intell. Rev.*, vol. 57, no. 12, p. 324, Oct. 2024, doi: 10.1007/s10462-024-10973-2.

[58] R. Mishra, G. Varshney, and S. Singh, ‘Jailbreaking Generative AI: Empowering Novices to Conduct Phishing Attacks’, in *2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S)*, Naples, Italy: IEEE, Jun. 2025, pp. 251–252. doi: 10.1109/DSN-S65789.2025.00022.

[59] P. V. Falade, ‘Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks’, *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 185–198, Oct. 2023, doi: 10.32628/CSEIT2390533.

[60] Z. Lin, J. Cui, X. Liao, and X. Wang, ‘Malla: Demystifying Real-world Large Language Model Integrated Malicious Services’, presented at the 33rd USENIX Security Symposium (USENIX Security 24), 2024, pp. 4693–4710. Accessed: Mar. 03, 2026. [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/lin-zilong

[61] F. Greco, G. Desolda, A. Esposito, and A. Carelli, ‘David versus Goliath: Can Machine Learning Detect LLM-Generated Text? A Case Study in the Detection of Phishing Emails’, in *Proceedings of the 8th Italian Conference on Cyber Security (ITASEC 2024)*, G. D’Angelo, F. Luccio, and F. Palmieri, Eds, in CEUR Workshop Proceedings, vol. 3731. Salerno, Italy: CEUR, Apr. 2024. Accessed: Mar. 02, 2026. [Online]. Available: https://ceur-ws.org/Vol-3731/#paper41

[62] F. Xiong, T. Markchom, Z. Zheng, S. Jung, V. Ojha, and H. Liang, 'NCL-UoR at SemEval-2024 Task 8: Fine-tuning Large Language Models for Multigenerator, Multidomain, and Multilingual Machine-Generated Text Detection', in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 163–169. doi: 10.18653/v1/2024.semeval-1.25.

[63] J. Thapa, G. Chahal, S. V. Gabreanu, and Y. Otoum, 'Phishing Detection in the Gen-AI Era: Quantized LLMs vs Classical Models', 2025, *arXiv*. doi: 10.48550/ARXIV.2507.07406.

[64] S. Tarapiah, L. Abbas, O. Mardawi, S. Atalla, Y. Himeur, and W. Mansoor, 'Evaluating the Effectiveness of Large Language Models (LLMs) Versus Machine Learning (ML) in Identifying and Detecting Phishing Email Attempts', *Algorithms*, vol. 18, no. 10, p. 599, Sep. 2025, doi: 10.3390/a18100599.

[65] J. Yan, 'Large Language Model (LLM) Text Generation Detection: A BERT-BiLSTM Approach with Attention Mechanisms', in *2024 4th International Conference on Electronic Information Engineering and Computer (EIECT)*, Shenzhen, China: IEEE, Nov. 2024, pp. 404–407. doi: 10.1109/EIECT64462.2024.10866488.

[66] M. Hosseinzadeh *et al.*, 'Improving phishing email detection performance through deep learning with adaptive optimization', *Sci. Rep.*, vol. 15, no. 1, p. 36724, Oct. 2025, doi: 10.1038/s41598-025-20668-5.

[67] L. Tian and N. Jiang, 'Research on Detection Methods for Text Generated by Large Language Models Based on Multi-Model Ensemble', *Appl. Comput. Eng.*, vol. 106, no. 1, pp. 59–67, Nov. 2024, doi: 10.54254/2755-2721/106/20241331.

[68] M. A. Wani, M. ElAffendi, and K. A. Shakil, 'AI-Generated Spam Review Detection Framework with Deep Learning Algorithms and Natural Language Processing', *Computers*, vol. 13, no. 10, p. 264, Oct. 2024, doi: 10.3390/computers13100264.

[69] E. Hotoğlu, S. Sen, and B. Can, 'A Comprehensive Analysis of Adversarial Attacks against Spam Filters', 2025, *arXiv*. doi: 10.48550/ARXIV.2505.03831.

[70] D. H. Kulal, C. P. Arannonu, A. Anwar, N. Rastogi, and Q. Niyaz, 'Robust ML-based Detection of Conventional, LLM-Generated, and Adversarial Phishing Emails Using Advanced Text Preprocessing', 2025, *arXiv*. doi: 10.48550/ARXIV.2510.11915.

[71] R. Ardeshirifar, 'Comparing hand-crafted and deep learning approaches for detecting AI-generated text: performance, generalization, and linguistic insights', *AI Ethics*, vol. 5, no. 4, pp. 4197–4209, Aug. 2025, doi: 10.1007/s43681-025-00699-4.

[72] L. Fröhling and A. Zubiaga, 'Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover', *PeerJ Comput. Sci.*, vol. 7, p. e443, Apr. 2021, doi: 10.7717/peerj-cs.443.

[73] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, 'DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature', 2023, *arXiv*. doi: 10.48550/ARXIV.2301.11305.

[74] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang, 'Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature', 2023, *arXiv*. doi: 10.48550/ARXIV.2310.05130.

[75] A. Hans *et al.*, 'Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text', 2024, *arXiv*. doi: 10.48550/ARXIV.2401.12070.

[76] V. Verma, E. Fleisig, N. Tomlin, and D. Klein, 'Ghostbuster: Detecting Text Ghostwritten by Large Language Models', in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 1702–1717. doi: 10.18653/v1/2024.naacl-long.95.

[77] Y. Xu, A. Liu, X. Hu, L. Wen, and H. Xiong, 'Mark Your LLM: Detecting the Misuse of Open-Source Large Language Models via Watermarking', 2025, *arXiv*. doi: 10.48550/ARXIV.2503.04636.
[78] S. Dathathri *et al.*, 'Scalable watermarking for identifying large language model outputs', *Nature*, vol. 634, no. 8035, pp. 818–823, Oct. 2024, doi: 10.1038/s41586-024-08025-4.
[79] K. C. Fraser, H. Dawkins, and S. Kiritchenko, 'Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods', *J. Artif. Intell. Res.*, vol. 82, pp. 2233–2278, Apr. 2025, doi: 10.1613/jair.1.16665.
[80] H. D. S. Gameiro, A. Kucharavy, and L. Dolamic, 'LLM Detectors Still Fall Short of Real World: Case of LLM-Generated Short News-Like Posts', 2024, *arXiv*. doi: 10.48550/ARXIV.2409.03291.
[81] J. Wang, R. Li, J. Yang, and C. Mao, 'RAFT: Realistic Attacks to Fool Text Detectors', 2024, *arXiv*. doi: 10.48550/ARXIV.2410.03658.
[82] A. Pedrotti *et al.*, 'Stress-testing Machine Generated Text Detection: Shifting Language Models Writing Style to Fool Detectors', in *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria: Association for Computational Linguistics, 2025, pp. 3010–3031. doi: 10.18653/v1/2025.findings-acl.156.
[83] J. Wu *et al.*, 'DetectRL: Benchmarking LLM-Generated Text Detection in Real-World Scenarios', 2024, *arXiv*. doi: 10.48550/ARXIV.2410.23746.
[84] X. He, X. Shen, Z. Chen, M. Backes, and Y. Zhang, 'MGTBench: Benchmarking Machine-Generated Text Detection', in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, Salt Lake City UT USA: ACM, Dec. 2024, pp. 2251–2265. doi: 10.1145/3658644.3670344.
[85] E. Crothers, N. Japkowicz, H. Viktor, and P. Branco, 'Adversarial Robustness of Neural-Statistical Features in Detection of Generative Transformers', in *2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy: IEEE, Jul. 2022, pp. 1–8. doi: 10.1109/IJCNN55064.2022.9892269.
[86] M. Bethany, B. Wherry, E. Bethany, N. Vishwamitra, A. Rios, and P. Najafirad, 'Deciphering Textual Authenticity: A Generalized Strategy through the Lens of Large Language Semantics for Detecting Human vs. Machine-Generated Text', presented at the 33rd USENIX Security Symposium (USENIX Security 24), 2024, pp. 5805–5822. Accessed: Mar. 03, 2026. [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/bethany
[87] A. Ismail, C. A. Ezzat, and A. M. El-Korany, 'A Survey on AI-Generated Text Detection: Methods, Challenges, and Future Directions', in *2025 1st Future International Conference on Artificial Intelligence and Cybersecurity (FICAC)*, Cairo, Egypt: IEEE, Nov. 2025, pp. 97–110. doi: 10.1109/FICAC65757.2025.11341850.
[88] A. A. Alethary and A. H. Aliwy, 'A Systematic Review of AI-Generated Text Detection: Approaches, Tools, and Datasets', *Al-Salam J. Eng. Technol.*, vol. 5, no. 1, pp. 145–165, Feb. 2026, doi: 10.55145/ajest.2026.05.01.012.
[89] H. M. Abbas, 'A Novel Approach to Automated Detection of AI-Generated Text', *J. Al-Qadisiyah Comput. Sci. Math.*, vol. 17, no. 1, Mar. 2025, doi: 10.29304/jqcsm.2025.17.11958.
[90] T. Xie, 'Comparing KNN, Logistic Regression, Random Forest and BERT Fine-Tuning for Scam Message Detection', *Finance Econ.*, vol. 1, no. 5, Oct. 2025, doi: 10.61173/e19gh209.
[91] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, 'Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment', *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 05, pp. 8018–8025, Apr. 2020, doi: 10.1609/aaai.v34i05.6311.

[92] S. Garg and G. Ramakrishnan, 'BAE: BERT-based Adversarial Examples for Text Classification', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 6174–6181. doi: 10.18653/v1/2020.emnlp-main.498.

[93] T. Gidatkar, O. Ajao, and M. Shardlow, 'Differential Robustness in Transformer Language Models: Empirical Evaluation Under Adversarial Text Attacks', 2025, *arXiv*. doi: 10.48550/ARXIV.2509.09706.

[94] T. Ali, A. Eleyan, M. Al-Khalidi, and T. Bejaoui, 'GAN-Guarded Fine-Tuning: Enhancing Adversarial Robustness in NLP Models', in *2025 5th IEEE Middle East and North Africa Communications Conference (MENACOMM)*, Byblos, Lebanon: IEEE, Feb. 2025, pp. 1–6. doi: 10.1109/MENACOMM62946.2025.10910951.

[95] X. Dong, A. T. Luu, M. Lin, S. Yan, and H. Zhang, 'How Should Pre-Trained Language Models Be Fine-Tuned Towards Adversarial Robustness?', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 4356–4369. Accessed: Mar. 03, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/22b1f2e0983160db6f7bb9f62f4dbb39-Abstract.html

[96] D. Hee Lee and B. Jang, 'Enhancing Machine-Generated Text Detection: Adversarial Fine-Tuning of Pre-Trained Language Models', *IEEE Access*, vol. 12, pp. 65333–65340, 2024, doi: 10.1109/ACCESS.2024.3396820.

[97] S. Fredriksson Karvelas and K. Frohm Krischel, 'Go Phish - Detecting Phishing Attacks with Prompt Engineered Large Language Models as Classification Tools', Stockholm University, 2025. Accessed: Mar. 03, 2026. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-245927

[98] E. Eilertsen, V. Mavroeidis, and G. Grov, 'LLM-Powered Intent-Based Categorization of Phishing Emails', 2025, *arXiv*. doi: 10.48550/ARXIV.2506.14337.

[99] J. Hua, P. Wang, and P. Lutchkus, 'HOW EFFECTIVE ARE LARGE LANGUAGE MODELS IN DETECTING PHISHING EMAILS?', *Issues Inf. Syst.*, 2024, doi: 10.48009/3_iis_2024_125.

[100] K. Afane, W. Wei, Y. Mao, J. Farooq, and J. Chen, 'Next-Generation Phishing: How LLM Agents Empower Cyber Attackers', in *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA: IEEE, Dec. 2024, pp. 2558–2567. doi: 10.1109/BigData62323.2024.10825018.

[101] F. Trad and A. Chehab, 'Prompt Engineering or Fine-Tuning? A Case Study on Phishing Detection with Large Language Models', *Mach. Learn. Knowl. Extr.*, vol. 6, no. 1, pp. 367–384, Feb. 2024, doi: 10.3390/make6010018.

[102] A. H. Nasution, W. Monika, A. Onan, and Y. Murakami, 'Benchmarking 21 Open-Source Large Language Models for Phishing Link Detection with Prompt Engineering', *Information*, vol. 16, no. 5, p. 366, Apr. 2025, doi: 10.3390/info16050366.

[103] H. Fang *et al.*, 'Your Language Model Can Secretly Write Like Humans: Contrastive Paraphrase Attacks on LLM-Generated Text Detectors', in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, 2025, pp. 8596–8613. doi: 10.18653/v1/2025.emnlp-main.433.

[104] A. Samadi and A. Sullivan, 'Evaluating Text Classification Robustness to Part-of-Speech Adversarial Examples', 2024, *arXiv*. doi: 10.48550/ARXIV.2408.08374.

[105] A. Fazla, L. Krauter, D. Guzman Piedrahita, and A. Michail, 'Robustness of Misinformation Classification Systems to Adversarial Examples Through BeamAttack', in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, vol. 16089, J. Carrillo-de-Albornoz, A. García Seco De Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi,

P. Rosso, D. Spina, G. Faggioli, and N. Ferro, Eds, in Lecture Notes in Computer Science, vol. 16089. , Cham: Springer Nature Switzerland, 2026, pp. 89–116. doi: 10.1007/978-3-032-04354-2_7.

[106] A. Nighojkar and J. Licato, 'Improving Paraphrase Detection with the Adversarial Paraphrasing Task', in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, 2021, pp. 7106–7116. doi: 10.18653/v1/2021.acl-long.552.

[107] T. Roth, I. J. Unanue, A. Abuadbba, and M. Piccardi, 'A Constraint-Enforcing Reward for Adversarial Attacks on Text Classifiers', 2024, *arXiv*. doi: 10.48550/ARXIV.2405.11904.

[108] D. Cuong, D. Le, and T. Le, 'A Curious Case of Searching for the Correlation between Training Data and Adversarial Robustness of Transformer Textual Models', in *Findings of the Association for Computational Linguistics ACL 2024*, Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, 2024, pp. 13475–13491. doi: 10.18653/v1/2024.findings-acl.800.

[109] V. Yadav, Z. Tang, and V. Srinivasan, 'PAG-LLM: Paraphrase and Aggregate with Large Language Models for Minimizing Intent Classification Errors', in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Washington DC USA: ACM, Jul. 2024, pp. 2569–2573. doi: 10.1145/3626772.3657959.

[110] R. Xiao, K. Zhang, Y. Zhang, Q. Wang, S. Hou, and L. Liu, 'Study on relationship between adversarial texts and language errors: a human-computer interaction perspective', *Behav. Inf. Technol.*, vol. 44, no. 9, pp. 2053–2068, May 2025, doi: 10.1080/0144929X.2024.2396437.

[111] M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, 'How the different explanation classes impact trust calibration: The case of clinical decision support systems', *Int. J. Hum.-Comput. Stud.*, vol. 169, p. 102941, Jan. 2023, doi: 10.1016/j.ijhcs.2022.102941.

[112] X. Wang and M. Yin, 'Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons', *ACM Trans. Interact. Intell. Syst.*, vol. 12, no. 4, pp. 1–36, Dec. 2022, doi: 10.1145/3519266.

[113] O. Rezaeian, A. E. Bayrak, and O. Asan, 'Explainability and AI Confidence in Clinical Decision Support Systems: Effects on Trust, Diagnostic Performance, and Cognitive Load in Breast Cancer Care', *Int. J. Human–Computer Interact.*, pp. 1–21, Aug. 2025, doi: 10.1080/10447318.2025.2539458.

[114] P. Wang and H. Ding, 'The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance', *Inf. Process. Manag.*, vol. 61, no. 4, p. 103732, Jul. 2024, doi: 10.1016/j.ipm.2024.103732.

[115] S. B. H. Pias, A. Freel, T. Trammel, T. Akter, D. Williamson, and A. Kapadia, 'The Drawback of Insight: Detailed Explanations Can Reduce Agreement with XAI', 2024, *arXiv*. doi: 10.48550/ARXIV.2404.19629.

[116] A. Papenmeier, D. Kern, G. Englebienne, and C. Seifert, 'It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI', *ACM Trans. Comput.-Hum. Interact.*, vol. 29, no. 4, pp. 1–33, Aug. 2022, doi: 10.1145/3495013.

[117] M. Abbaspour Onari, 'From Explanation to Trust: Modeling and Measuring Trust in Explainable Decision Support', Phd Thesis 1 (Research TU/e / Graduation TU/e), Eindhoven University of Technology, Eindhoven, 2025. [Online]. Available: https://research.tue.nl/files/369143413/20251104_Abbaspour_Onari_mail.pdf

[118] A. A. Tutul, E. H. Nirjhar, and T. Chaspari, ‘Investigating Trust in Human-AI Collaboration for a Speech-Based Data Analytics Task’, *Int. J. Human–Computer Interact.*, pp. 1–19, Mar. 2024, doi: 10.1080/10447318.2024.2328910.
[119] C. Panigutti *et al.*, ‘Co-design of Human-centered, Explainable AI for Clinical Decision Support’, *ACM Trans. Interact. Intell. Syst.*, vol. 13, no. 4, pp. 1–35, Dec. 2023, doi: 10.1145/3587271.
[120] E. A. Katsarakes, ‘Making Explanations Make Sense: XAI for SMiShing Detection’, Master’s Thesis, Old Dominion University, 2025. doi: 10.25777/hq9j-5k20.
[121] K. Singh, P. Aggarwal, P. Rajivan, and C. Gonzalez, ‘Cognitive elements of learning and discriminability in anti-phishing training’, *Comput. Secur.*, vol. 127, p. 103105, Apr. 2023, doi: 10.1016/j.cose.2023.103105.
[122] E. A. Cranford, C. Lebiere, P. Rajivan, P. Aggarwal, and C. Gonzalez, ‘Modeling cognitive dynamics in end-user response to phishing emails’, in *Proceedings of the 17th ICCM*, Applied Cognitive Science Lab State College, PA, 2019. [Online]. Available: https://iccm-conference.neocities.org/2019/proceedings/papers/ICCM2019_paper_57.pdf
[123] A. Adhikari, D. M. J. Tax, R. Satta, and M. Faeth, ‘LEAFAGE: Example-based and Feature importance-based Explanations for Black-box ML models’, in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, New Orleans, LA, USA: IEEE, Jun. 2019, pp. 1–7. doi: 10.1109/FUZZ-IEEE.2019.8858846.
[124] T. Saka, K. Vakali, A. D. G. Jenkins, N. Kokciyan, and K. Vaniea, ‘Judging Phishing Under Uncertainty: How Do Users Handle Inaccurate Automated Advice?’, in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, Apr. 2025, pp. 1–18. doi: 10.1145/3706598.3714267.
[125] G. Desolda, J. Aneke, C. Ardito, R. Lanzilotti, and M. F. Costabile, ‘Explanations in warning dialogs to help users defend against phishing attacks’, *Int. J. Hum.-Comput. Stud.*, vol. 176, p. 103056, Aug. 2023, doi: 10.1016/j.ijhcs.2023.103056.
[126] P. Buono, G. Desolda, F. Greco, and A. Piccinno, ‘Let warnings interrupt the interaction and explain: designing and evaluating phishing email warnings’, in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 2023, pp. 1–6. doi: 10.1145/3544549.3585802.
[127] R. Hunter, R. Moulange, J. Bernardi, and M. Stein, ‘Monitoring Human Dependence On AI Systems With Reliance Drills’, 2024, *arXiv*. doi: 10.48550/ARXIV.2409.14055.
[128] A. Klingbeil, C. Grützner, and P. Schreck, ‘Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI’, *Comput. Hum. Behav.*, vol. 160, p. 108352, Nov. 2024, doi: 10.1016/j.chb.2024.108352.
[129] R. A. Hagen, L. Øverlier, and K. Helkala, ‘Human Factors in AI-Driven Cybersecurity: Cognitive Biases and Trust Issues’, *Digit. Threats Res. Pract.*, vol. 6, no. 4, pp. 1–20, Dec. 2025, doi: 10.1145/3759260.
[130] J. Tilbury, S. Flowerday, G. Bott, Y. T. Chua, E. Olson, and B. Foltz, ‘Human-Factor Vulnerabilities of Automation in SOCs: A Mixed-Methods Multigroup Analysis’, *Comput. Secur.*, p. 104871, Mar. 2026, doi: 10.1016/j.cose.2026.104871.
[131] R. Fok and D. S. Weld, ‘In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making’, *AI Mag.*, vol. 45, no. 3, pp. 317–332, Sep. 2024, doi: 10.1002/aaai.12182.
[132] P. Spitzer, J. Holstein, K. Morrison, K. Holstein, G. Satzger, and N. Kühl, ‘Don’t Be Fooled: The Misinformation Effect of Explanations in Human–AI Collaboration’, *Int. J. Human–Computer Interact.*, pp. 1–29, Nov. 2025, doi: 10.1080/10447318.2025.2574511.
[133] T. M. Peters and R. W. Visser, ‘The Importance of Distrust in AI’, in *Explainable Artificial Intelligence*, vol. 1903, L. Longo, Ed., in Communications in Computer and

Information Science, vol. 1903. , Cham: Springer Nature Switzerland, 2023, pp. 301–317. doi: 10.1007/978-3-031-44070-0_15.
[134] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, 'Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application', in *2023 ACM Conference on Fairness, Accountability, and Transparency*, Chicago IL USA: ACM, Jun. 2023, pp. 77–88. doi: 10.1145/3593013.3593978.
[135] R. Serafini, A. Yardim, and A. Naiakshina, 'Exploring the Impact of Intervention Methods on Developers' Security Behavior in a Manipulated ChatGPT Study', in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, Apr. 2025, pp. 1–26. doi: 10.1145/3706598.3713989.
[136] B. Lit, E. Crowder, D. Vogel, and H. Khan, '"I know it's not right, but that's what it said to do": Investigating Trust in AI Chatbots for Cybersecurity Policy', Oct. 10, 2025, *arXiv*: arXiv:2510.08917. doi: 10.48550/arXiv.2510.08917.
[137] D. A. Mathew, 'Human–AI Collaboration in Security Operations: Measuring Alert Trust, Automation Bias, and Analyst Upskilling in AI-Augmented SOC Environments', *Int. J. Comput. Technol. Electron. Commun.*, vol. 8, no. 5, pp. 11375–11380, Oct. 2025, doi: 10.15680/IJCTECE.2025.0805011.
[138] D. Humphreys, A. Koay, D. Desmond, and E. Mealy, 'AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business', *AI Ethics*, vol. 4, no. 3, pp. 791–804, Aug. 2024, doi: 10.1007/s43681-024-00443-4.
[139] I. J. Chowdhury and M. A. Y. Tanvir, 'Calibrated Trust in AI for Security Operations: A Conceptual Framework for Analyst–AI Collaboration', Dec. 23, 2025, *Computer Science and Mathematics*. doi: 10.20944/preprints202512.2112.v1.
[140] D. Ganye and K. Smith, 'Examining the effects of cognitive load on information systems security policy compliance', *Internet Res.*, vol. 35, no. 1, pp. 380–418, Jan. 2025, doi: 10.1108/INTR-04-2023-0329.
[141] J. L. Jenkins, B. B. Anderson, A. Vance, C. B. Kirwan, and D. Eargle, 'More Harm Than Good? How Messages That Interrupt Can Make Us Vulnerable', *Inf. Syst. Res.*, vol. 27, no. 4, pp. 880–896, Dec. 2016, doi: 10.1287/isre.2016.0644.
[142] B. B. Anderson, A. Vance, C. B. Kirwan, D. Eargle, and J. L. Jenkins, 'How users perceive and respond to security messages: a NeuroIS research agenda and empirical study', *Eur. J. Inf. Syst.*, vol. 25, no. 4, pp. 364–390, Jul. 2016, doi: 10.1057/ejis.2015.21.
[143] P. L. Morgan, E. J. Williams, N. A. Zook, and G. Christopher, 'Exploring Older Adult Susceptibility to Fraudulent Computer Pop-Up Interruptions', in *Advances in Human Factors in Cybersecurity*, vol. 782, T. Z. Ahram and D. Nicholson, Eds, in Advances in Intelligent Systems and Computing, vol. 782. , Cham: Springer International Publishing, 2019, pp. 56–68. doi: 10.1007/978-3-319-94782-2_6.
[144] M. N. H. Chowdhury, 'Understanding the impact of time pressure on human cybersecurity behavior', thesis, University of Newcastle Australia, 2025. Accessed: Mar. 11, 2026. [Online]. Available: https://openresearch.newcastle.edu.au/articles/thesis/Understanding_the_impact_of_time_pressure_on_human_cybersecurity_behavior/29013368/1
[145] I. Malik, 'Decision Fatigue and Cybersecurity Behaviors: A Qualitative Study of University Students', *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 58s, pp. 441–456, Aug. 2025, doi: 10.52783/jisem.v10i58s.12613.
[146] P. A. Hancock, A. D. Kaplan, K. R. MacArthur, and J. L. Szalma, 'How effective are warnings? A meta-analysis', *Saf. Sci.*, vol. 130, p. 104876, Oct. 2020, doi: 10.1016/j.ssci.2020.104876.

[147] K. S. Jones, N. R. Lodinger, B. P. Widlus, A. S. Namin, and R. Hewett, 'Do Warning Message Design Recommendations Address Why Non-Experts Do Not Protect Themselves from Cybersecurity Threats? A Review', *Int. J. Human–Computer Interact.*, vol. 37, no. 18, pp. 1709–1719, Nov. 2021, doi: 10.1080/10447318.2021.1908691.

[148] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri, 'Bridging the Gap in Computer Security Warnings: A Mental Model Approach', *IEEE Secur. Priv. Mag.*, vol. 9, no. 2, pp. 18–26, Mar. 2011, doi: 10.1109/MSP.2010.198.

[149] A. Vance, D. Eargle, D. Eggett, D. W. Straub, and K. Ouimet, 'Do Security Fear Appeals Work When They Interrupt Tasks? A Multi-Method Examination of Password Strength', *MIS Q.*, vol. 46, no. 3, pp. 1721–1738, Sep. 2022, doi: 10.25300/MISQ/2022/15511.

[150] B. B. Anderson, A. Vance, C. B. Kirwan, J. L. Jenkins, and D. Eargle, 'From Warning to Wallpaper: Why the Brain Habituates to Security Warnings and What Can Be Done About It', *J. Manag. Inf. Syst.*, vol. 33, no. 3, pp. 713–743, Jul. 2016, doi: 10.1080/07421222.2016.1243947.

[151] S. Pham, G. Lenzini, and D. Pöhn, 'How Johnny Experiences Phishing Warnings: A Qualitative Study Investigating the Impact of Design Decisions on the User', *IEEE Access*, vol. 13, pp. 211960–211980, 2025, doi: 10.1109/ACCESS.2025.3642897.

[152] B. Maram, S. Balija, M. C. Sekhar, L. G. Rao, M. S. S. Raj, and V. E. Sathishkumar, 'XAI-Driven Security Systems: Enhancing Trust and Clarity in Automated Threat Detection and Response', in *2025 2nd International Conference on Intelligent Systems for Cybersecurity (ISCS)*, Gurugram, India: IEEE, Nov. 2025, pp. 1–7. doi: 10.1109/ISCS69371.2025.11386435.

[153] D. Honeycutt, M. Nourani, and E. Ragan, 'Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy', *Proc. AAAI Conf. Hum. Comput. Crowdsourcing*, vol. 8, pp. 63–72, Oct. 2020, doi: 10.1609/hcomp.v8i1.7464.

[154] J. Chen, S. Mishler, and B. Hu, 'Automation Error Type and Methods of Communicating Automation Reliability Affect Trust and Performance: An Empirical Study in the Cyber Domain', *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 5, pp. 463–473, Oct. 2021, doi: 10.1109/THMS.2021.3051137.

[155] N. Van Berkel, J. Opie, O. F. Ahmad, L. Lovat, D. Stoyanov, and A. Blandford, 'Initial Responses to False Positives in AI-Supported Continuous Interactions: A Colonoscopy Case Study', *ACM Trans. Interact. Intell. Syst.*, vol. 12, no. 1, pp. 1–18, Mar. 2022, doi: 10.1145/3480247.

[156] J. Y. Kim, C. Lester, and X. J. Yang, 'Beyond Binary Decisions: Evaluating the Effects of AI Error Type on Trust and Performance in AI-Assisted Tasks', *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 67, no. 10, pp. 1062–1083, Oct. 2025, doi: 10.1177/00187208251326795.

[157] R. Kocielnik, S. Amershi, and P. N. Bennett, 'Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–14. doi: 10.1145/3290605.3300641.

[158] S. Egelman, L. F. Cranor, and J. Hong, 'You've been warned: an empirical study of the effectiveness of web browser phishing warnings', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Florence Italy: ACM, Apr. 2008, pp. 1065–1074. doi: 10.1145/1357054.1357219.

[159] K. Krol, M. Moroz, and M. A. Sasse, 'Don't work. Can't work? Why it's time to rethink security warnings', in *2012 7th International Conference on Risks and Security of Internet*

and Systems (CRiSIS), Cork, Ireland: IEEE, Oct. 2012, pp. 1–8. doi: 10.1109/CRISIS.2012.6378951.

[160] D. Modic and R. Anderson, 'Reading this may harm your computer: The psychology of malware warnings', *Comput. Hum. Behav.*, vol. 41, pp. 71–79, Dec. 2014, doi: 10.1016/j.chb.2014.09.014.

[161] G. Bal, 'Explicitness of Consequence Information in Privacy Warnings: Experimentally Investigating the Effects on Perceived Risk, Trust, and Privacy Information Quality', in *35th International Conference on Information Systems (ICIS 2014)*, Auckland, New Zealand: ICIS, 2014. Accessed: Mar. 11, 2026. [Online]. Available: https://www.researchgate.net/publication/268122722_Explicitness_of_Consequence_Information_in_Privacy_Warnings_Experimentally_Investigating_the_Effects_on_Perceived_Risk_Trust_and_Privacy_Information_Quality

[162] S. Egelman, D. Molnár, N. Christin, A. Acquisti, C. Herley, and S. Krishnamurthi, 'Please Continue to Hold: An Empirical Study of User Tolerance of Security Delays', presented at the Workshop on the Economics of Information Security, 2010. Accessed: Mar. 11, 2026. [Online]. Available: https://cs.brown.edu/~sk/Publications/Papers/Published/emcahk-pl-cont-hold-sec-delay/

[163] C. Bravo-Lillo, L. F. Cranor, J. Downs, S. Komanduri, and M. Sleeper, 'Improving Computer Security Dialogs', in *Human-Computer Interaction – INTERACT 2011*, vol. 6949, P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, and M. Winckler, Eds, in Lecture Notes in Computer Science, vol. 6949. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 18–35. doi: 10.1007/978-3-642-23768-3_2.

[164] R. Epstein and V. R. Zankich, 'The surprising power of a click requirement: How click requirements and warnings affect users' willingness to disclose personal information', *PLOS ONE*, vol. 17, no. 2, p. e0263097, Feb. 2022, doi: 10.1371/journal.pone.0263097.

[165] B. Kaiser, J. Wei, E. Lucherini, K. Lee, J. N. Matias, and J. Mayer, 'Adapting Security Warnings to Counter Online Disinformation', presented at the 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 1163–1180. Accessed: Mar. 11, 2026. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/kaiser

[166] G. Desolda, F. Greco, and L. Vigano, 'APOLLO: A GPT-based tool to detect phishing emails and generate explanations that warn users', *Proc. ACM Hum.-Comput. Interact.*, vol. 9, no. 4, pp. 1–33, Jun. 2025, doi: 10.1145/3733049.

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I, Farkas Pongrácz

1 grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis „Evaluating AI-Based Defences Against AI-Generated Social Engineering: A Technical and Human-Centred Analysis“, supervised by Dr. Adrian Nicholas Venables

1.1 to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

1.2 to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2 I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3 I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

[1] The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix A – Phase 1 Data

Phase 1 results can be found under this repository: https://github.com/pongraczfarkas/phishing-study-phase1

Appendix B – Phase 2 Data

Phase 2 results can be found under this repository: https://github.com/pongraczfarkas/phishing-study-phase2