Securing the Weakest Link: Exploring Psychological Vulnerabilities in Phishing Emails with LLMs

Faithful C. Onwuegbuche, Rajesh Titung, Esa Rantanen, Anca D. Jucrut, Cecilia Alm, Liliana Pasquale

1 PROBLEM STATEMENT:

Despite efforts at curbing phishing, individuals and organisations still fall victim to phishing attacks [1,2].

Research Objective: we addressed how attackers exploit susceptibility factors such as fear and greed, referred to as “psychological vulnerabilities” (PV) in phishing emails.

METHODOLOGY:

Figure 1: System Design

- Figure 1: shows the system design used in the study. We proposed a taxonomy of PV (fear, urgency, greed, curiosity, trust, compassion) inspired by previous theories [3,4] on human susceptibility to scams and fraud.
- Using a dataset targeting 6 universities [5], we assessed how LLMs (GPT 4, Llama2, and GeminiPro) automatically detect vulnerabilities and valence.
- We evaluated the performance of LLMs to human annotations using reliability statistics and analysed LLM hallucinations.

2 RESULTS: VULNERABILITY AND VALENCE ANALYSIS

Figure 2: Vulnerability Distributions for Email Subjects using all Labels.

- Figure 2 – Attackers commonly exploit curiosity and urgency in email subjects.

Figure 3: Vulnerability Distributions for Email Bodies using all Labels.

- Figure 3 – Attackers commonly exploit urgency, trust, and fear in email bodies.

3 RESULTS: PAIRED ACCOMPANIED VULNERABILITIES:

Figure 4: Subject Pairing Venn Diagrams.

- Figures 7 & 8: Attackers use a single vulnerability for email subjects and multiple for the body. Urgency-Fear pair is prevalent in subjects, while Urgency-Trust pair is more exploited in the body.

Figure 5: Body Pairing Venn Diagrams.

4 RESULTS: VALENCE-AROUSAL MAPPING:

Figure 6: Subject Valence-Arousal Mapping.

- Figures 9 & 10: Urgency & Fear exhibit the highest arousal levels in both email subjects and bodies, while Trust & Curiosity show lower arousal.

Figure 7: Body Valence-Arousal Mapping.

Table 2: Inter-rater reliability analysis for LLMs and Humans.

- Table 2 – All LLMs show agreement with human annotators.
- GPT-4 outperforms GeminiPro and Llama2 overall, with higher Cohen’s Kappa and Krippendorf’s Alpha values.

Future Work:

- Expand the study to incorporate datasets from diverse sectors beyond universities.
- Evaluate whether identifying PV in phishing emails can improve the performance of automated machine learning phishing detection approaches.

5 RESULTS: LLMS PERFORMANCE VS HUMAN

Table 2 – All LLMs show agreement with human annotators.

- GPT-4 outperforms GeminiPro and Llama2 overall, with higher Cohen’s Kappa and Krippendorf’s Alpha values.

Future Work:

- Expand the study to incorporate datasets from diverse sectors beyond universities.
- Evaluate whether identifying PV in phishing emails can improve the performance of automated machine learning phishing detection approaches.

References