Software for a better world

University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

AWARE-AI NSF Research Traineeship Program

A World Leading SFI Research Centre

Science Foundation Ireland For what's next

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

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

Connect with Faithful

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:

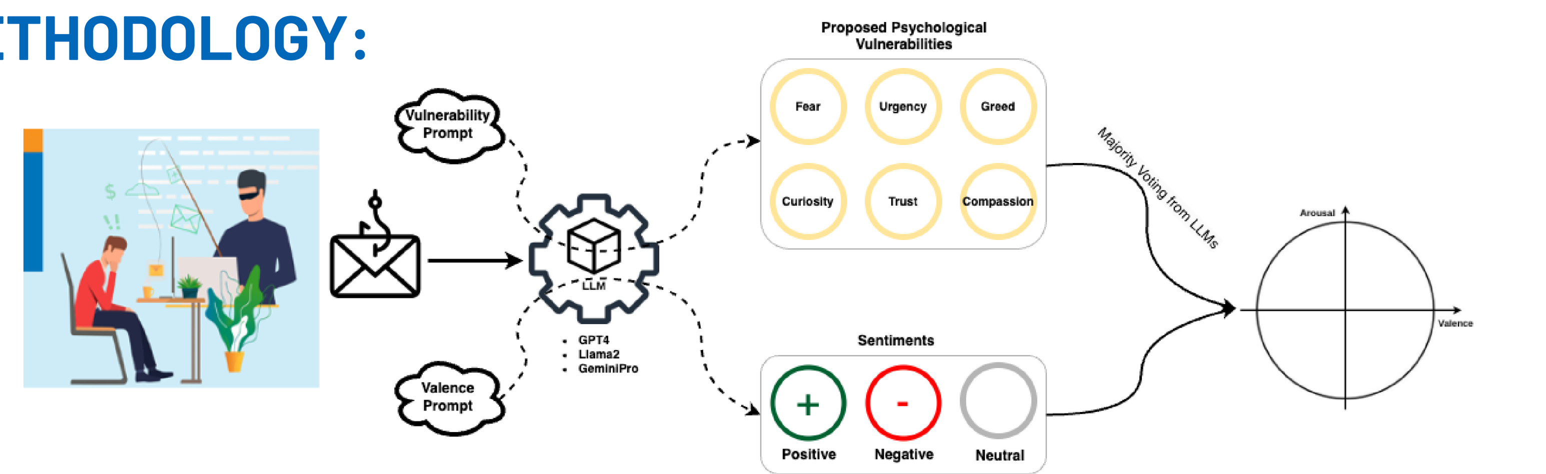

Figure 1: System Design

- 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

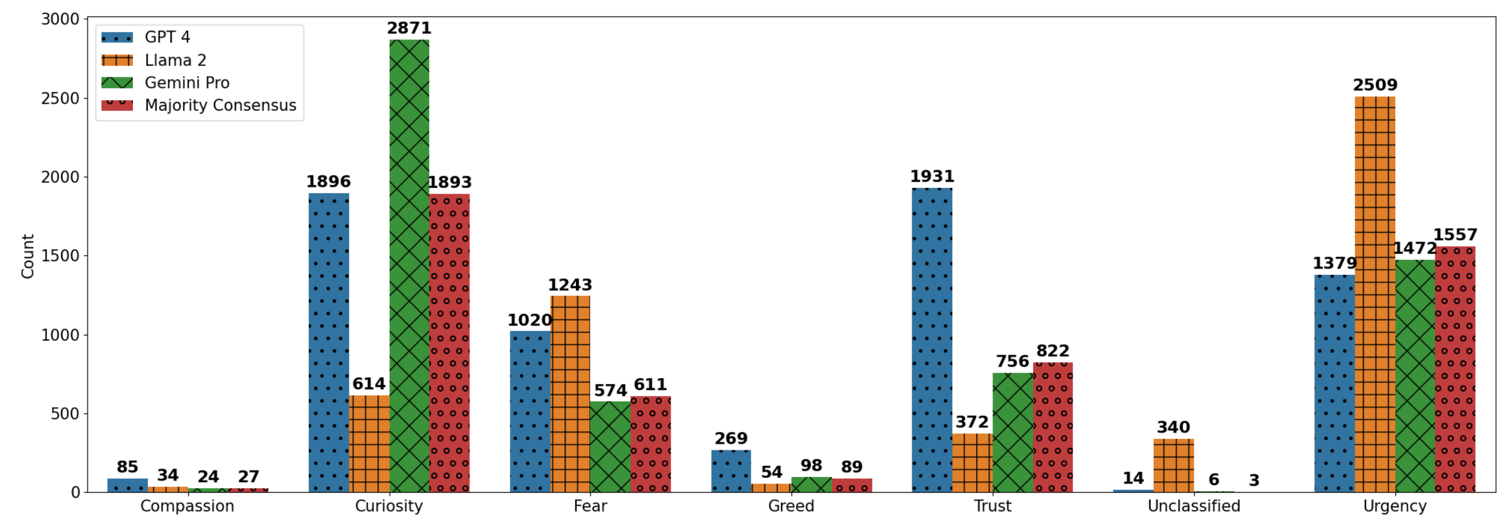

Figure 2: Vulnerability Distributions for Email Subjects using all Labels.

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

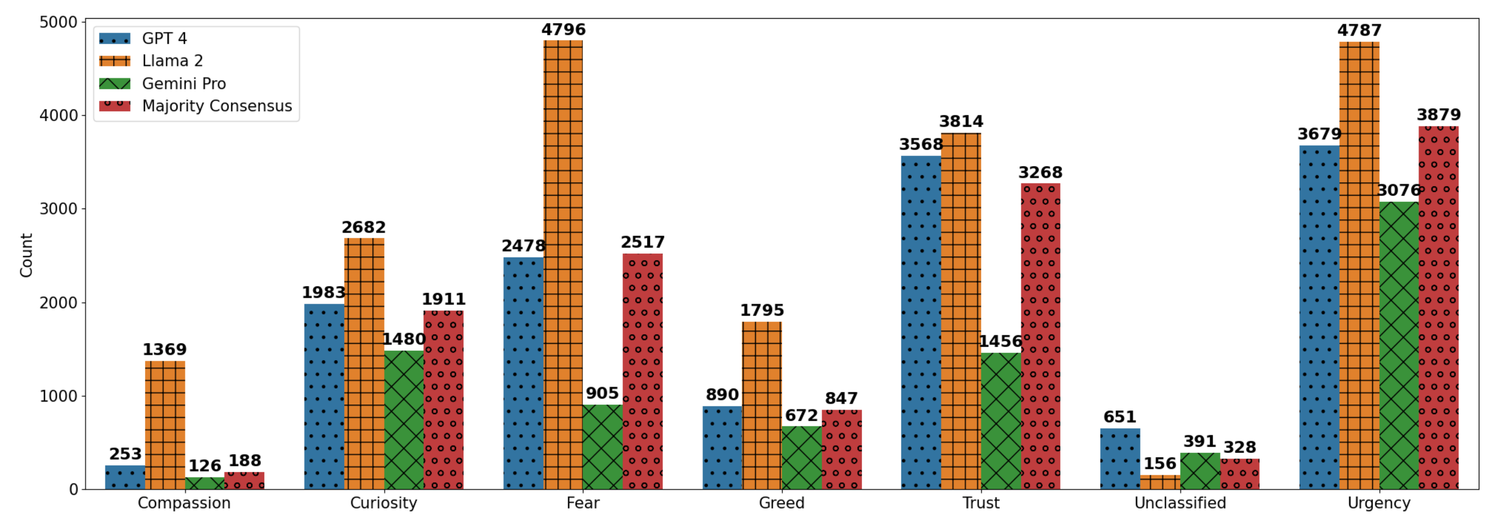

Figure 3: Vulnerability Distributions for Email Bodies using all Labels.

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

3 RESULTS: PAIRED ACCOMPANIED VULNERABILITIES:

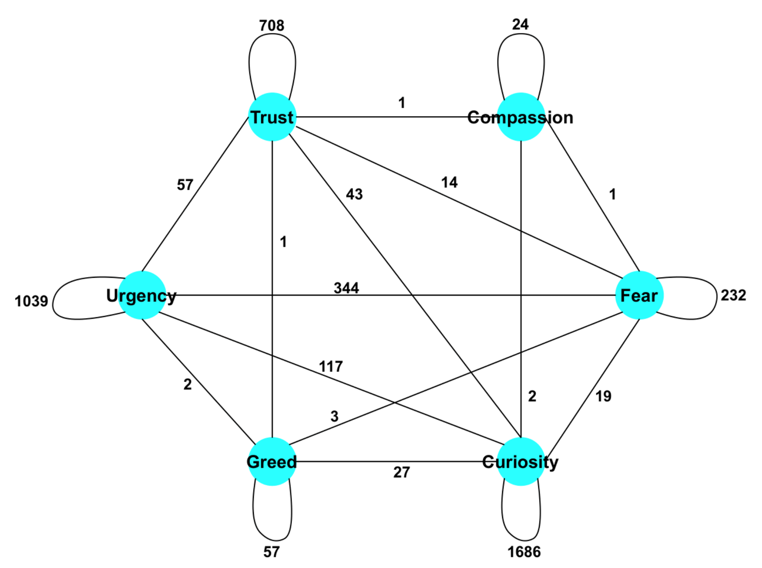

Figure 7: Subject Paired Vulnerabilities.

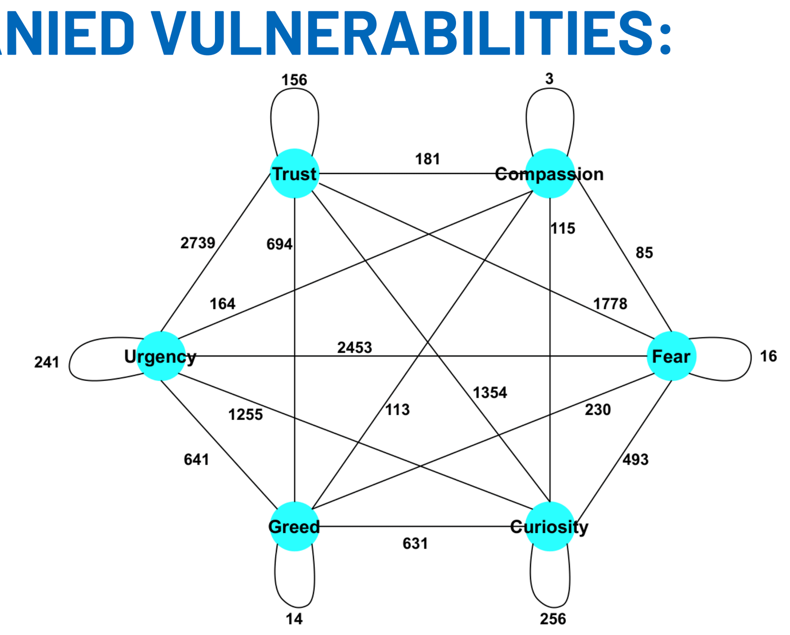

Figure 8: Body Paired Vulnerabilities.

- 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.

RESULTS: VALENCE-AROUSAL MAPPING:

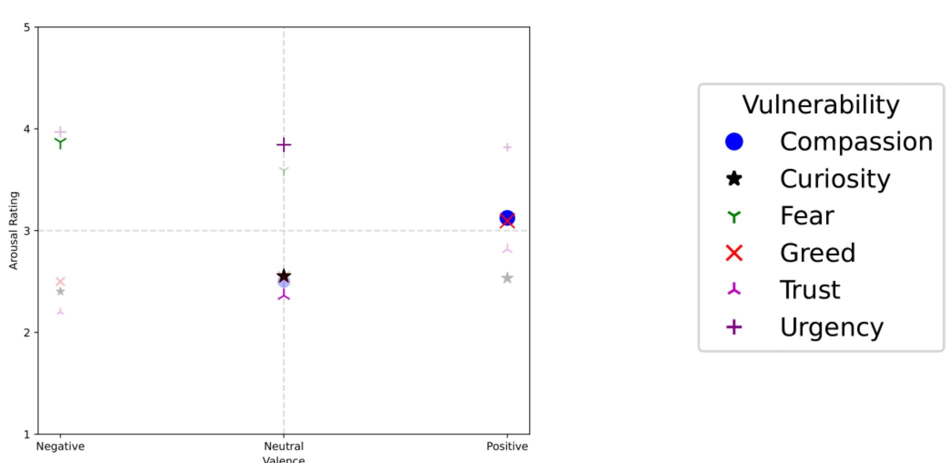

Figure 9: Subject Mapping

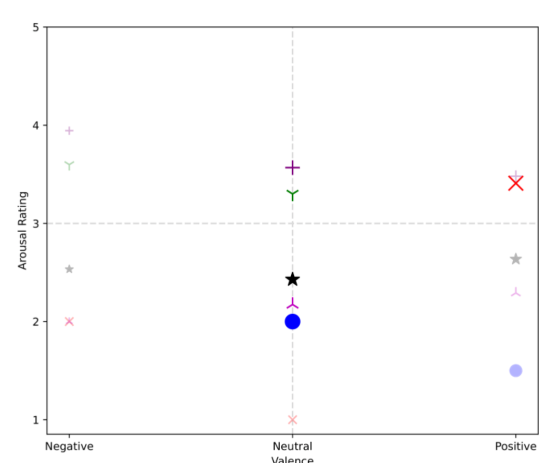

Figure 10: Body Mappings

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

4 RESULTS: LLMS PERFORMANCE VS HUMAN

Analysis	Email	N	κ or α	Soft (%)	Hard (%)
Reliability Between Humans and LLMs					
Psy. Vul.	Subject	133	0.2471	56.39	43.61
	Body	200	0.0179	88.38	17.17
Sentiment	Subject	137	0.4092	-	90.51
	Body	113	0.4015	-	78.76
Reliability Between Humans and GPT4					
Psy. Vul.	Subject	133	0.229	64.66	42.11
	Body	200	0.0174	84.85	19.7
Sentiment	Subject	137	0.3957	-	91.97
	Body	113	0.4401	-	88.5
Reliability Between Humans and GeminiPro					
Psy. Vul.	Subject	133	0.3303	67.67	54.14
	Body	200	-0.0055	77.78	5.05
Sentiment	Subject	137	0.2597	-	82.48
	Body	113	0.2222	-	60.18
Reliability Between Humans and Llama2					
Psy. Vul.	Subject	133	-0.0837	18.8	14.29
	Body	200	-0.0281	92.93	10.61
Sentiment	Subject	137	0.3409	-	87.59
	Body	113	0.3412	-	73.45

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 Krippendorff'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

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

[2] PROOFPOINT. 2022 State of Phishing. https://go.proofpoint.com/en-2022-state-of-the-phish.html

[3] CIALDINI, R. B., AND CIALDINI, R. B. *Influence: The psychology of persuasion*, vol. 55. Collins New York, 2007.

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

[5] CIAMBRONE, G., AND WILSON, S. Creation and analysis of a corpus of scam emails targeting universities. In *Companion Proceedings of the ACM Web Conference 2023* (New York, NY, USA, 2023), WWW '23 Companion, Association for Computing Machinery, p. 24–27.

HOST INSTITUTION

UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

PARTNER INSTITUTIONS

FUNDED BY:

European Union
European Regional Development Fund

Science Foundation Ireland For what's next