PhishGuard: A Generative Pre-Trained Transformer (GPT)-Powered Phishing Email Detection Solution

Angelina M. Messina (messinaan@rider.edu), Zhengping Jay Luo (zluo@rider.edu)
Department of Computer Science & Physics, Rider University

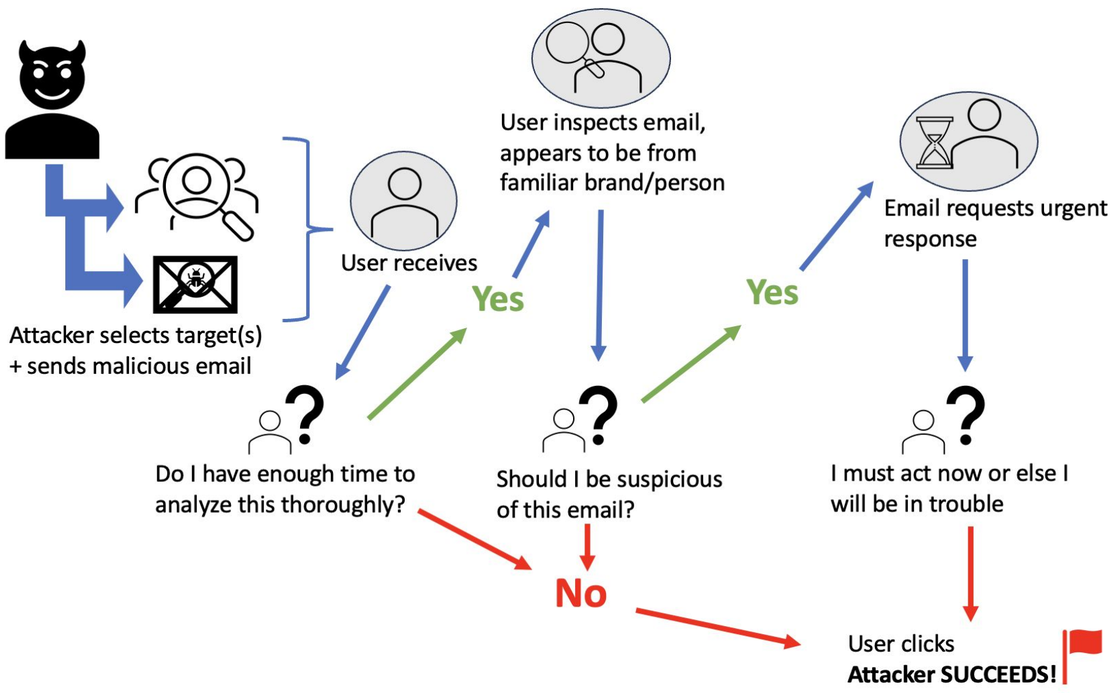
Abstract

Phishing attacks remain one of the most prevalent and evolving cyber threats, deceiving millions of users each year and causing billions in financial losses. Users often fall victim due to three primary factors: (i) lack of attention or time, (ii) visual deception, and (iii) insufficient security awareness. This poster proposes a user-friendly and effective strategy for identifying phishing emails. Training a user-friendly generative pre-trained transformer (GPT), referred to as PhishGuard throughout the poster, can dynamically identify phishing indicators from images or text inputs. This allows users to develop security knowledge and learn on an individualized basis the indicators of phishing content. Our results demonstrate that the proposed GPT model efficiently detects phishing attempts, offering an informative and accessible tool for a diverse range of users.

Research Question

How can personalized GPT-based phishing detection models reduce the success rate of phishing attacks?

Methodology



- To understand the common phishing indicators, python script parses a sample set of 500 real-word corporate emails from the *Enron Email Dataset* [1].
- The analysis procedure includes natural language processing (NLP)-based string matching of header and email content extraction to assess the likelihood that an email is phishing [2].
- After computing a summary containing a keyword analysis report, a dashboard visually portrays the findings. The keywords that appear most frequently are represented in the bar graph (Figure 1) and frequencies of mismatched headers are represented in the pie chart (Figure 2).

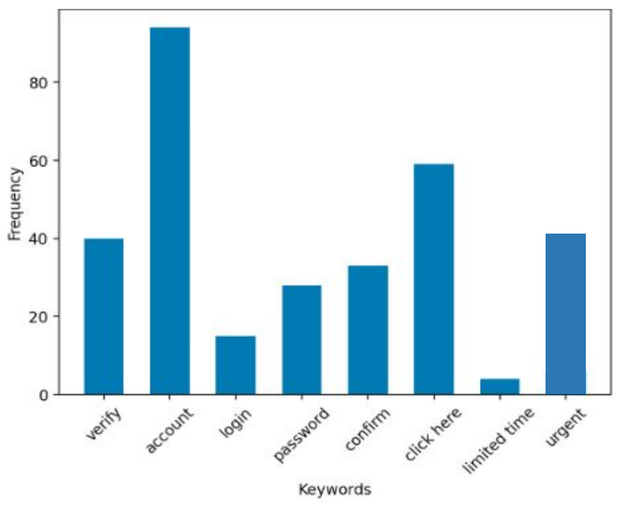


Fig 1. Frequency of phishing keywords

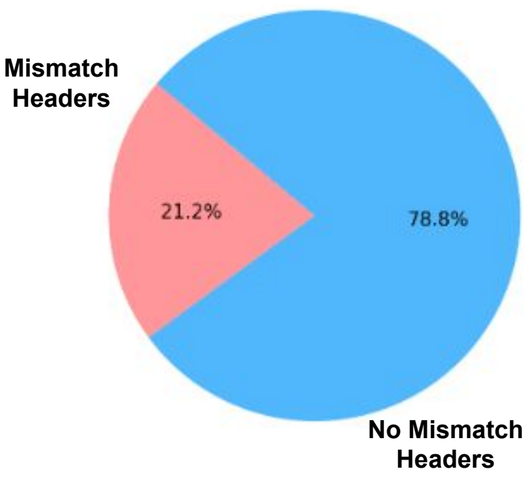


Fig 2. Frequency of mismatched email headers indicate spoofing of sender's identity

- Considering the keyword and mismatch header indicators, a study was conducted including 30 college students. The study assessed the speed and ability of participants to correctly categorize phishing emails.
- Participants received a timed quiz with a combination of phishing and legitimate emails and were tasked with classifying them.

Results

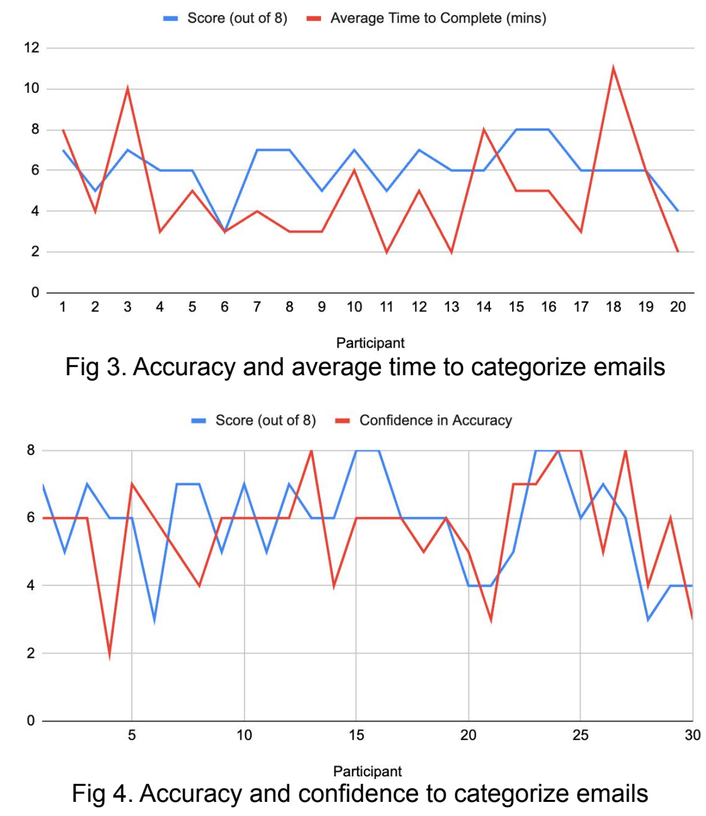
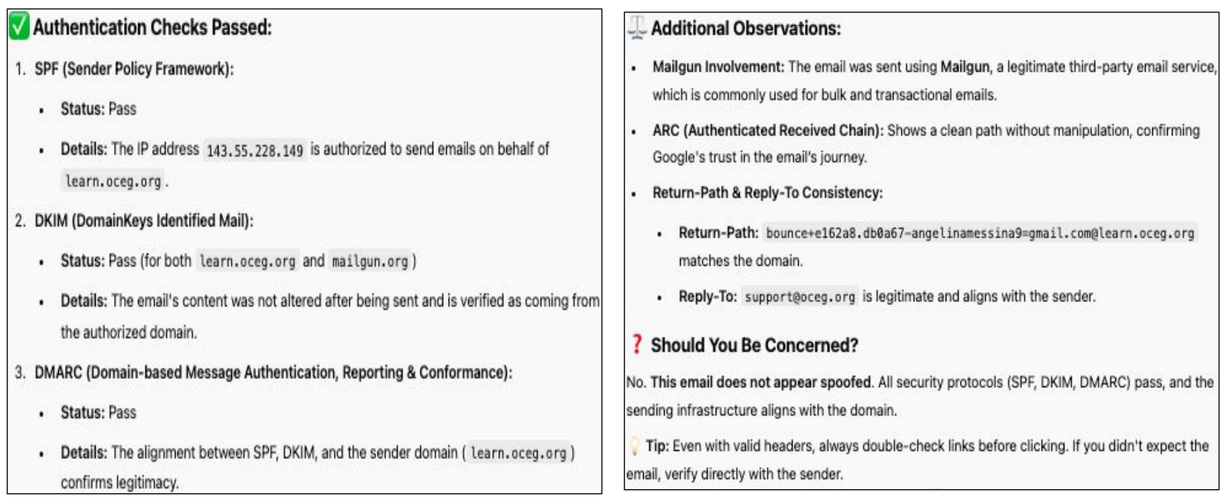


Fig 3. Accuracy and average time to categorize emails

Fig 4. Accuracy and confidence to categorize emails

- There is a correlation between accuracy and time spent analyzing the email (Figure 3).
- The average correctness score was 73.75% and the average time to complete as 4.8 minutes.
- Users more confident in their selections tend to have higher accuracy in identifying phishing emails (Figure 4).
- The average confidence score of users was 71.25%.
- Participants report emails as visually deceiving and indicate interest in a tool to guide decision-making.



Authentication Checks Passed:

1. SPF (Sender Policy Framework):
   - **Status:** Pass
   - **Details:** The IP address `143.55.228.149` is authorized to send emails on behalf of `learn.oceg.org`.
2. DKIM (DomainKeys Identified Mail):
   - **Status:** Pass (for both `learn.oceg.org` and `mailgun.org`)
   - **Details:** The email's content was not altered after being sent and is verified as coming from the authorized domain.
3. DMARC (Domain-based Message Authentication, Reporting & Conformance):
   - **Status:** Pass
   - **Details:** The alignment between SPF, DKIM, and the sender domain (`learn.oceg.org`) confirms legitimacy.

Additional Observations:

- **Mailgun Involvement:** The email was sent using **Mailgun**, a legitimate third-party email service, which is commonly used for bulk and transactional emails.
- **ARC (Authenticated Received Chain):** Shows a clean path without manipulation, confirming Google's trust in the email's journey.
- **Return-Path & Reply-To Consistency:**
  - **Return-Path:** `bounce+e162a8.db0a67-angelinamessina9=gmail.com@learn.oceg.org` matches the domain.
  - **Reply-To:** `support@oceg.org` is legitimate and aligns with the sender.

Should You Be Concerned?

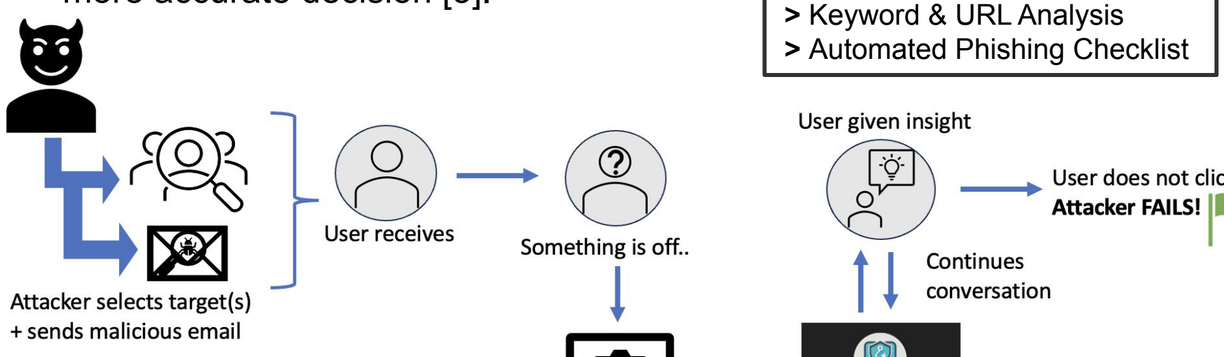
No. **This email does not appear spoofed.** All security protocols (SPF, DKIM, DMARC) pass, and the sending infrastructure aligns with the domain.

Tip: Even with valid headers, always double-check links before clicking. If you didn't expect the email, verify directly with the sender.

Example Output: GPT Authenticating Email Header

- A GPT encourages users to pursue a safer course of action, significantly reducing the human analysis time and leading to a more accurate decision [3].

Capabilities
> File Metadata Analysis
> Domain Reputation Check
> Email Header Analysis
> Keyword & URL Analysis
> Automated Phishing Checklist



- PhishGuard GPT reduces decision time, making the AI-assisted evaluation 2.5 times faster.
- The GPT guides user decision and boosts confidence levels, increasing phishing identification accuracy.

Conclusion

- Common phishing terms can be extracted using NLP tools provided in Python through extracting email semantics and analyzing headers.
- Our Generative Pre-Trained Transformer PhishGuard is capable of analyzing inputted text, images, and documents to provide actionable insights and a verdict on the phishing likelihood while also conversing with the user.
- Our PhishGuard solution provides an intuitive way to educate users, improve confidence in decision-making, and reduce vulnerability and costly impact associated with falling victim to phishing.

References

1. Enron Email Dataset. https://www.cs.cmu.edu/~enron/.
2. Gascon, Hugo, et al. "Reading between the lines: content-agnostic detection of spear-phishing emails." Research in Attacks, Intrusions, and Defenses: 21st International Symposium, RAID 2018, Heraklion, Crete, Greece, September 10-12, 2018, Proceedings 21. Springer International Publishing, 2018.
3. Chavez, Martin R., et al. "Chat generative pre-trained transformer: why we should embrace this technology." American journal of obstetrics and gynecology 228.6 (2023): 706-711.