THE WEB CONFERENCE ACM

MAY 16, 2024

# "Are Adversarial Phishing Webpages a Threat in Reality?"

## Understanding the Users' Perception of Adversarial Webpages

Ying Yuan, Qingying Hao, Giovanni Apruzzese, Mauro Conti, Gang Wang





Università degli Studi di Padova

University of Illinois Urbana-Champaign

Universität Liechtenstein

# Would you give your information to this website?

# Landscape of Phishing

○ Phishing websites are continuously increasing and polluting the Web

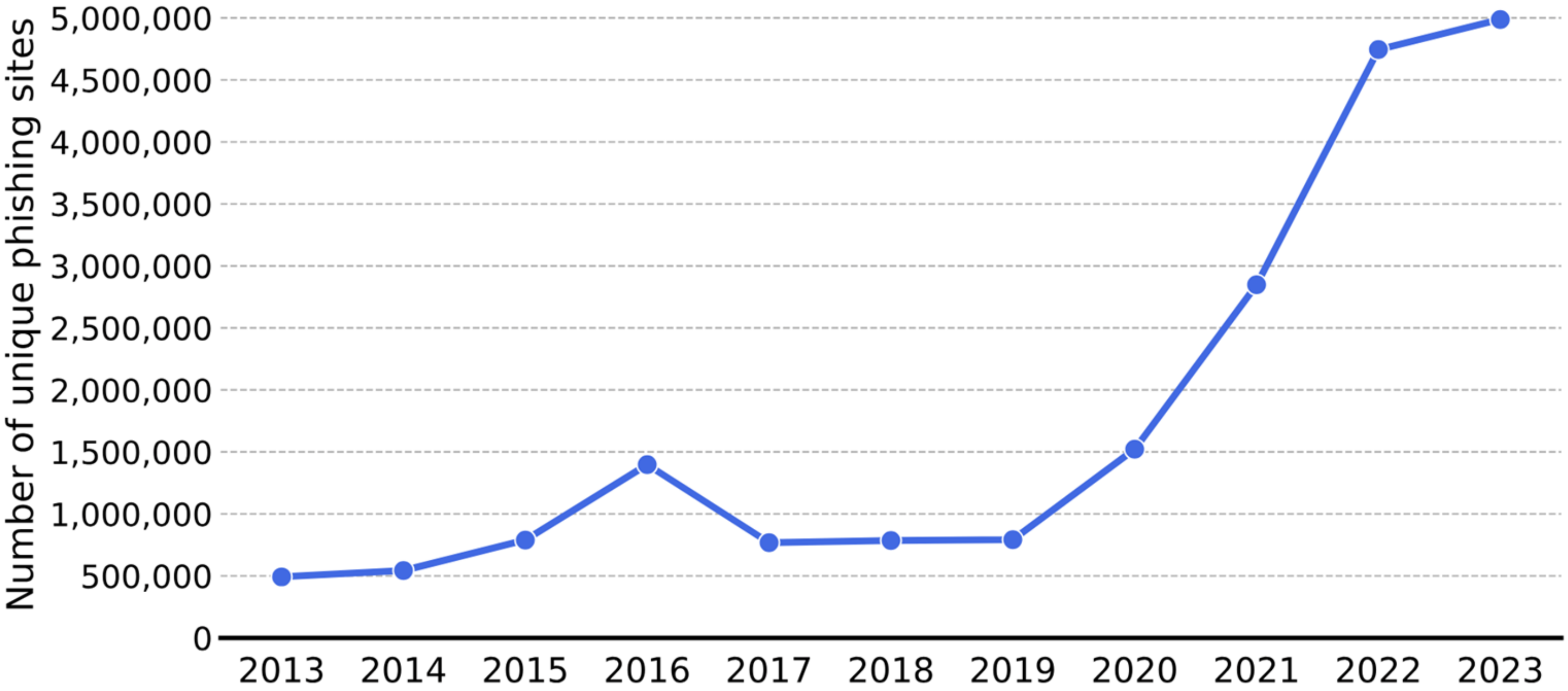
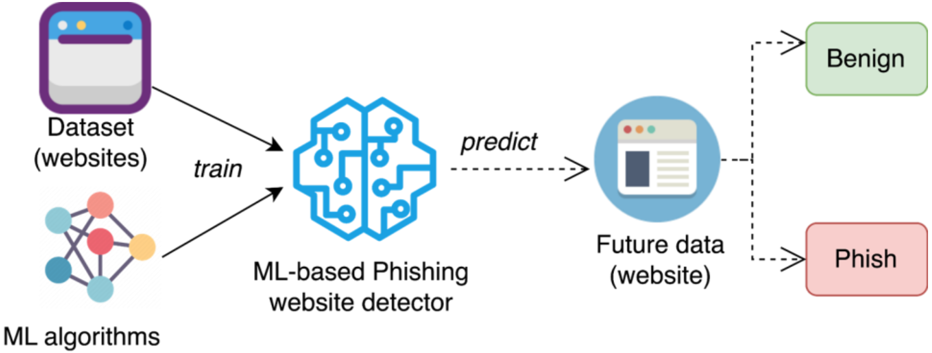



Image reference: APWG, Phishing activity trends report, 2013-2023

# Landscape of Phishing – Countermeasures

- **Blocklist-driven**
  - Low false positive rate, but cannot detect zero-day phishing [1]
- **Data-driven (Machine Learning)**
  - Detect previously unseen phishing
  - Even popular web-browser (Google Chrome) use it [2]

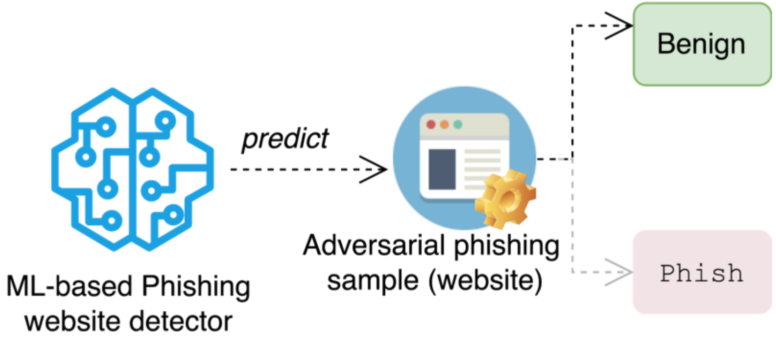




[1] Ke Tian, et al. "Needlein  a haystack: Tracking down elite phishing domains in the wild." In *IMC*, 2018
[2] Google product updates,  https://blog.google/products/chrome/building-a-more-helpful-browser-with-machine-learning/. 2022

# Adversarial Attacks Against ML-PWD

- ML-based Phishing website detector (ML-PWD) are good …
- …but prone to evasion attacks




SpacePhish: The Evasion-space of Adversarial Attacks against Phishing Website Detectors using Machine Learning [3]

Adversarial Sampling Attacks Against Phishing Detection [4]

Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning [5]

Cracking Classifiers for Evasion: A Case Study on the Google's Phishing Pages Filter [6]

Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers [7]

“Real Attackers Don't Compute Gradients”: Bridging the Gap Between Adversarial ML Research and Practice [8]

[3] In *ACSAC*, 2022
[4] In *DBSec*, 2019
[5] In *CCS*, 2018
[6] In *WWW*, 2016
[7] *International Journal of Intelligent Systems* 36, 2021
[8] In *SaTML*, 2023

# Motivation

- Practitioners' viewpoint
  - *"I never thought about securing my machine learning models" [9]*
- To convince them
  - What is the impact of adversarial ML on the end-users in practice?

In the context of Phishing:

- Goal: trick a ***human user*** to input their sensitive data
- '*successful*' evasion attack:
  - bypass the phishing detector…
  - and deceive *human users*

**Would you give your information to this website?**


[9] Boenisch Franziska, et al. "I Never Thought About Securing My Machine Learning Systems": A Study of Security and Privacy Awareness of Machine Learning Practitioners. In *Proceedings of Mensch Und Computer,* 2021

# Research Questions

1. Do **adversarial webpages fool users** as much as they fool ML phishing detectors? (Are adversarial phishing webpages a threat in reality?)

2. Are some **perturbations** more likely to **deceive users**?

3. How do users **perceive adversarial phishing webpages**?
   (e.g., What cues are indicative of users' suspicion, and What perturbations deceive also the human eye?)

# Methodology

Data Collection → ML Phishing Classifier → Adversarial Phish. Pages → User Study → Analysis

**Data Collection**

- 30k benign & phish
- 100 real adversarial sample

**ML Phishing Classifier**

- Custom ML-PWD
- Commercial ML-PWD

**Adversarial Phish. Pages**

- Custom Adversarial Phish. (APW-Lab)
- Real Adversarial Phish. (APW-Wild)

**User Study**

- Baseline study
- Adversarial study
- Recruited N=470

**Analysis**

- Thematic analysis
- Statistical analysis

# Candidate Webpages

We consider **fifteen popular brands** (commonly targeted by phishers)

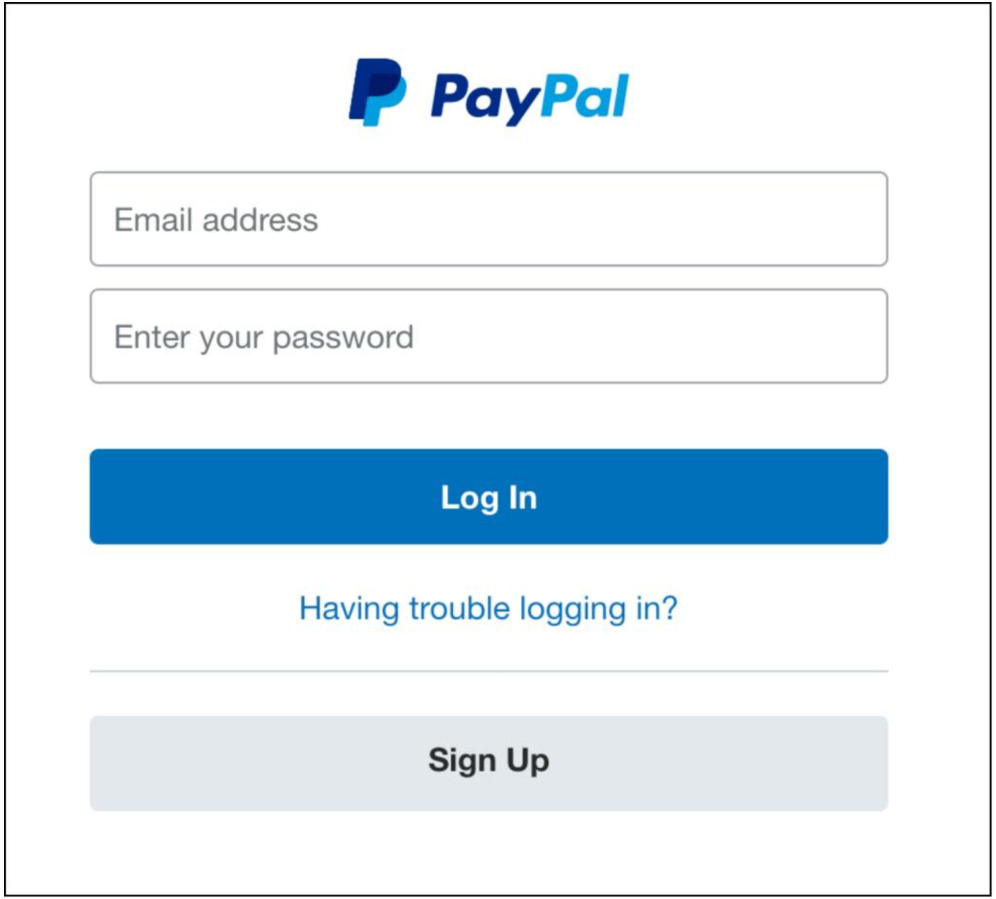
- Adobe, Amazon, Apple, AT&T, Bank of America, DHL, Dropbox, eBay, Facebook, Google, Microsoft, Outlook, Paypal, Wells Fargo, Yahoo

## Classes of Webpages

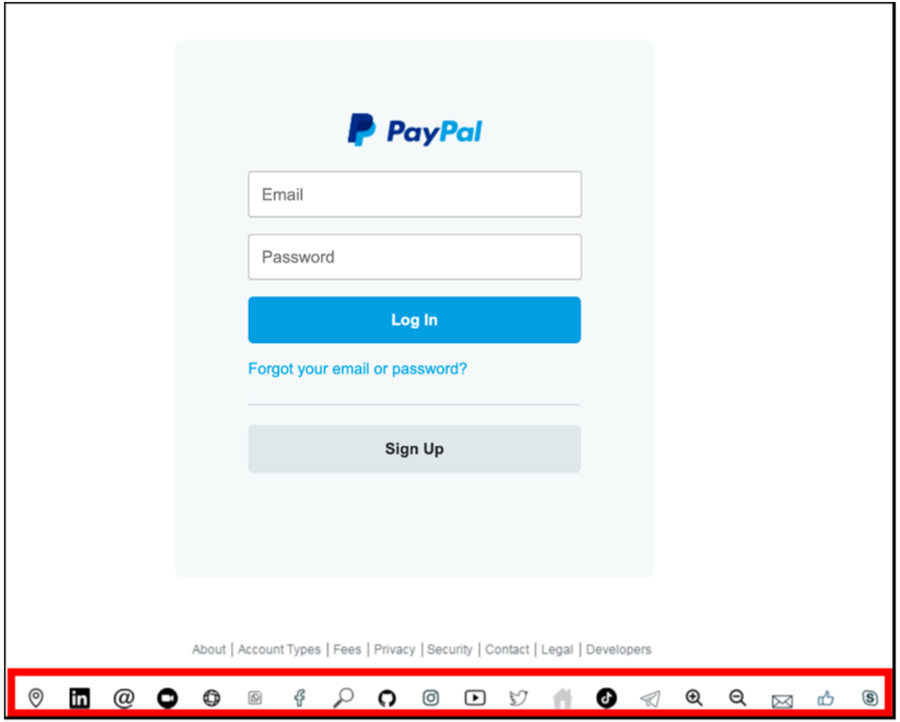
- Legitimate
- Unperturbed Phishing
- Custom Adversarial Phish.
    - APW-Lab_img, APW-Lab_typo, APW-Lab_pswd, APW-Lab_bg
- Real Adversarial Phish. [8]

[10] Similarweb. https://www.similarweb.com/top-websites/. 2023

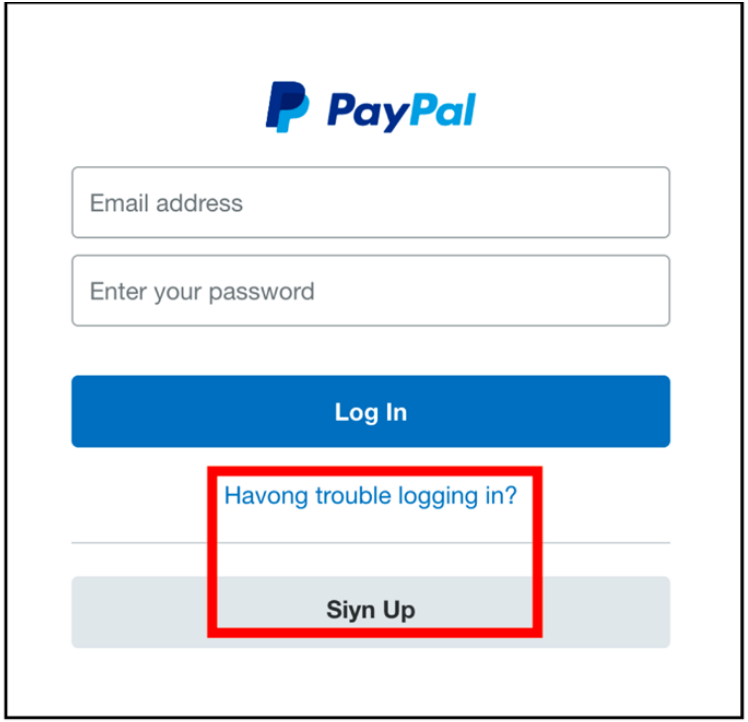
# Candidate Webpages – Unperturbed Phishing



# Candidate Webpages – Custom Adversarial Phish.

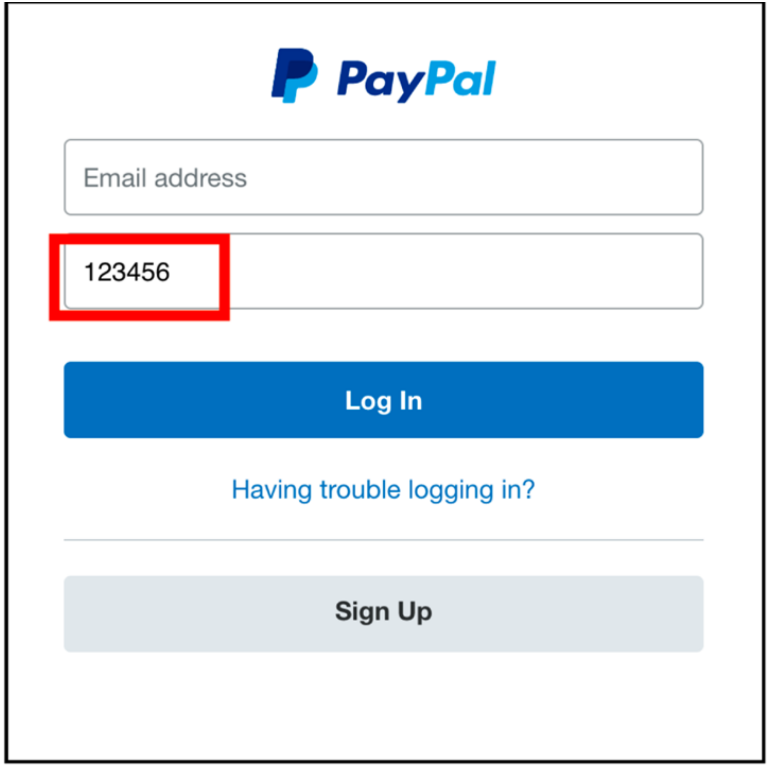



(a) APW-Lab_img

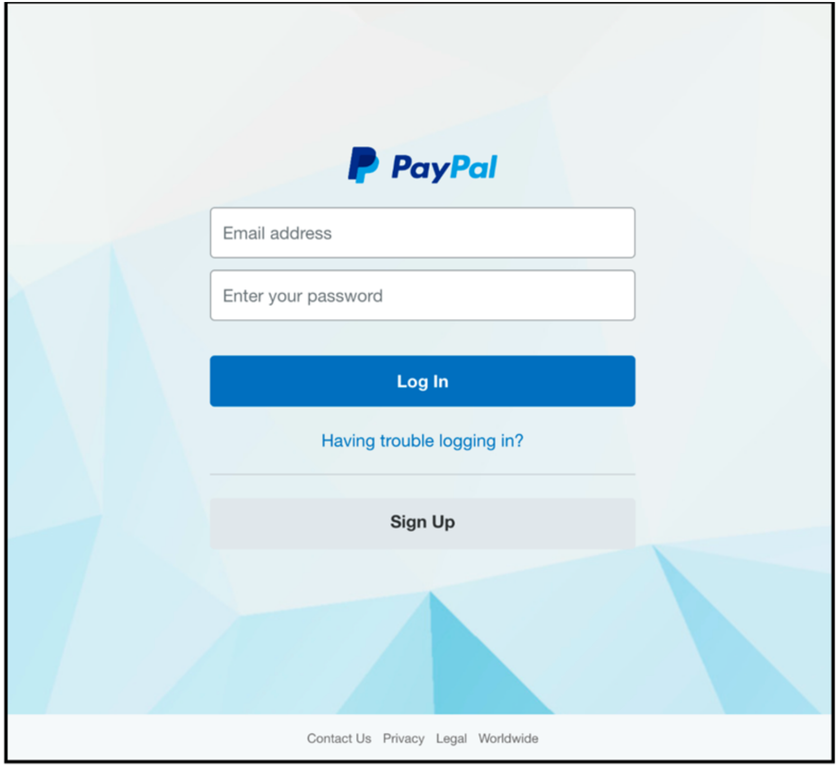



(b) APW-Lab_typo

# Candidate Webpages – Custom Adversarial Phish.




(c) APW-Lab_pswd




(d) APW-Lab_bg

# Participant Task

- Participate once
- Review 15 webpages
  - Rate the legitimacy
  - Provide reasons (open-text)

How do you rate the legitimacy of this webpage?

| | | | | | |
|---|---|---|---|---|---|
| 1 (definitely phishing) | 2 (very probably phishing) | 3 (probably phishing, but not sure) | 4 (probably legitimate, but not sure) | 5 (very probably legitimate) | 6 (definitely legitimate) |

What specific components/indicators on the webpage have influenced your choice?

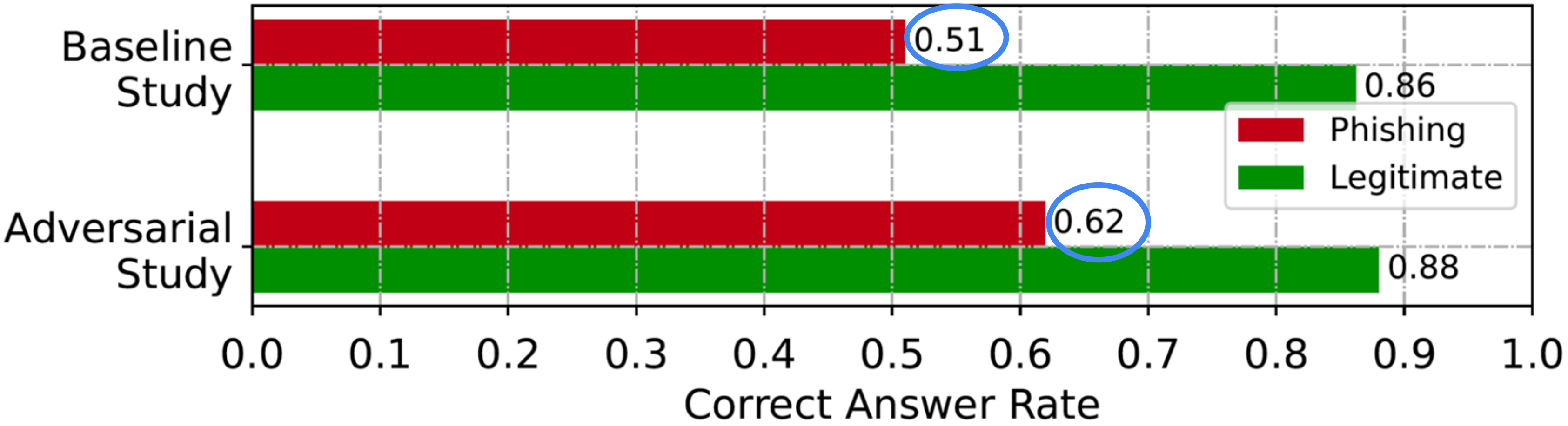
# Research Questions

1. Do **adversarial webpages fool users** as much as they fool ML phishing detectors? (Are adversarial phishing webpages a threat in reality?)

2. Are some **perturbations** more likely to **deceive users**?

3. How do users **perceive adversarial phishing webpages**?
   (e.g., What cues are indicative of users' suspicion, and What perturbations deceive also the human eye?)

# Overall Correct Answer Rate (RQ1)




- Respondents can more easily discern adversarial phishing webpages (62%) than "unperturbed" ones (51%)
- However, 38% of adversarial webpages can still fool users

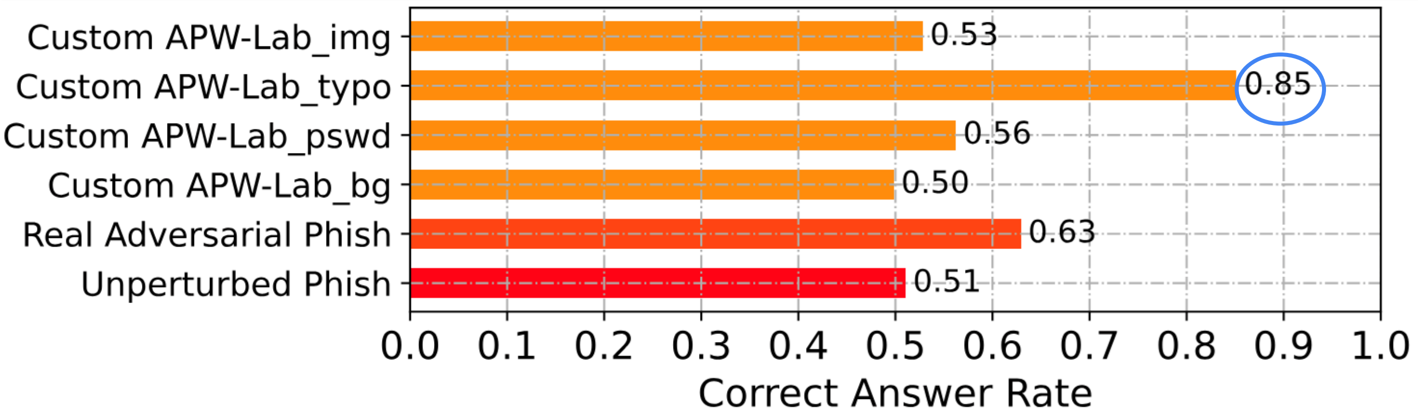
# Research Questions

1. Do **adversarial webpages fool users** as much as they fool ML phishing detectors? (Are adversarial phishing webpages a threat in reality?)

2. Are some **perturbations** more likely to **deceive users**?

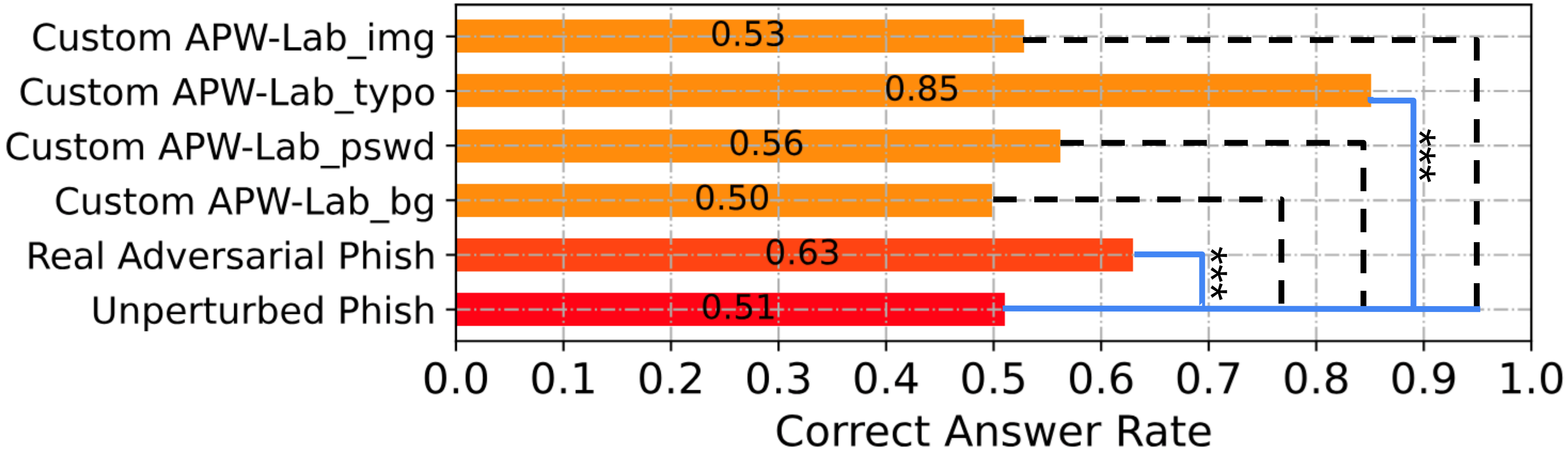
3. How do users **perceive adversarial phishing webpages**?
   (e.g., What cues are indicative of users' suspicion, and What perturbations deceive also the human eye?)

# Detection Rate for Phishing (RQ2)




- Not all adversarial perturbations equally deceive users
- Adversarial phishing webpages with typos are more likely to be perceived

# Detection Rate for Phishing (RQ1/RQ2) – Statistical Analysis




Statistical significance is denoted by *** (*P* < 0.001), **(*P* < 0.01), and * (*P* < 0.05) under binary mixed effect regression

- Except for APW-Lab_typo, adversarial phishing webpages still deceive users

# Research Questions

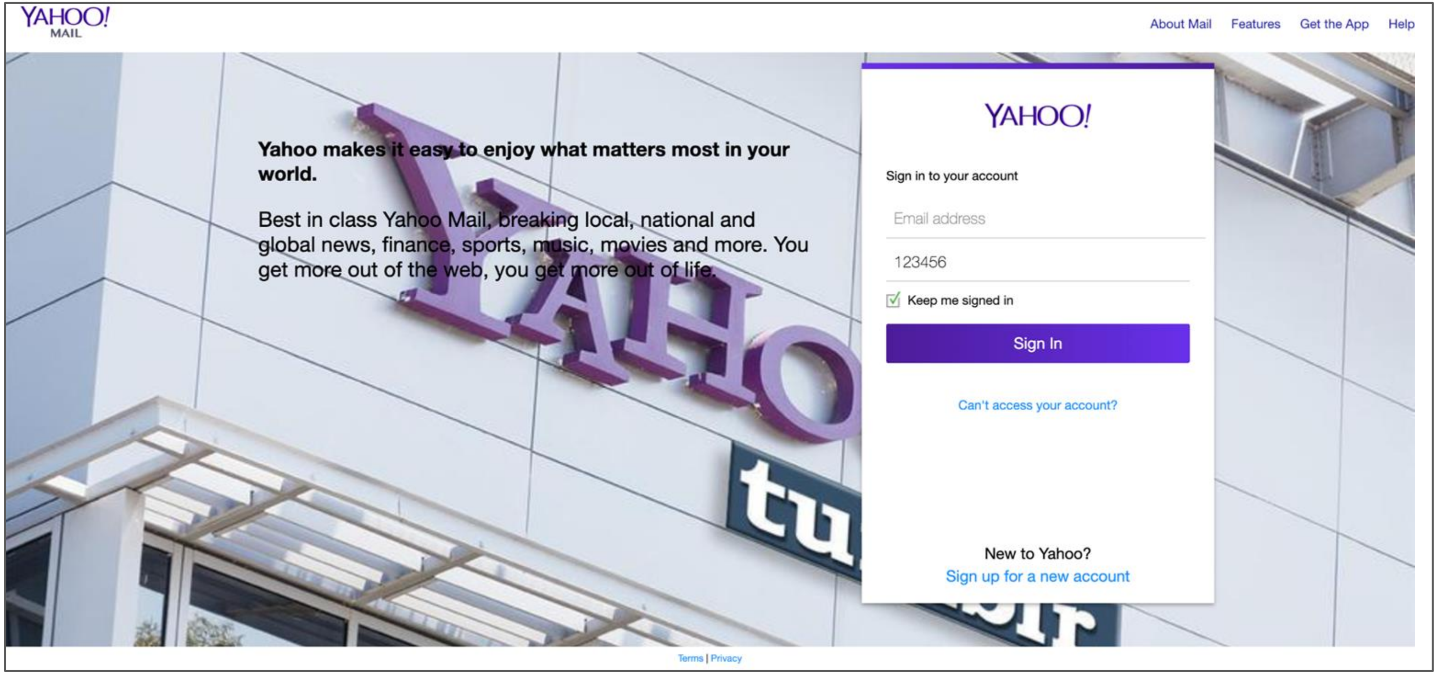
1. Do **adversarial webpages fool users** as much as they fool ML phishing detectors? (Are adversarial phishing webpages a threat in reality?)

2. Are some **perturbations** more likely to **deceive users**?

3. How do users **perceive adversarial phishing webpages**?
   (e.g., What cues are indicative of users' suspicion, and What perturbations deceive also the human eye?)

# Users' Assessment Strategies – Exemplary (RQ3)

What specific components/indicators on the webpage have influenced your choice?




"**icons**, **photo** and **sign in info** look correct" – P560

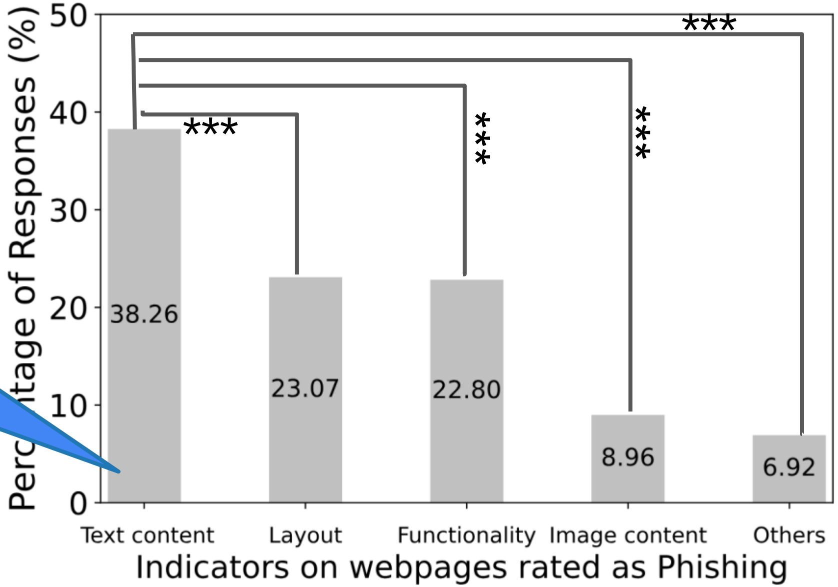
Thematic analysis

- coding 1,307 (37%) answers

# Users' Assessment Strategies – Rated as Phishing (RQ3)

- Text content is the most prevalent factor
- Few answers mention image content

**Textual content** significantly influences the perceived credibility of webpages.




Statistical significance is denoted by *** ($P < 0.001$), **($P < 0.01$), and * ($P < 0.05$) under pairwise Chi-squared tests

# Summary

## Adversarial Phishing Webpages

- A threat in reality
- Vary in artifacts

## Perturbations

- Typos increase suspicion
- Visual perturbation deceive users

## User Perception

- Mostly rely on textual, layout, functionality
- Rarely based on image/misinformed cues
- Affect by phishing knowledge & visiting frequency

*Thanks!*

*Check out our paper!*

[https://threatadvphish.github.io](https://threatadvphish.github.io)

*ying.yuan@phd.unipd.it*

**Ying Yuan, Qingying Hao, Giovanni Apruzzese, Mauro Conti, Gang Wang**