E-PhishGen: Unlocking Novel Research in Phishing Email Detection

Luca Pajola*†, Eugenio Caripoti*†, Stefan Banzer§, Simeone Pizzi†, Mauro Conti*‖, Giovanni Apruzzese§¶

*University of Padua, †SpritzMatter S.R.L., ‖Orebro University, §University of Liechtenstein, ¶Reykjavik University
(luca.pajola, eugenio.caripoti, simeone.pizzi)@spritzmatter.com, mauro.conti@unipd.it, (stefan.banzer, giovanni.apruzzese)@uni.li

Abstract

Every day, our inboxes are flooded with unsolicited emails, ranging between annoying spam to more subtle phishing scams. Unfortunately, despite abundant prior efforts proposing solutions achieving near-perfect accuracy, the reality is that countering malicious emails still remains an unsolved dilemma.

This "open problem" paper carries out a critical assessment of scientific works in the context of phishing email detection. First, we focus on the *benchmark datasets* that have been used to assess the methods proposed in research. We find that most prior work relied on datasets containing emails that—we argue—are not representative of current trends, and mostly encompass the English language. Based on this finding, we then re-implement and re-assess a variety of *detection methods reliant on machine learning* (ML), including large-language models (LLM), and release all of our codebase—an (unfortunately) uncommon practice in related research. We show that most such methods achieve near-perfect performance when trained and tested on the same dataset—a result which intrinsically hinders development (how can future research outperform methods that are already near perfect?). To foster the creation of "more challenging benchmarks" that reflect current phishing trends, we propose E-PhishGEN, an LLM-based (and privacy-savvy) framework to generate novel phishing-email datasets. We use our E-PhishGEN to create E-PhishLLM, a novel phishing-email detection dataset containing 16616 emails in three languages. We use E-PhishLLM to test the detectors we considered, showing a much lower performance than that achieved on existing benchmarks—indicating a larger room for improvement. We also validate the quality of E-PhishLLM with a user study (n=30). To sum up, we show that phishing email detection is still an open problem—and provide the means to tackle such a problem by future research.

RQ1: *what benchmark datasets are used in related literature to assess previously-proposed phishing email detectors?*

We answer RQ1 via a literature review across 562 papers retrieved from top-tier conferences (WWW, S&P/EuroS&P, USENIX SEC, ACSAC, NDSS, CCS, AsiaCCS, IMC, CHI, WDSM) between 2014—2024 and complemented with Google-Scholar searches

> **Answer to RQ1.** Previously proposed methods are evaluated on datasets (such as SpamAssassin, SpamBase, Enron, Nazario, LingSpam) that have old (and monolingual) emails—which hardly resemble current phishing trends. Moreover, related literature often does not release their codebase: this is problematic especially given that it prevents accurate replication of the testbed.

Ancillary Findings:

- No clear naming for datasets,
- Usage of datasets that mix spam with phishing emails,
- Other (rarely used) datasets are continuously mutating, or are not available anymore, or are not open source.

RQ2: *what performance do existing detectors achieve on some previously-used benchmark datasets for phishing email detection?*

We take 8 existing datasets (CEAS, TREC, Chataut, SpamAssassin, and two variants of Enron and LingSpam) and use them to re-assess existing ML-based phishing email detectors, spanning across 5 ML models (i.e., RF, LR, NB, SVM, MLP) using TF-IDF features, a feature-agnostic BERT-based model (DistilBERT) fine-tuned on these datasets, and two commercial LLMs (Gemini-2.0 flash and GPT-4o-mini) in a zero-shot fashion. We assess the performance in a "same-dataset" setting, as well as in a "cross-dataset" setup, and even in a "all-vs-one" scenario in which a model is trained on all datasets but one, and tested on the remaining one. This required us to put all datasets in a common format. We release our entire implementation in our GitHub.

> **Answer to RQ2.** We confirm that models requiring a training phase achieve near-perfect performance when tested on data from the same dataset. Yet, such models struggle when tested on data from other datasets—but DistilBERT seems to have a better generalization power. Zero-shot-prompted LLMs exhibit high F1-scores when tasked to detect phishing emails from our considered datasets, but the performance on Chataut is poor.

[Subject] Help!

You have been specially selected to qualify for the following:
Premium Vacation Package and Pentium PC Giveaway
To review the details, please click on the link below using the confirmation number:
http://www.1chn.net/wintrip
Confirmation Number: **Lh340**
Please confirm your entry within 24 hours of receiving this confirmation.
Wishing you a fun-filled vacation!
If you have any additional questions or cannot connect to the site, do not hesitate to contact me:

`vacation@btamail.net.cn`

Email 1. An email in the popular dataset SpamAssassin [39] (from 2005).

RQ3: *what is a way to overcome the shortcomings of existing datasets, without raising privacy concerns?* ➔ E-PhishGEN & E-PhishLLM

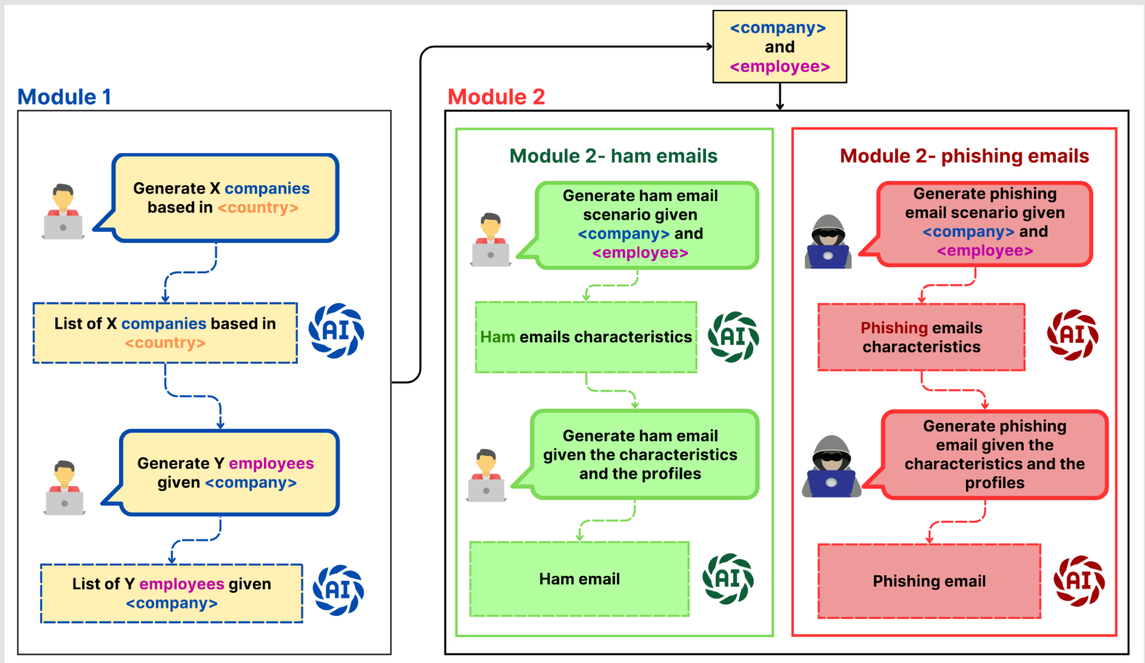

Fig. 3: Overview of E-PhishGEN.. The framework is composed of two modules: Module 1 generates synthetic company and employee profiles based on high-level input prompts; Module 2 takes these profiles as input and uses an LLM to generate realistic emails, including both benign and phishing variants, tailored to the organizational context and user roles.

[Subject] Scheduling a Call for Supply Chain Adjustments

Dear Marco, I hope this message finds you well. We need to schedule a video call to discuss some adjustments and potential delays in the supply chain affecting our current project components. Could you please inform me of your availability this week? Looking forward to hearing from you. Best regards, TomJohnson

Email 2. Illustrative example of a benign email in E-PhishLLM.

[Subject] Urgente: Verifica delle Credenziali dell'Account

Ciao Marco, Ti scrivo per conto del tuo manager per richiedere un'urgenteverifica delle tue credenziali aziendali. È molto importante che tuproceda al controllo immediato della correttezza delle informazioni d'accesso personali. Si prega di seguire il link di verifica di seguito e aggiornare qualsiasi informazione necessaria quanto prima:«link» Grazieperla tua collaborazione. Cordiali saluti, Federica Rossi Responsabile IT Fabbri Tech Automazione

Email 3. Illustrative example of a malicious email in E-PhishLLM.

RQ4: *what performance do previous methods achieve on E-PhishLLM?*

> **Answer to RQ4.** Detectors trained on any combinations of our eight considered datasets struggle to detect the phishing emails in our E-PhishLLM dataset. LLMs, however, are much more effective. These results show that our proposed E-PhishLLM dataset represents a better "benchmark" than existing datasets to test previously proposed detectors.

RQ5: *does E-PhishLLM contain phishing emails of a higher quality than those included in some previously-proposed phishing email datasets?*

We carry out a user study via an online questionnaire. We received 30 responses, of which only 16% consider themselves as "beginners" from a viewpoint of IT expertise. Participants were roughly split between 18—25 and 26—40 years of age. The questionnaire included 20 emails, 5 from E-PhishLLM and 5 from SpamAssassin, Nazario, and Enron. Participants had to answer one question for each email: "How would your rate the overall phishing quality of this email?" (Answer in a 1—5 Likert Scale). The average score for E-PhishLLM was 3.41, whereas the others all scores significantly ($p<.05$) lower (Nazario=2.65, SpamAssassin=1.57, Enron=1.45).

> **Answer to RQ5.** Yes, E-PhishLLM contains phishing emails of higher quality than those of SpamAssassin, Nazario, and Enron.

Repository: https://github.com/pajola/e-phishGen