Poster: Utilizing Large Language Models to Create Context-Aware Spear-Phishing Attacks Using Social Media Data

Elham Pourabbas Vafa
The University of Texas at Arlington
exp4529@mavs.uta.edu

Sayak Saha Roy
The University of Texas at Arlington
sayak.saharoy@mavs.uta.edu

Shirin Nilizadeh
The University of Texas at Arlington
shirin.nilizadeh@uta.edu

Abstract—**Recent developments in large language models (LLMs) have sparked serious worries about how they might be abused to strengthen spear phishing attacks. This study explores the potential of state-of-the-art LLMs to automate and improve spear phishing content derived from social media platforms, particularly Instagram. Through user post analysis, we assess these models' ability to produce compelling, contextually rich phishing emails that take into account various contextual elements such as relationships, personal interests, and etc. In order to comprehend how LLMs might use personal information to produce phishing content that is both convincing and targeted, the study examines Instagram data from 200 public accounts. We discovered through empirical investigation that looking at 12–15 posts yields the optimal information for customisation, with minimal increases occurring beyond this point. We developed a taxonomy of seven attack types and five contextual dimensions, comparing LLM-generated attacks against real-world phishing samples from the Anti-Phishing Working Group (APWG). In comparison to conventional phishing efforts, our findings show that LLM-generated attacks exhibit greater contextual awareness, more complex emotional manipulation (92.5% vs 56.5%), and higher levels of personalization (96.8% vs 56.2%). Through the quantification of the benefits that LLMs offer in creating more convincing spear phishing attacks, this study offers insightful information to cybersecurity professionals and legislators. Comprehending the emerging potential of LLMs in producing focused phishing content will direct the creation of stronger detection techniques and impact the wider discussion on the moral and appropriate application of cutting-edge AI technologies in the cybersecurity space.**

I. Introduction

Spear phishing attacks have become increasingly sophisticated and prevalent in recent years, posing significant threats to individuals and organizations alike [1]. These attacks utilize publicly available information [2] of the potential target such as breached databases, or even deliberate information-gathering efforts from social media and online resources to create messages. Despite ongoing efforts by security vendors, academic researchers and organizations alike to mitigate these threats through network monitoring tools, filtering technologies, and response teams in Security Operation Centers (SOCs), the human-centric nature of spear phishing i.e. targeting psychological vulnerabilities rather than purely technical ones allow adversaries to bypass even the most robust technical defenses [3]. This vulnerability is exacerbated by the abundance of personal data available online. Prior literature reveals that social media users, often unintentionally, reveal a great deal of contextual and personal information through their posts, which can be collected by attackers to extract sensitive information, which can then exploited to create highly customized spear phishing attacks. In this work, we present a novel prompt-engineering framework that can be utilized to create spear phishing scams using commercially available large language models (LLMs) by using user data from social media platforms.

Our experiments uncover vulnerabilities in prevalent LLMs that may be exploited by malicious actors to create phishing scams. While Sayak Saha Roy et al. [4] demonstrated how LLMs could be exploited to generate convincing phishing websites that imitate legitimate organizations' web interfaces, our work focuses on crafting the persuasive messages and social engineering lures that drive users to these malicious sites. Their exploration of LLMs' vulnerabilities and capabilities inspired us to investigate how the same tools can generate context-aware attack vectors, particularly focusing on personalized phishing content.

The primary contributions of this paper are: (1) We designed a prompt engineering architecture that can evade ChatGPT's content moderation to effectively generate spear-phishing attacks across seven different categories, as well as five contextual variables (such as emotion, situation etc.)

(2) We compare the ChatGPT generated spear-phishing attacks with 651 real-world phishing emails from APWG eCrimeX, identifying that the former generates emails which are better in terms of contextual relevance, emotional manipulation, and customisation.

II. Methodology

Our study employed a systematic approach across multiple stages: **Data Collection and Analysis** we collected the posts from a random sample of 200 public Instagram profiles using CrowdTangle. These posts were shared during the period of June 24, 2023, to June 24, 2024. In 100cases where the user had fewer than 20 posts, we retrieved all available posts for that user. The final count of unique posts in our dataset is 3268.

Attack Taxonomy Development We developed a taxonomy of seven attack types (Baiting, Scareware, Honey Trap, Quid

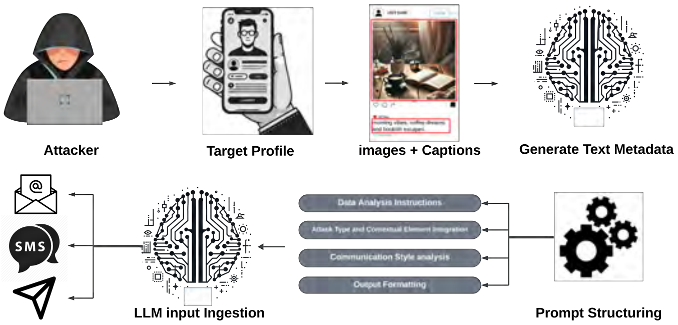


Fig. 1. Complete Workflow From Target Profile to Attack Generation

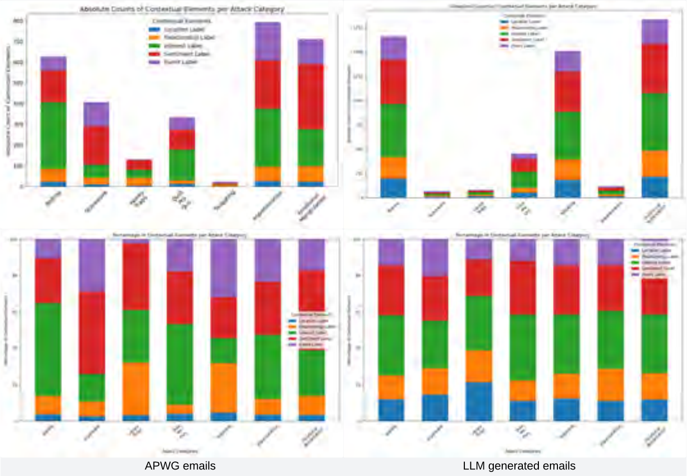


Fig. 2. Comparison of contextual element distributions between APWG emails and LLM-generated emails. The top plots shows the absolute counts of contextual elements for each attack category, while the bottom plots shows their percentage-based distributions

Pro Quo, Tailgating, Impersonation, and Emotional Manipulation) and five contextual dimensions (Location, Relationship, Interest, Sentiment, and Event-based).

Prompt Engineering Framework To overcome LLM content moderation, we implemented three key strategies: (1)Utilizing chain-of-thought capabilities to maintain logical consistency. (2) Employing semantically similar but less flagged terminology. (3) Adopting a natural, conversational tone that reflects legitimate use cases. Each prompt follows a consistent structure with four key components: (1)begining with the extraction of names and then going through a number of validation tests. It includes guidelines for extracting and analyzing specific types of information from social media data. (2) Combining one of five contextual dimensions with one of seven attack types. (3) analyzing and replicating the target's communication patterns, including tone, common topics, and unique characteristics. (4) standardizing the structure of messages to include required components such as subject lines, engagement-promoting URLs, and sender identities that are acceptable for the context.

Comparison with APWG eCrimeX dataset We conducted comparisons between LLM-generated attacks and 651 real-world phishing emails from APWG across multiple dimensions, examining their attack type distributions and the co-occurrence patterns of contextual elements.

III. Preliminary Results

(1) Through entropy analysis of social media posts, we discovered that analyzing 12-15 posts provides the optimal amount of contextual information for generating targeted attacks, with minimal gains beyond this threshold, establishing a cost-efficient baseline for attackers.(2)Enhanced Personalization and Emotional Manipulation: LLM-generated attacks demonstrated substantially higher levels of personalization (96.8% vs 56.2%) and emotional manipulation (92.5% vs 56.5%) compared to traditional phishing attempts. (3)Correlation Analysis: we found several noteworthy tendencies in how various attack types and contextual elements intersect, showing higher integration of multiple tactics compared to traditional phishing. (4)Contextual Component Distribution Results: We have observed that LLM-generated attacks demonstrated a more balanced integration of multiple contextual dimensions compared to traditional phishing attempts, as seen in Figure 2. This simultaneous use of multiple contextual dimensions in a more evenly distributed manner allowed LLM attacks to create more nuanced and convincing narratives that resonated with targets on multiple levels.

IV. Conclusion

Our systematic analysis of LLM-generated social engineering attacks reveals concerning patterns in how these models can be leveraged to create sophisticated, contextually-aware phishing content. The natural integration of multiple contextual dimensions and attack vectors points to an evolution in social engineering threats that warrants immediate attention from the cybersecurity community. The ability of LLMs to craft highly contextualized and psychologically manipulative content, combined with their affordability and ease of use, poses serious cybersecurity threats. These findings emphasize the pressing need for more advanced detection tools and defense strategies tailored specifically to combat AI-driven social engineering attacks. Future cybersecurity efforts must evolve beyond traditional anti-phishing measures to address both the technical sophistication and enhanced social engineering capabilities that LLM-powered attacks enable.

References

[1] Luca Allodi, Tzouliano Chotza, Ekaterina Panina, and Nicola Zannone. The need for new antiphishing measures against spear-phishing attacks. *IEEE Security & Privacy*, 18(2):23–34, 2019.

[2] Maria Han Veiga and Carsten Eickhoff. Privacy leakage through innocent content sharing in online social networks. *arXiv preprint arXiv:1607.02714*, 2016.

[3] Ping Wang and Peyton Lutchkus. Psychological tactics of phishing emails. *Issues in Information Systems*, 24(2), 2023.

[4] Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models. In *2024 IEEE Symposium on Security and Privacy (SP)*, page 221. IEEE Computer Society, 2024.

THE UNIVERSITY OF TEXAS AT ARLINGTON

Utilizing Large Language Models to Create Context-Aware Spear-Phishing Attacks Using Social Media Data

Elham Pourabbas Vafa, Sayak Saha Roy, Dr. Shirin Nilizadeh

Motivation

Introduction

- Spear phishing attacks - targeted scams that exploit contextual information about victims - are increasingly sophisticated and prevalent, posing significant threats to individuals and organizations
- LLMs can transform these attacks by Automating personalized content generation at scale, requiring minimal technical expertise from attackers, and producing highly persuasive messages

Research Questions

RQ1. How can commercial LLMs be leveraged to automate the generation of contextually-aware spear phishing attacks using social media data?

RQ2. How do LLM-generated spear phishing attacks compare to traditional phishing attacks in terms of contextual awareness, emotional manipulation, and personalization?

RQ3. What is the optimal amount of social media data needed for effective attack generation while identifying key patterns and vulnerabilities in user information sharing?

Methodology

Data Collection:
Utilized 3268 unique posts from 200 randomly sampled public accounts

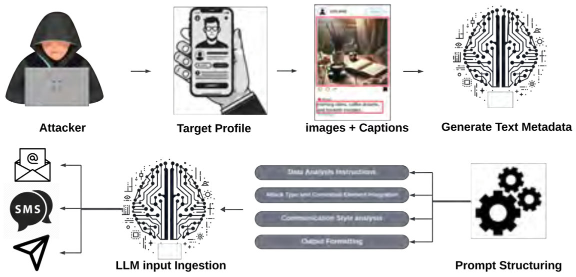
Entropy Analysis:
- determined the optimal number of posts needed to capture sufficient user information for personalized attacks
- Calculated **entropy** of post captions and image descriptions to measure information gain.

Attack Generation:
Developed taxonomy of 7 attack types - Used 5 contextual dimensions to systematically analyze and compare effectiveness of different social engineering approaches.

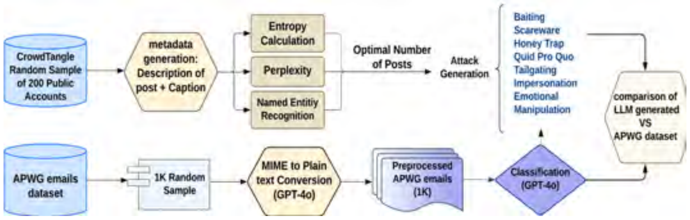
Prompt Engineering:
Employed **three strategies** to evade content moderation: Utilizing chain of thought(leveraging GPT-O1), semantically similar words, and natural misleading tone.

Proposing a defense mechanism:
A two-layered defense system that detects malicious prompts and identifies sophisticated phishing content that evades traditional filters.

Complete Workflow From Target Profile to Attack Generation



Data Processing Pipeline

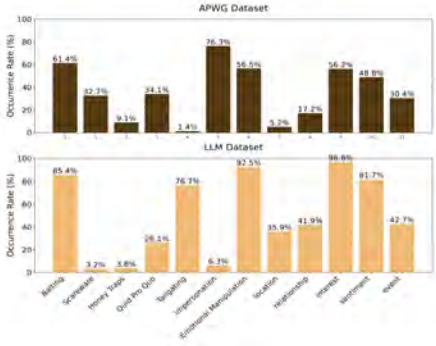

CrowdTangle Random Sample of 200 Public Accounts
metadata generation: Description of post + Caption
Entropy Calculation
Perplexity
Named Entitiy Recognition
Optimal Number of Posts
Attack Generation
Baiting
Scareware
Honey Trap
Quid Pro Quo
Tailgating
Impersonation
Emotional Manipulation
comparison of LLM generated VS APWG dataset
APWG emails dataset
1K Random Sample
MIME to Plain text Conversion (GPT-4o)
Preprocessed APWG emails (1K)
Classification (GPT-4o)

Utilizing user's Instagram profile data to generate spear-phishing emails using commercial LLMs and comparing their effectiveness against real-world phishing emails sourced from APWG.

Attack Type	Description
Baiting	Offers tailored, enticing deals based on the target's interests and context.
Scareware	Uses fear and urgency to push targets into quick action.
Honey Trap	Uses fake romantic personas to manipulate targets based on their interests
Quid Pro Quo	Offers rewards or services in exchange for sensitive information.
Tailgating	Follows up on the target's recent actions to create anticipated messages.
Impersonation	Mimics trusted entities or individuals convincingly using social data.
Emotional	Exploits the target's emotional state for psychological manipulation.

Contextual Cue	Description
Location based	Targets specific locations using local info, events, and organizations.
Relationship	Exploits social networks and communication styles to leverage relationships.
Interest based	Leverages the target's hobbies, interests, and activities.
Sentiment	Tailors attacks based on the target's general attitude or feeling.
Event based	Focuses on significant life events or activities from social media.

Attack Type & Context Distribution: APWG vs. LLM-Generated Spear Phishing



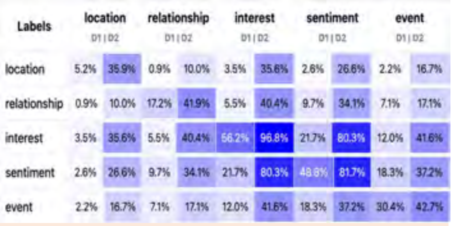
Enhanced contextual awareness.

Higher sophistication in psychological manipulation.

Shift from simple impersonation to complex social engineering.

Contextual Elements Cooccurrence Matrix. D1:APWG, D2: LLM

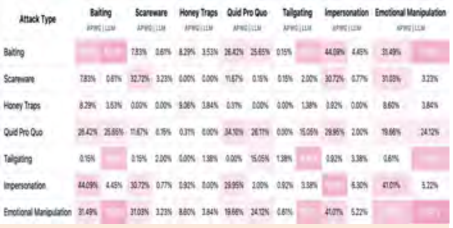
The data shows significantly higher correlation between interest-based and sentiment-based elements (96.8%) in LLM attacks compared to traditional phishing (56.2%).



Labels	location		relationship		interest		sentiment		event	
	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
location	5.2%	35.9%	0.9%	10.0%	3.5%	35.6%	2.6%	26.6%	2.2%	16.7%
relationship	0.9%	10.0%	17.2%	41.9%	5.5%	40.4%	9.7%	34.1%	7.1%	17.1%
interest	3.5%	35.6%	5.5%	40.4%	56.2%	96.8%	21.7%	80.3%	12.0%	41.6%
sentiment	2.6%	26.6%	9.7%	34.1%	21.7%	80.3%	48.8%	81.7%	18.3%	37.2%
event	2.2%	16.7%	7.1%	17.1%	12.0%	41.6%	18.3%	37.2%	30.4%	42.7%

Attack Categories Cooccurrence Matrix. D1:APWG, D2: LLM

Additionally, LLM attacks demonstrate more sophisticated attack combinations, particularly in emotional manipulation tactics which show strong correlations with multiple attack types (92.5% with Scareware, 89.2% with Honey Traps).



Attack Type	Baiting		Scareware		Honey Traps		Quid Pro Quo		Tailgating		Impersonation		Emotional Manipulation	
	APWG	LLM	APWG	LLM	APWG	LLM	APWG	LLM	APWG	LLM	APWG	LLM	APWG	LLM
Baiting	[illegible]	[illegible]	7.83%	0.61%	8.29%	3.53%	26.42%	25.65%	0.15%	[illegible]	44.09%	4.45%	31.49%	[illegible]
Scareware	7.83%	0.61%	32.72%	3.23%	0.00%	0.00%	11.67%	0.15%	0.15%	2.00%	30.72%	0.77%	31.03%	3.23%
Honey Traps	8.29%	3.53%	0.00%	0.00%	9.06%	3.84%	0.31%	0.00%	0.00%	1.38%	0.92%	0.00%	8.60%	3.84%
Quid Pro Quo	26.42%	25.65%	11.67%	0.15%	0.31%	0.00%	34.10%	26.11%	0.00%	15.05%	29.95%	2.00%	19.66%	24.12%
Tailgating	0.15%	[illegible]	0.15%	2.00%	0.00%	1.38%	0.00%	15.05%	1.38%	[illegible]	0.92%	3.38%	0.61%	[illegible]
Impersonation	44.09%	4.45%	30.72%	0.77%	0.92%	0.00%	29.95%	2.00%	0.92%	3.38%	[illegible]	6.30%	41.01%	5.22%
Emotional Manipulation	31.49%	[illegible]	31.03%	3.23%	8.60%	3.84%	19.66%	24.12%	0.61%	[illegible]	41.01%	5.22%	[illegible]	[illegible]

Our analysis reveals that LLM-generated attacks demonstrate more sophisticated combinations of attack types and contextual elements. This advanced integration of tactics and contextual awareness indicates a concerning evolution in phishing capabilities through LLM automation.

Proposed Defense Mechanisms

To combat the misuse of LLMs for spear phishing, we propose a **two-layered defense system**:

1 Malicious Prompt Detection

- **Objective**: Detect and block malicious prompts before they generate phishing content.
- **Next Step:** Test prompts across multiple LLMs to evaluate their moderation practices.

2 Spear Phishing Email Detection

- **Objective**: Secondary defense layer for catching malicious content that evades prompt detection
- **Approach:** Focus on identifying phishing indicators such as urgency, emotional manipulation, etc.

Conclusion

Our study reveals significant vulnerabilities in state-of-the-art language models, showcasing their potential to generate **highly personalized spear phishing attacks** with minimal effort.
Implying an urgent need for **robust moderation systems** to prevent malicious use of these models.