POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Artificial Scams: On the risks of Fully Agentic Spear Phishing

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Manuela Maroni

Advisor: Prof. Stefano Longari

Co-advisors: Prof. Michele Carminati, Francesco Panebianco

Academic year: 2024-2025

1. Introduction and motivation

Phishing remains one of the most widespread and successful forms of cyberattack, relying on social engineering and persuasion techniques rather than technical abilities. While traditional campaigns rely on generic templates, spear phishing attacks leverage information on the target to increase credibility. Large Language Models (LLMs) have further transformed this landscape by enabling the automated generation of coherent, grammatically correct, and personalized phishing messages at scale.

Recent studies show that LLMs can produce highly convincing phishing emails and may enable partially or fully automated spear phishing campaigns [3]. The dual nature of these models has also been highlighted in research exploring both offensive and defensive applications of LLMs in phishing contexts [4]. However, empirical evidence is limited on the effectiveness of LLM-driven spear phishing when operating under realistic attacker constraints, particularly when starting from minimal initial information such as the email address.

This thesis addresses this gap by designing and evaluating a fully automated multi-agent system that generates personalized phishing emails using only the target's email address and publicly available information. The goal is to assess whether reconnaissance and persuasive email generation can be integrated into an end-to-end pipeline, and whether such automation significantly increases phishing effectiveness compared to generic campaigns.

Adopting this realistic threat model introduces several challenges: inferring identity from an email address is inherently ambiguous, publicly available information may be sparse or incomplete, and real-world email delivery constraints can affect experimental measurement. This work systematically addresses these challenges, providing a quantitative evaluation of LLM-based spear phishing while raising awareness and educating participants.

The main contributions of this work are:

- A fully automated multi-agent system that generates personalized phishing emails starting from the target's email address,
- A tool that generates personal reports containing the information inferred by the pipeline at the request of study participants for awareness and educational purposes,
- An empirical evaluation demonstrating that LLM-generated personalized phishing emails are more effective than generic ones and a quantitative analysis on the effect of different personalization factors.

2. Background

Unlike purely technical exploits, phishing primarily relies on social engineering and psychological manipulation. Classic research by Cialdini [2] identifies key persuasion principles, namely authority, scarcity, reciprocity, consistency, liking, unity, and social proof, that systematically influence human decision-making. These principles are widely exploited in phishing emails to induce urgency, trust or perceived legitimacy. Phishing success is strongly correlated with the strategic use of these techniques, especially when combined with contextual cues that increase credibility.

LLMs have introduced new capabilities for generating persuasive and personalized phishing content. Heiding et al. [4] compare LLM-generated phishing emails with manually crafted ones, showing that the former can achieve competitive or higher Click-Through Rates (CTR). Additionally, the same models demonstrate strong performance in detecting phishing intent, highlighting the double nature of LLMs in this domain.

Continuing in this line of work, Heiding et al. [3] evaluate the capability of LLMs to launch fully automated spear phishing campaigns validated on human subjects. Their results indicate that AI-generated personalized emails can match the effectiveness of human experts in terms of CTR, suggesting that automation does not necessarily reduce attack quality. The study also examines the reconnaissance phase, showing that automated profiling can successfully extract background information in a large majority of cases. These findings demonstrate that LLMs can support end-to-end spear phishing workflows under semi-realistic conditions.

Bethany et al. [1] explore the scalability of LLM-driven phishing by conducting a large-scale campaign targeting thousands of individuals within a university setting. Their work highlights how LLMs can be integrated into automated pipelines capable of generating tailored messages at scale, while also evaluating the limitations of existing detection infrastructures against such attacks.

Together, these studies show that LLMs substantially reduce the cost and effort required to conduct effective spear phishing campaigns. However, they assume access to structured background information or at least basic knowledge about the target. The question of how effective a fully automated spear phishing system can be when starting from minimal input, such as only the email address, remains underexplored, motivating the research presented in this thesis.

3. Proposed approach

The proposed approach automates the entire spear phishing workflow, integrating reconnaissance, interests inference, and persuasive message generation into a single end-to-end process. Rather than treating phishing simply as a text-generation problem, we model it as a multi-stage pipeline that begins with minimal input, i.e. the target's email address, and progressively enriches it using publicly available information. At a high level, the system mirrors the operational logic of a real attacker. It first infers possible identities associated with the email address, then collects information about social media profiles from publicly accessible online sources, extracts relevant data from them, and generates a phishing email tailored to the inferred profile. The entire workflow is designed to be operated autonomously and to rely exclusively on information gathered through Open Source INTelligence (OSINT) techniques, ensuring that it reflects realistic external attacker capabilities without assuming privileged access. Furthermore, the division into logically independent steps allows us to analyze, test, and validate each phase separately as well as to improve interpretability. Beyond message generation, the approach integrates a controlled delivery and measurement framework. Generated emails are distributed through a dedicated sending module and URL clicks redirect recipients to dedicated landing pages that record interaction events. Additionally, an on-demand report mechanism is included to transparently demonstrate how publicly available information can be aggregated and exploited. This transforms the system from a purely offensive simulation into a research tool that also supports awareness and defensive considerations. The entire workflow for personalized phishing emails is shown in Figure 1.

The personalized workflow is complemented by a traditional phishing baseline, which shares the same delivery and logging infrastructure but

relies on non-personalized, real-world phishing content. This baseline enables a controlled comparison between generic and automated personalized phishing under consistent experimental conditions.

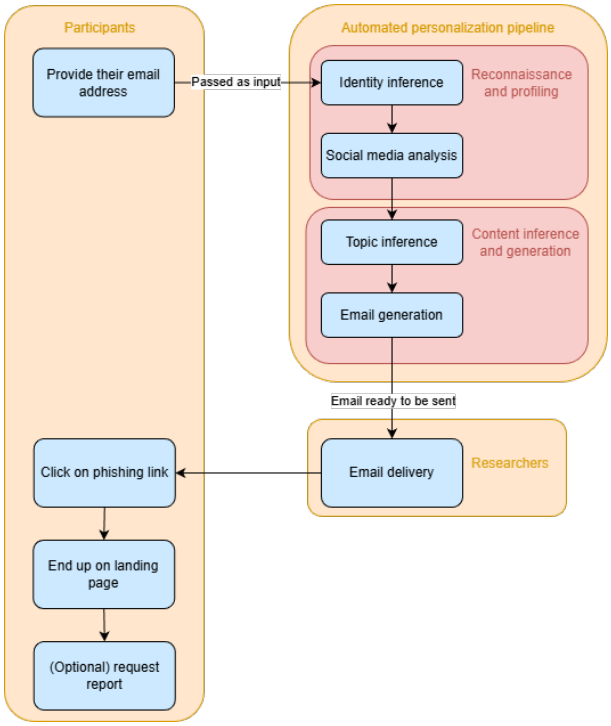

Figure 1: Personalized phishing workflow.

4. Implementation

The proposed approach is implemented through a multi-agent system that orchestrates LLM-based reasoning with deterministic processing components and web automation tools.

The system is composed of dedicated agents, each responsible for a specific operational task. The flow of these agents, which we are about to describe, is shown in Figure 2. The `EmailAddressAnalyzer` agent analyzes the structure of the input email address and generates candidate names and potential organizational affiliations. The `SocialsFinder` agent performs automated search queries and selects candidate social media profiles consistent with the inferred identity. `InstagramScraper` and `LinkedinScraper` extract publicly available information from the social media profiles previously identified using browser automation techniques. When multiple candidate identities are identified, the `NameValidator` agent evaluates consistency between the inferred names and the scraped content to select the most plausible profile. The `TopicFinder` agent processes the extracted data and identifies recurring themes, professional contexts, or personal interests to select a relevant topic, a persuasion strategy and the language in which to write the message. Finally, the `EmailGenerator` agent writes a coherent phishing email aligned with the selected topic.

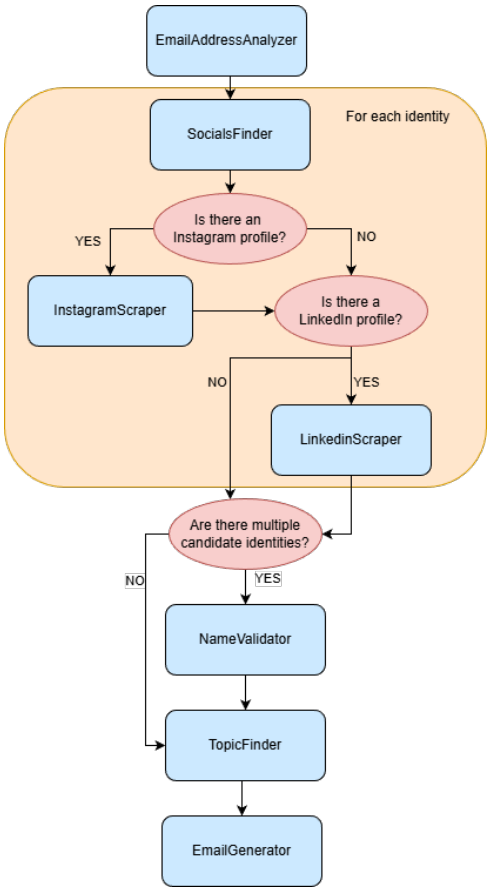

Figure 2: Agents' flow for personalized email generation.

Agent coordination is handled through a graph-based orchestration framework that enforces sequential execution, which mirrors the dependencies of each step with the previous ones, and structured data exchange between components. Each agent operates through controlled system prompts and produces machine-readable outputs, reducing ambiguity and limiting uncontrolled error propagation.

Outside of the email generation pipeline, additional modules support experimental deployment. A dedicated email delivery script sends messages via authenticated API access and embeds unique identifier tokens in the phishing

links. A landing page and logging infrastructure record click events, with interaction metadata, and report requests. A separate report generation module re-executes the pipeline upon user request and produces a structured summary of the inferred personal information and its potential exploitation in social engineering attacks.
The system is deployed in a controlled laboratory environment with temporary data retention, automated deletion of intermediate artifacts, and safeguards to avoid access to restricted information.

5. Experimental evaluation

The proposed system was evaluated through structured experiments designed to asses both the technical reliability of the individual components and the real-world effectiveness of fully automated personalized phishing. The evaluation combines offline validation of the pipeline's agents with a controlled human-subject experiment comparing personalized and traditional phishing emails.

5.1. Component validation

Before involving human participants, each agent of the automated pipeline was validated independently to evaluate its reliability and identify potential weaknesses. Validation experiments were carried out offline using two datasets: a subset of the Enron dataset [5], which provides structured corporate email addresses, and a custom dataset built from publicly available personal websites and blogs containing email addresses, full names, and links to associated social media profiles.
The `EmailAddressAnalyzer` agent was tested on 2000 internal Enron addresses (emails with `@enron.com` domain), 1821 external Enron addresses, and 152 addresses from the custom-built dataset. Results show a strong dependency on address structure: Top-N accuracy reaches 90.70% for internal Enron addresses, but drops to 46.68% for external Enron addresses and 55.92% for personal email addresses. These results demonstrate that the agent performs reliably in structured environments, whereas performance on external and personal email addresses is inherently limited by their higher ambiguity and lack of standardized naming conventions.
The performance of the `SocialsFinder` agent varies across platforms: LinkedIn achieved the best results, with 56.68% of correctly identified profiles and relatively low false positives, while Instagram and X show moderate correct identification rates (35.38% and 33.57%, respectively) with non-negligible wrong identifications. Facebook presents the highest ambiguity, with only 11.55% of correctly identified profiles and a significant rate of incorrect attributions. These results highlight the intrinsic difficulty of reliably linking publicly available information to a specific individual, especially in the presence of common names and incomplete profile data.
Context extraction through `InstagramScraper` is technically robust: profiles were successfully scraped in 96.28% of cases, bios were available in 75.60% of profiles, and image descriptions were generated in 96.82% of posts. This ensures that, once a correct profile is identified, sufficient information is usually available for topic selection.
The `NameValidator` agent correctly selected the target full name in 50.66% of cases, highlighting the difficulty of identity disambiguation. The `TopicFinder` agent was evaluated both against predicted and ground-truth identities. It generated relevant topics in 85.50% of cases when evaluated against the profiles identified by the pipeline, but this value drops to 50.50% when compared to the ground truth profiles, quantifying the impact of previous agents' identification errors. Finally, the `EmailGenerator` agent produced well-written and coherent emails in 93.50% of cases, with correct name usage in 78.50%, the latter being primarily limited by name selection errors in previous agents.
This validation phase confirms the technical feasibility of the automated pipeline while highlighting structural limitations inherent to OSINT-based identity inference.

5.2. Personalized vs. generic phishing

To assess the real-world impact of automated personalization on phishing attacks, we conducted a controlled human-subject experiment involving two distinct student population. The personalized phishing group consisted of 28 Computer Science and Engineering students who voluntarily participated in the study through a recruitment form, while the generic phishing group included 28 Engineering students

selected from the institutional phishing awareness program. Each participant received a single phishing email: the experimental group received individually generated personalized messages, whereas the control group received a single non-personalized email derived from a real-world phishing example. Effectiveness was primarily measured using the CTR, complemented by secondary indicators of engagement such as email replies and automated URL inspections.
The results, summarized in Table 1, show a substantial increase in the engagement for personalized emails compared to the generic baseline. The personalized campaign achieved a CTR of 28.57%, compared to 3.57% in the generic group. Personalized emails also triggered additional forms of interaction, including three email replies, three explicit automated inspections of embedded links, and two report requests. Although replies in some cases complained about personalization inaccuracies, they also indicate a deeper engagement with the content. No replies were observed from participants that received the generic email, and only one participant inspected the link with an automated URL checker.
These findings show that automated personalization significantly increases user interaction, even within a technically aware population.

5.3. Personalization factors analysis

To better understand which elements of personalization influence behavior the most strongly, we performed a subgroup analysis within the personalized phishing experiment (results in Table 2). The analysis focused on two core factors explicitly handled by the pipeline: the correct usage of the recipient's name and the relevance of the selected topic with respect to the recipient's publicly available interests.
The results reveal a clear layered effect of personalization. Correct use of the name appears to be a necessary credibility condition: emails containing an incorrect recipient name resulted in no clicks, while 34.78% of participants who were correctly addressed in the email clicked on the link, indicating that even minor identity inconsistencies can immediately undermine trust. Topic relevance had an even stronger impact on engagement. Messages aligned with the recipient's inferred interests achieved a CTR of 50.00%, compared to 7.14% for non-relevant or generic topics. This difference highlights that, while surface-level identity cues prevent early rejection, semantically meaningful personalization actively drives the interaction with the phishing links.
These findings suggest that effective spear phishing operates on multiple levels: basic identity consistency establishes plausibility, while topic relevance increases persuasive power.

6. Conclusions

This thesis demonstrates that large language models can be integrated into a modular, fully automated spear phishing pipeline capable of significantly increasing attack effectiveness through personalization starting from only the email address. The system shows that reconnaissance, identity inference, interest extraction and phishing email generation can be combined into an end-to-end process that requires no human intervention. Component-level validation shows that linguistic generation quality is consistently high, while identity disambiguation and personal interests inference remain the main technical challenges. Controlled human-subject experiments confirm that personalized LLM-generated phishing emails achieve a substantially higher engagement than generic messages, even within a technically aware student population.
Several limitations must be acknowledged. The participant pool consisted mainly of students from a technical university, possibly leading to a conservative estimate of real-world susceptibility. Moreover, the pipeline relies exclusively on publicly available data, making its performance dependent on the richness and accuracy of an individual's digital footprint. Finally, ethical and legal constraints required debriefing and transparency measures that partially reduced the realism of the simulated attacks. Real-world adversaries would not operate under such constraints and may exploit additional sensitive or semi-private information.
These findings underline important security implications. The automation of reconnaissance and persuasive content generation lowers the cost of spear phishing and increases its scalability. As AI-generated phishing messages become more polished and tailored to targets' interests,

Human-subject experiment results

Metric	Personalized phishing	Generic phishing
Participants	28	28
Emails received in inbox	28	27
Unique clicks	8	1
Click-through rate (CTR)	**28.57%**	**3.57%**
Email replies	**3**	**0**
Automated link inspections	**3**	**1**
Report requests	**2**	**N.A.**

Table 1: Personalized vs. generic phishing email effectiveness.

Subgroup analysis results

Personalization factor	Participants	CTR	Email replies
Correct recipient name used	**23**	**34.78%**	**2**
Incorrect name	5	0%	1
Relevant topic selected	**14**	**50.00%**	**1**
Non-relevant or generic topic	14	7.14%	2

Table 2: Impact of individual personalization factors on click-through rate.

traditional detection strategies based on superficial textual anomalies are likely to become less effective. Future research should therefore extend the population considered in this study to a more diverse demographic and improve identity disambiguation and interest inference mechanisms, while simultaneously advancing defensive techniques capable of detecting highly personalized AI-generated phishing content. Proactive investigation of offensive automation remains essential to anticipate evolving threats and to design robust, adaptive cybersecurity defenses.

References

[1] Mazal Bethany, Athanasios Galiopoulos, Emet Bethany, Mohammad Bahrami Karkevandi, Nishant Vishwamitra, and Peyman Najafirad. Large language model lateral spear phishing: A comparative study in large-scale organizational settings. *arXiv preprint arXiv:2401.09727*, 2024.

[2] R.B. Cialdini. *Influence: The Psychology of Persuasion*. Collins Business Essentials. HarperCollins e-books, 2009.

[3] Fred Heiding, Simon Lermen, Andrew Kao, Bruce Schneier, and Arun Vishwanath. Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects. *arXiv preprint arXiv:2412.00586*, 2024.

[4] Fredrik Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter S Park. Devising and detecting phishing emails using large language models. *IEEE Access*, 12:42131–42146, 2024.

[5] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer, 2004.