# A Large-Scale Study of Personalized Phishing using Large Language Models

Stefan Czybik
*BIFOLD & TU Berlin*

Anne Josiane Kouam
*Inria & TU Berlin*

Peter Heubl
*Ruhr-University Bochum*

Jan Magnus Nold
*Ruhr-University Bochum*

Konrad Rieck
*BIFOLD & TU Berlin*

## Abstract

Large Language Models (LLMs) can generate fluent and persuasive text, making them valuable tools for communication. However, this capability also renders them attractive for malicious purposes. While several studies have shown that LLMs can support generic phishing, their potential for *personalized attacks* at scale has not been explored and quantified yet. In this study, we thus evaluate the effectiveness of *LLM-based spear phishing* in an experiment with 7 700 participants. Using the target email addresses as queries, we collect personal information through web searches and automatically generate emails tailored to each participant. Our findings reveal a concerning situation: LLM-based spear phishing almost triples the click rate compared to generic phishing strategies. This effect is consistent, regardless of whether the generic emails are written by humans or generated by LLMs as well. Moreover, the cost of personalization is minimal, with approximately \$0.03 per email. Given that phishing is still a major attack vector against IT infrastructures, we conclude that there is a pressing need to strengthen existing defenses, for example, by limiting publicly available information linkable to email addresses and incorporating personalized phishing into awareness trainings.

## 1 Introduction

Large Language Models (LLMs) have enabled remarkable progress in natural language processing in recent years and are at the core of widely used services such as chatbots [11], online translators [2], and search engines [33]. Their ability to produce fluent and coherent text has set a new standard for human–machine interaction [34]. However, these same capabilities also make LLMs attractive to adversaries, since generating persuasive language aligns closely with techniques used in social engineering. Phishing is a prominent example of this threat, whose success depends on crafting credible emails that establish trust and prompt user interaction, such as clicking a link.

Several studies have explored how LLMs can enhance phishing strategies, either by automatically generating convincing email content [e.g., 6, 17, 24, 28, 30, 32] or by refining the language of existing campaigns [e.g., 1, 14]. These efforts have primarily focused on *generic phishing*, in which a single message is broadcast to many recipients. In contrast, *spear phishing*, where each message is tailored to its recipient, has received little attention in the context of LLMs so far. In practice, however, personalized attacks are markedly more effective: Although spear phishing accounts for only 0.1% of all phishing emails, it is linked to 66% of data breaches, according to a recent study by Barracuda Networks [5].

Given this context, the capabilities of LLMs raise a critical question: *Can these models enable personalization in large-scale phishing campaigns?* A few recent studies have begun to explore this possibility, typically through isolated demonstrations involving high-profile individuals, such as politicians or celebrities [16, 25]. While these examples demonstrate the feasibility of personalization, they offer no evidence that such attacks can be conducted automatically at scale for a broader group of victims. As a result, the threat posed by LLMs-based personalization and its quantitative impact on the security of email users remain largely unexplored.

In this study, we address this gap and explore the capabilities of LLMs in conducting large-scale spear phishing. Specifically, we perform an experiment with 7 700 participants to compare the effectiveness of personalized phishing with that of traditional phishing approaches. To enable automatic personalization, we use the participants' email addresses as queries in web searches and retrieve available public information. This data is then used by LLMs to generate user profiles and compose tailored emails to each recipient.

To ensure ethical compliance, our experiment is embedded within the phishing awareness training of a university and conducted in close coordination with the institution's data protection office, security team, and administration. The results reveal a concerning situation: LLM-based spear phishing almost triples the click rate compared to generic phishing. While untargeted emails achieve a success rate of 3.9 %, automatically personalized messages reach an average of 10.0 %. This effect is consistent, regardless of whether the generic emails are written by humans or also generated by LLMs. Moreover, emails personalized by humans achieve a success rate of 24.2 %, suggesting that the effectiveness of LLM-based attacks may even improve further with model capabilities.

To contextualize our findings, we assess the quality of the generated profiles and estimate the cost of automated spear phishing. Users with a low to medium amount of public information yield the highest click rates. Furthermore, the cost of personalization using LLMs is minimal, at $0.03 per email. We require a budget of only $150 to send personalized emails to all 3 310 users with publicly available information.

Given that phishing is still a major attack vector against IT infrastructures, we argue that improved countermeasures are urgently needed. However, detecting phishing emails remains a notoriously hard problem due to the constant evolution of attack strategies and the mimicry of legitimate communication [1, 7, 10, 23]. This highlights the need to go beyond purely technical detection and to adopt a layered defense strategy. Specifically, we recommend strengthening complementary measures in three directions: reducing the availability of linkable personal data, increasing user awareness of personalized phishing, and mitigating the impact of human errors through technical safeguards.

In summary, this paper provides a systematic and large-scale study of LLM-based spear phishing, making the following major contributions: <

- **Large-scale evaluation:** We present the first study to move beyond anecdotal evidence, quantifying the threat of LLM-generated spear phishing through a field experiment involving over 7 700 recipients.
- **Comparative analysis:** We compare LLM-based spear phishing against both generic LLM-generated and human-authored phishing emails, showing that personalization yields 2.74 higher click rates than traditional methods.
- **Insights and recommendations:** We identify key factors that influence attack success by analyzing user profiles and associated attack costs. Based on these findings, we recommend different defense strategies, including limiting publicly available information and enhancing user training to better address personalized threats.

**Roadmap.** We review previous studies on LLM-based phishing in Section 2. Our research scope and the workflow for generating personalized emails are presented in Section 3 and Section 4, respectively. The study setup, including ethical and technical constraints, is detailed in Section 5. We present our findings in Section 6 and discuss insights and recommendations in Section 7. Limitations are addressed in Section 8, and we conclude in Section 9.

## 2 Previous Studies

Research on LLM-based phishing has rapidly expanded in recent years. To structure our discussion, we categorize prior work using four criteria that reflect their methodological approach to phishing and its evaluation (see Table 1):

C1 **Personalization of emails:** This criterion indicates whether the study considers personalized phishing emails using LLMs instead of generic attacks.

C2 **Large-scale evaluation:** This criterion indicates whether the study includes a real-world experiment with a substantial number of recipients (typically over 1 000).

C3 **Baseline generic phishing:** This criterion assesses whether LLM-generated emails are compared against manually crafted emails for generic phishing.

C4 **Baseline spear phishing:** This criterion assesses whether LLM-generated emails are compared against human-written spear phishing emails.

Table 1: Comparison of related work on LLM-based phishing.

| **Study** | **Participants** | **C1** | **C2** | **C3** | **C4** |
|---|---|---|---|---|---|
| *Generic phishing* | | | | | |
| Heiding et al. [17] | 112 | ✗ | ✗ | ✓ | ✗ |
| Weinz et al. [32] | 36 699 | ✗ | ✓ | ✓ | ✗ |
| Sniegowski [30] | 23 743 | ✗ | ✓ | ✓ | ✗ |
| Bethany et al. [6] | 9 129 | ✗ | ✓ | ✓ | ✗ |
| Olea et al. [24] | 160 | ✗ | ✗ | ✓ | ✗ |
| *Spear phishing* | | | | | |
| Hazell [16] | – | ✓ | ✗ | ✗ | ✗ |
| Roy et al. [28] | – | ✓ | ✗ | ✗ | ✗ |
| Qi et al. [26] | – | ✓ | ✗ | ✓ | ✗ |
| Pourabbas Vafa et al. [25] | – | ✓ | ✗ | ✓ | ✗ |
| Our study | 7 741 | ✓ | ✓ | ✓ | ✓ |

**Generic phishing.** Several studies have investigated how LLMs enhance generic, untargeted phishing attacks through controlled experiments and large-scale deployments.

One of the earliest studies by Heiding et al. [17] conducts a controlled lab experiment with 112 participants, comparing emails generated by GPT-4, human-written messages, and hybrid versions. The study shows that LLM-generated emails are significantly more convincing than standard templates, achieving click rates of up to 44%. However, emails crafted by humans using psychological manipulation techniques still perform better, reaching up to 79%. Similarly, Olea et al. [24] conducts a user study focusing on perception of phishing emails. Involving 160 undergraduate students, the study shows that LLM-generated emails are often perceived as legitimate and hence sufficient for phishing campaigns.

Moving beyond the lab, Weinz et al. [32] sends phishing emails generated by LLMs across organizations of varying sizes. Their results demonstrate click rates exceeding 30%, particularly in smaller companies, with LLM-based messages in some cases outperforming traditional phishing formats. In a similar vein, Sniegowski [30] explores fine-tuned LLMs and finds that their outputs often appear more credible due to contextual relevance and stylistic polish. Bethany et al. [6] extends these findings to a university setting, showing that LLM-generated lateral phishing emails can match the effectiveness of messages crafted by humans.

Collectively, these studies provide strong evidence that LLM-generated phishing is effective at scale and represents a viable attack. However, the studies consistently rely on static email templates, lacking user-level personalization (C1).

**Spear phishing.** A second line of work has thus started to investigate the use of LLMs for spear phishing, where emails are personalized and tailored to each victim individually.

Hazell [16] are among the first to generate targeted phishing emails for over 600 members of the UK Parliament using data scraped from Wikipedia. Their study highlights the threat of such attacks, with each email costing only a few cents. However, the work remains conceptual, lacking real-world experimentation, as the generated emails are not sent to the intended targets. Extending this work, Qi et al. [26] propose a two-LLM pipeline, where one model generates phishing emails and a second model acts as a critic, iteratively refining the messages until they are no longer flagged as suspicious. Their evaluation combines automated filters and human raters, showing that the generated emails are both highly deceptive and readable. Likewise, Pourabbas Vafa et al. [25] use Instagram data from 200 public accounts to craft manipulative emails that reflect personal interests, relationships, and behavioral cues. Their results show that LLM-generated emails outperform real-world phishing in both emotional complexity and personalization. Finally, Roy et al. [28] examines how four commercial LLMs can be prompted to generate phishing emails and websites, demonstrating that these models can imitate trusted brands and implement evasive tactics.

While these studies demonstrate progress toward scalable spear phishing, they stop short of measuring the effectiveness of personalized emails under realistic conditions and at large scale (C2). Moreover, none compare the success of LLM-generated emails with that of traditional, manually crafted spear phishing (C4).

**Positioning.** To the best of our knowledge, we are the first to evaluate the efficacy of personalized phishing at scale and to compare it against other phishing strategies, including manually written emails. This closes an important research gap and provides empirical evidence of how LLMs can amplify the threat of phishing.

# 3 Research Scope

Before presenting our study and its technical workflow, we first outline the general research questions, threat model, and methodological approaches underlying our experiments.

## 3.1 Research Questions

Our study investigates the real-world impact of LLM-generated spear phishing at large scale. To this end, our analysis is guided by the following research questions:

**RQ1** Can LLMs personalize phishing emails at scale using only the victims' email addresses as reference?

**RQ2** How effective and scalable is LLM-generated spear phishing relative to manually crafted attacks?

**RQ3** How does the amount of publicly available information influence the success of LLM-based spear phishing?

To answer these questions, we design and conduct a real-world phishing experiment in collaboration with a partner university. Our approach relies on LLMs to generate emails using only publicly accessible information associated with the victims' email addresses. We begin by formalizing the underlying threat model, specifying the attacker's capabilities and constraints. Building on this model, we then introduce the phishing approaches evaluated in our study, including both LLM- and human-generated baselines.

## 3.2 Threat Model

We consider a threat scenario in which an adversary launches a spear phishing campaign targeting many members of an institution. The attacker's goal is to maximize user engagement by leveraging publicly available personal information and LLMs to craft tailored phishing emails that mimic the tone and context of legitimate communication. Although such attacks may ultimately aim at credential theft, malware delivery, or infrastructure compromise, our study focuses on initial user engagement measured through click rates. Consequently, we assume the attacker possesses the following capabilities:

- **Email discovery:** The attacker can identify and collect publicly listed email addresses of institutional employees, typically found on academic department pages, staff directories, mailing list archives, or event websites.

- **Web search:** The attacker can collect publicly available information about each victim through search engines, including professional biographies, affiliation details, personal activities, and other data for building a profile.

- **Local LLMs:** To maintain operational secrecy, the attacker avoids cloud-based LLM APIs (e.g., ChatGPT) and instead uses locally hosted LLMs running on private infrastructure to process data and get phishing content.

- **Email delivery:** Phishing emails are sent from outside the institution. To enhance credibility and impersonate legitimate entities, the attacker may employ lookalike domains or compromised accounts, reusing domains across victims to limit excessive registrations.

The attacker automates this workflow end-to-end: after gathering email addresses, they collect associated public information, build user profiles, and generate individualized phishing content. This enables the large-scale generation of personalized emails with minimal manual effort. A overview of this workflow is described in the following Section 4.

This threat scenario does not aim to benchmark LLMs or optimize attacks, but to evaluate what an attacker can realistically achieve under practical constraints. The adversary uses only open-source models on local infrastructure, avoiding advanced evasion or prompt engineering, yielding a conservative lower bound on the effectiveness of personalized attacks (see Section 8).

### 3.3 Phishing Approaches

As the basis for our study, we define a set of *phishing approaches* that span different levels of personalization, ranging from human-written generic emails to automated spear phishing with LLMs. For comparability, all approaches operate under the same threat model: the attacker is external to the institution, and emails originate from domains that appear credible but are not part of the institution. The approaches vary only in the target of the emails (generic vs. spear) and the level of automation (manual vs. LLMs). Technical details of these approaches are provided in Section 4, and their experimental setup is described in Section 5.

**LLM-based spear phishing.** This is the main condition of interest in our study, where we use an LLM-generated personal phishing email. To analyze the impact of personalization depth, we divide recipients into three subgroups based on the amount of public information used during generation.

**Manual spear phishing.** To contrast LLM-based personalization with human-crafted phishing, we manually generate spear phishing emails for a randomly selected set of recipients. These messages are composed in the same way an attacker might write targeted phishing emails by hand.

**LLM-based generic phishing.** In this baseline we prompt a language model to generate broadly applicable emails tailored to the institutional context and select the most convincing output based on clarity, plausibility, and contextual fit.

**Manual generic phishing.** Finally, we include manually written generic phishing emails derived from real-world campaigns as a traditional, low-cost attack strategy.

By examining these four approaches together, we can compare LLMs-generated with manually written emails as well as contrast generic with spear phishing within a single experiment, thereby addressing all criteria C1–C4 in Section 2.

## 4 LLM-based Spear Phishing

We implement a multi-stage workflow that mirrors how a real-world adversary would automate the generation of personalized phishing emails using LLMs. The workflow is designed to reflect a best-effort attack strategy using only publicly available information, general-purpose LLMs, and commodity infrastructure. Each stage of the workflow—from data collection to email generation—is designed for scale.

**Workflow overview.** Figure 1 illustrates the full workflow used in our experiment. We assume the adversary has already obtained a list of target email addresses. Then, the process consists of three main steps: (A) *data collection*, where publicly available information is gathered for each target; (B) *target profiling*, where the collected data is used to generate a brief persona for each individual; and (C) *email generation*, where a tailored phishing message is crafted.

Different workflow steps use different LLMs, selected from locally hosted models, including variants of Llama, Deepseek, Gemma, Mistral, and Phi. Model choice at each stage is based on a qualitative assessment of outputs for a sample of email addresses and corresponding runtime measurements. This implies a practical trade-off: larger models can produce richer outputs but incur higher computational cost, while smaller models are faster but may lose detail. A discussion and runtime reports are provided in Appendix D and the prompts in F. This balanced approach follows our threat model and avoids extensive benchmarking or calibration.

Furthermore, each step is executed in a batch-processing manner over all targets, allowing for parallelization and efficient resource use. To manage LLM interactions and orchestration, we use LangChain[1] together with locally hosted models via Ollama[2].

### 4.1 Data Collection

The first step of the workflow involves identifying and gathering publicly available information about potential targets. We assume an attacker that relies exclusively on open information, such as institutional websites, search results, and public directories to support later stages of the workflow.

**Web data extraction.** With a list of target email addresses, we begin by collecting personal or contextual information that could be used to personalize phishing content. For each email

[1] https://github.com/langchain-ai/langchain/
[2] https://ollama.com/

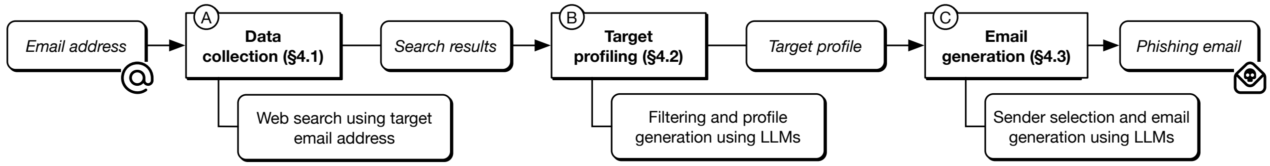


Figure 1: Workflow for LLM-based phishing: Given a target's email address, data is collected via web search. Local LLMs filter and summarize the retrieved content into a target profile, which is then used by other local LLMs to generate a personalized phishing email.

address, we query a search engine and download the content of the top-ranked pages using a lightweight HTML parser[3]. This approach enables fast, automated extraction of text without requiring JavaScript rendering or login access. While more advanced OSINT methods (e.g., scraping JavaScript-heavy platforms like LinkedIn or Facebook) could yield richer profiles, they introduce higher operational cost and greater risk of detection. In contrast, the use of publicly visible content provides a reasonable trade-off between scale and stealth.

For our experiments, we use the Google API to retrieve a list of webpages for a given target address. We find that some individuals have rich public profiles spanning several entries, whereas others appear in only a few. While dynamic filtering of results would be ideal, we adopt a fixed threshold and analyze only the first five entries, based on a preliminary experiment with 100 random email addresses. As we model a best-effort attacker, this choice establishes a conservative lower bound on the efficacy of personalization using LLMs.

## 4.2 Target Profiling

Once raw data is collected from public webpages, the next step is to extract personal details and generate user profiles. These profiles serve as input for subsequent LLM-based phishing email generation. This step is critical, as well-formed profiles directly affect the credibility and specificity of personalized phishing messages and thus the overall attack effectiveness.

**Information filtering.** The retrieved webpage content typically contains a mix of useful information and extraneous elements such as navigation menus, footers, and unrelated staff listings. To isolate target-specific data, we use an LLM to process each page and extract only information relevant to the individual. Each webpage is paired with the target's email address and provided to the model along with instructions to identify content corresponding to that person. The full prompt for this step is shown in Appendix F.

This simple approach supports several webpage formats, including profile pages, faculty directories, and departmental listings. As we use LLMs without specialized tuning, we rely on their native ability to resolve ambiguity and identify the most likely individual, even when multiple names or roles appear on the same page. To further aid disambiguation in cases involving common names, we flag pages where the target's email address appears directly in the text and assign such content greater weight during downstream processing.

To accommodate the LLM's context window limitations, we cap the input length at 4 096 tokens. Excess content is truncated from the end, which often contains boilerplate text such as legal disclaimers or navigation links. In our study, 96.61 % of the 20 050 webpages fall within this limit.

**Profile generation.** After extracting the relevant content, we use an LLM to synthesize a profile for each target. These profiles capture key personal and professional attributes such as name, interests, and affiliated institutions, which are intended to enhance the credibility of the phishing message. To prepare the input, we merge the extracted text segments from the selected webpages into a single block, ensuring that the combined content fits within the LLM's context window. The prompt for this generation is also shown in Appendix F.

While profile quality naturally varies depending on the amount and clarity of publicly available information, the selected model produces outputs that are generally coherent and plausible. Although some errors, such as minor blending of similarly named individuals or overly generic phrasing, do occur, the generated profiles remain sufficiently realistic.

## 4.3 Email Generation

The final stage of our workflow uses the profiles to generate spear phishing emails. This stage consists of two components: selecting a plausible sender identity including a sending domain, and composing a personalized email.

**Sender selection.** To enhance plausibility, the sender identity must align with the target's context. In our university-based scenario, we define five domain categories commonly associated with academic communication: *academic collaboration*, *conference invitations*, *training opportunities*, *institutional programs*, and *international exchange*. Each category is linked to one or more external domains matching its topic, such as an academic conference, training program, or research institution. Based on the domains, we define sender profiles,

[3] https://www.crummy.com/software/BeautifulSoup/

which are virtual identities containing full name, email address, affiliation, and job description. An LLM then selects the most suitable sender for each recipient using the target profile as input for the prompt from Appendix F.

The selected categories are sufficiently broad to plausibly engage staff across research, administration, and support roles in a university. When targeting other types of organizations, the adversary can define a corresponding set of categories.

**Phishing email generation.** Once a sender is selected, the final step is to generate a personalized email. The generation prompt includes the structured profile of the target as well as the selected sender identity. The topics of the emails were automatically selected by the LLM based on the victim and sender profiles. Given that both the sender identities and the publicly available profile information of the participants are rooted in an academic environment, the model predominantly produces university-related themes, such as conference invitations, collaboration requests, seminar announcements, or internal administrative matters. As a consequence, the distribution of topics reflects the communication norms of a university workplace and may differ from what an attacker would generate in a corporate, governmental, or non-academic setting. Nonetheless, this topic selection allows us to simulate how an adversary might tailor messages to the contextual cues obtained from publicly exposed data. This can create realistic spear-phishing scenarios within the university domain.

To ensure compatibility with our email delivery system, each message must include a placeholder token `{{link}}` indicating the phishing URL location. The placeholder will be replaced by the email delivery system and will consist of the domain of the selected sender and a 22-character random string for anonymous click rate collection.

In preliminary tests, we find that smaller open-source models often struggle to generate well-structured and contextually appropriate emails. To address this, we create a synthetic dataset consisting of fictitious user profiles and matching phishing emails, generated using GPT-4o. We then fine-tune local models on this synthetic dataset, with the goal of improving structural consistency and task adherence without altering the models' general language behavior. To achieve this, we employ low-rank adaptation [20], a parameter-efficient fine-tuning method.

This completes our workflow, from data collection to personalized email generation. Our approach reflects the capabilities of a scalable and practical adversary that relies solely on publicly available data and general-purpose LLMs. Instead of focusing on a single model, we consider several open-source options, such as Llama, Deepseek, Gemma, Mistral, and Phi, and select the most suitable one for each step based on a small sample of target emails. This strategy mimics an adversary with moderate resources who cannot afford an extensive, time-consuming comparison of models.

# 5 Study Design

Having described our LLM-based spear-phishing workflow, we now set the stage for our experimental study, outlining how we design and conduct the evaluation. We first describe our measures to address ethical requirements, then the experimental setup and methodology to assess phishing success.

## 5.1 Constraints and Ethics

Our study is conducted within a partner university (TU Braunschweig) and involves a diverse group of participants, including researchers, administrative staff, technical personnel, and working students. To ensure ethical and legal compliance, we collaborate closely with all relevant institutional stakeholders from the outset. Specifically, we work with the data protection and security offices of the partner university, as well as the data protection office of our own institution.

The security office of the partner university routinely conducts phishing awareness trainings formally approved by the university's administration and staff committee. We integrate our study into this process, so that participants are not only exposed to phishing emails but also receive educational material before the training and after a successful attack. This integration ensures compliance with all applicable regulations and strengthens the overall resilience of the university's employees against phishing attacks.

Furthermore, we follow best practices to minimize potential harm to participants. First, we implement stringent technical and organizational measures to ensure data security throughout our workflow. Second, all participants receive a briefing before the study, informing them about the phishing training, and a debriefing afterwards, clarifying the purpose of the study and reinforcing its educational intent. Although these briefings introduce a bias by making participants aware of potential phishing, they are mandated by the university's administration and personnel council.

All emails are manually reviewed to exclude inappropriate content, such as offensive, harmful, or controversial material. Across the entire study, this affected only four emails, which were removed. We suspect that the cause was hallucinations. Moreover, all landing pages shown after clicking a phishing link provide educational guidance on recognizing and avoiding phishing attempts. To protect privacy, participant clicks are analyzed in aggregate, preserving anonymity. Further ethical considerations are discussed in Appendix A.

## 5.2 Experimental Setup

We proceed to describe the practical setup of our study, starting with the identification and grouping of participants, followed by the deployment of the different phishing approaches.

**Participant groups.** Figure 2 shows the number of participants and their assignment to different groups. Participants are selected following the approach of a real-world attacker: We scrape publicly available institutional email addresses from the partner university, starting with staff directories and departmental websites. To broaden coverage, we generate additional candidate addresses by combining common names and validate them via search engine results. Using these heuristics, we identify 4 010 email addresses linked to public data out of 7 741 employees, corresponding to 51.8%. The remaining 3 731 employees lack public information and are therefore targeted only with generic phishing approaches in our study.

The availability of public information for an email address serves as the primary split criterion in our study design. Addresses without any online information are unsuitable for spear phishing and are therefore randomly divided into two equally sized groups for generic phishing: 1 866 participants are assigned to traditional manual phishing and 1 865 participants to LLM-based generic phishing.

The assignment of users with online information (w/ result) follows a structured process. First, we randomly select 100 email addresses for manual spear phishing to establish a direct human baseline. Second, we divide the remaining addresses into three subgroups of approximately equal size—*low*, *medium*, and *high*—based on the number of tokens in the generated user profiles. These subgroups form the basis for LLM-based spear phishing. Specifically, we define three quantiles (0–476, 477–656, and >656 tokens) to categorize the amount of available profile data.

To detect potential bias from the availability of public information, we randomly select 200 addresses within each subgroup as controls for generic phishing. These assignments, shown as smaller flows in Figure 2, enable us to assess potential differences in the effectiveness of generic phishing between users with and without public information. In total, we assign 3 310 participants to LLM-based spear phishing, 2 166 to the generic manual phishing variant, and 2 165 to the LLM-based generic variant.

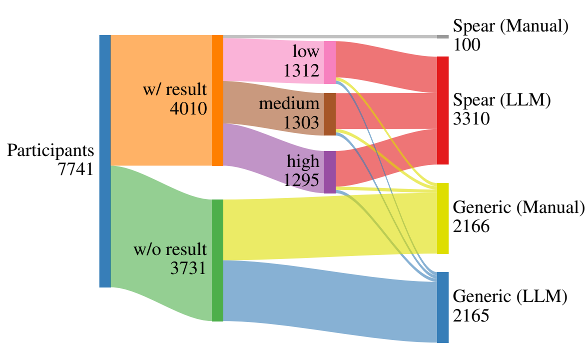


Figure 2: Distribution of email addresses to the groups.

**Manual spear phishing.** To establish a human baseline for spear phishing, we manually compose emails using exactly the same information available to the LLMs. For each of the 100 randomly selected targets, we follow a strict, time-limited procedure with two stages: research and writing. In the research phase, we allocate five minutes to collect publicly available information from five websites per target. In the subsequent two-minute writing phase, we compose the phishing email based on the collected information. The human author uses the same sender profiles as in the automated condition, resulting in similar message topics and contextual framing. In total, producing the full set of manually generated spear-phishing emails takes approximately 11 hours and 40 minutes. The task is performed by a senior researcher with expertise in phishing and security research.

**Manual generic phishing.** For generic manual phishing, we use a template previously employed in the awareness trainings of the partner university. It follows a common password-reset scam pattern targeting widely used services such as Outlook or WebEx. To ensure realism and comparability, we adapt the template to the university's popular file-sharing platform, Nextcloud, recreating a phishing attempt in the same style. The full email message is provided in Appendix H.

**LLM-based generic phishing.** For LLM-based generic phishing, we prompt GPT-4o to produce three candidate phishing messages. We select the message that best aligns with our experimental design: a survey invitation on the university's internal communication practices, a topic broad enough to plausibly engage the entire campus population. To maximize clarity and inclusiveness, the email is provided in both English and the local language. The exact prompt and model outputs are included in Appendix G.

**LLM-based spear phishing.** Finally, for spear phishing using LLMs, we follow the workflow detailed in Section 4 for each of the three subgroups.

## 5.3 Execution of Study

Our study begins with a briefing email sent to all participants as part of the regular phishing awareness training. This message includes links to training materials designed to improve phishing detection skills but deliberately omits the exact timing of the phishing campaign in the following weeks. Participants are instructed to delete any suspicious emails and to avoid clicking on embedded links, even if they suspect the messages to be part of the exercise. Phishing emails are then sent over a two-week period, starting one week after the briefing. All messages are sent during regular working hours. Finally, one week after the last phishing email is delivered, participants receive a debriefing message informing them that the exercise has concluded.

**Email delivery.** The phishing emails are sent from an external server, with sender domains configured to resolve to it. The server hosts a commercial platform used by the university for phishing awareness training. To avoid interference from automated defenses, the sending server is placed on the spam filter's whitelist, ensuring that no emails are flagged during delivery (see Section 7.2). Because the email addresses are obtained via scraping, we cannot verify in advance whether each one is still active, a limitation that equally applies to real-world attackers. Nonetheless, delivery reports confirm successful transmission for 89.5 % of all emails. In the subsequent analysis, we therefore use the *delivered emails* as the reference for the analysis of click rates.

**Landing page.** When participants click on a phishing link during the campaign, they are redirected to a landing page that explains the email is part of an awareness training and poses no security risk. The page provides educational material, along with information about the study. It also includes a link to a short user survey as described in the next paragraph.

**User feedback.** To complement the quantitative data, participants are invited via the landing page to complete a voluntary Qualtrics survey. The questionnaire assesses their perceptions of the phishing emails, awareness of social-engineering techniques, and self-reported reactions. It comprises three parts: (i) a plausibility scale measuring how credible participants find the emails in their work context; (ii) a social-engineering scale evaluating persuasive cues such as authority, urgency, or familiarity; and (iii) a motivation scale identifying factors participants believe drove them to click.

# 6 Results

We structure the results around our research questions. We first compare the effectiveness of phishing strategies, then analyze how the amount of public information influences attack success, present insights from the user survey, and finally assess the practical costs of executing attacks at scale.

## 6.1 Efficacy of Phishing Approaches

We start by analyzing the overall performance of the different phishing approaches in Table 2. A complete breakdown for all groups and subgroups is provided in Appendix C. LLM-based spear phishing substantially outperforms both generic variants, achieving a click rate of 10.0 %. This rate is nearly *three times higher* than the generic LLM baseline (3.7 %) and more than *twice as high* as the manually crafted generic phishing email (4.1 %). All differences are statistically significant ($p < 10^{-5}$), demonstrating that personalization via LLMs markedly increases phishing effectiveness under real-world conditions. Statistical significance was assessed using two-proportion $z$-tests on independent recipient groups, and all reported $p$-values refer to these tests. These findings directly answer RQ1, establishing a clear advantage of LLM-based personalization over generic approaches.

Table 2: Delivery and click performance across phishing approaches. More details with all results are presented in Appendix C.

| Campaign | Sent # | Delivered # | Delivered % | Clicked # | Clicked % |
|---|---|---|---|---|---|
| Spear (LLM) | 3 310 | 3 177 | 96.0 % | 330 | 10.0 % |
| Spear (Manual) | 100 | 99 | 99.0 % | 24 | 24.2 % |
| Generic (LLM) | 2 165 | 1 735 | 80.1 % | 65 | 3.7 % |
| Generic (Manual) | 2 166 | 1 918 | 88.6 % | 79 | 4.1 % |

When compared to manual spear phishing, however, the picture shifts. Human-crafted spear phishing achieves the highest click rate overall, at 24.2 %, corresponding to a *2.4× higher* success rate than the LLM-based variant (10.0 %). This difference is also statistically significant. While this underscores the superior effectiveness of human-authored spear phishing, such attacks do not scale: they require sustained manual effort and can only be carried out on a limited set of recipients. We return to this scalability aspect in Section 6.4.

**Effect of available information.** To examine how the amount of publicly available information affects phishing success, we analyze click rates across the previously defined subgroups. As shown in Figure 3, the results reveal an unexpected pattern: individuals in the high-profile group are significantly less likely to click on phishing links. In contrast, the low- and medium-profile groups exhibit comparable success rates (12.7 % and 12.1 %, respectively). This drop is statistically significant ($p < 10^{-5}$) and counterintuitive, as one might expect that richer public information would enable more convincing and effective personalized attacks. We investigate this discrepancy in more detail in Section 6.2.

No clear trend emerges for generic phishing, where all recipients receive the same non-personalized message. As shown in Figure 3, success rates range from 3.3% to 7.6% across groups. Participants with high-profile information show slightly lower click rates, suggesting they may behave differently, potentially due to greater phishing awareness or distinct email engagement patterns. However, this effect is modest and statistically non-significant ($p = 0.59$).

**Click rates over time.** Our study is embedded within a series of awareness trainings conducted at the partner university, enabling a direct comparison between our phishing attacks and earlier campaigns. Table 3 summarizes the corresponding results. In 2024, two prior campaigns were carried out:

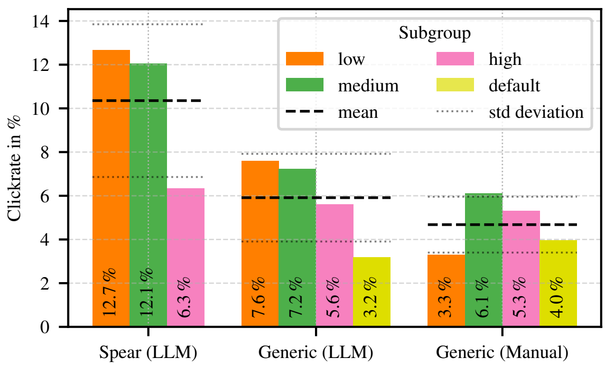


Figure 3: Results for the profile information subgroups. Horizontal dashed bars indicate the average click rate per phishing approach.

the first (OWA Password) targeted academic staff with a Microsoft Outlook Web Access reset scenario, while the second (Webex Password) focused on non-academic employees using a human-resource phishing lure. Both campaigns reported click rates between 5% and 7%, indicating a moderate level of susceptibility across groups.

Table 3: Click rates of different awareness trainings.

| | Delivered | | Clicked | |
|---|---|---|---|---|
| **Campaign** | **Target** | **#** | **#** | **%** |
| OWA Password (2024) | † | 3652 | 192 | 5.3 % |
| Webex Password (2024) | ∗ | 2607 | 182 | 7.0 % |
| Generic (Manual) | All | 1918 | 79 | 4.1 % |
| Generic (LLM) | All | 1735 | 65 | 3.8 % |
| Spear (LLM) | All | 3310 | 330 | 10.0 % |

† Scientific Staff; ∗ Non-Scientific Staff

Our manually crafted phishing email (Nextcloud Password) was closely modeled after the earlier awareness messages. Despite this similarity, it achieved a slightly lower click rate of 4.1 %, which may reflect improved employee awareness over time. However, other factors — such as a higher proportion of employees on vacation or inactive accounts — may also have contributed to the reduced engagement. A comparable trend is observed for the generic LLM-based phishing emails, which reached a click rate of 3.7 %. Together, these findings suggest that employees are becoming more proficient at detecting templated, generic phishing attempts, regardless of whether they are written by humans or LLMs.

In contrast, fully automated LLM-based spear phishing achieves a substantially higher success rate of 10.0 %, despite requiring no manual curation. This gap indicates that while traditional awareness training appears effective against generic phishing threats, personalized LLM-driven attacks introduce a new class of challenges that may be harder to anticipate and defend against.

## 6.2 Impact of Public Information

We uncover a counterintuitive pattern: individuals with the most publicly available information are *least* likely to fall for LLM-generated spear phishing emails. While one might expect richer profiles to enable more persuasive and tailored messages, the observed drop in success rates suggests that having more information does not necessarily translate into more effective attacks. To understand this discrepancy, we analyze how the LLM processes public information throughout the workflow, focusing on the transition from collected personal information to the final phishing emails. We introduce a set of metrics to capture different aspects of this generation process: content reuse of the LLMs across profiles and emails, as well as topical coherence between them.

**Token overlap.** We first investigate the extent to which the model reuses input text when constructing a user profile. To quantify this, we compute the cosine similarity between token frequency vectors of the filtered input text and the generated target profile. We deliberately rely on the filtered input rather than the raw webpages, as the latter often contain unrelated elements such as navigation menus, disclaimers, or other noise that could distort this metric. Higher similarity scores indicate stronger reuse of original content, whereas lower scores suggest greater abstraction or hallucination.

As shown in Figure 4(a), the overlap is highest for the *low-information* group, with a 60th percentile of 0.4, compared to 0.28 and 0.2 for the medium and high groups. This suggests that when little data is available, the LLM tends to reuse the same vocabulary from the input, producing profiles that are short but lexically consistent with the source. As the amount of input grows, however, the overlap in word usage decreases. This indicates that the model introduces more varied phrasing and shifts in emphasis, sometimes reformulating appropriately, but also at the risk of omitting key details or drifting toward speculative content.

**Information retention.** Next, we examine to what extent the content of a generated profile actually carries over into the final phishing email. Intuitively, a higher degree of reuse should strengthen personalization, as details about the target remain visible in the message. Conversely, lower reuse suggests that the LLM is rephrasing or abstracting more heavily, which could either improve fluency or dilute the connection to the target. To quantify this, we compute the fraction of profile tokens that also appear in the corresponding phishing email. This simple overlap ratio provides a lexical view of how directly the profile informs the generated message.

Figure 4(b) shows that retention is lowest for the high-information group (median 0.17), compared to 0.18 for both low and medium groups. While the numerical difference is modest, the trend aligns with the earlier findings on token overlap: when profiles are rich, the LLM tends to recontex-

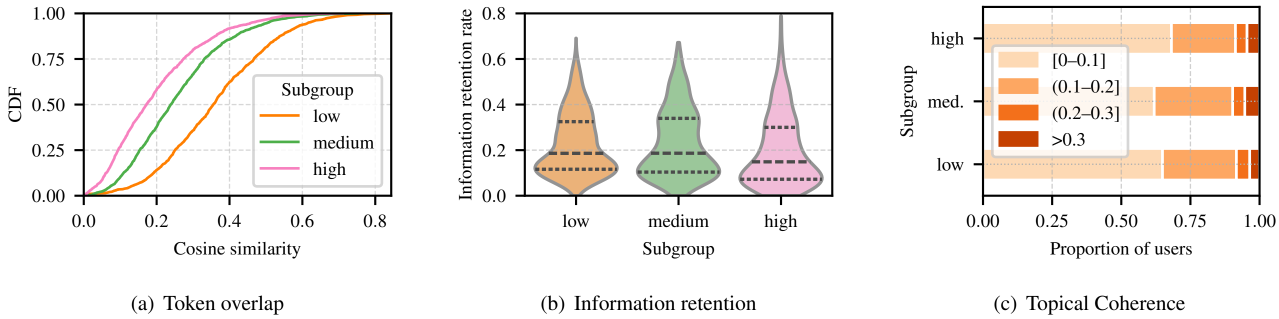


Figure 4: Distributions of interpretability metrics across low-, medium-, and high-information user groups.

tualize rather than copy directly. In practice, this means that abundant input does not translate into more explicitly personalized emails; instead, the model paraphrases and reshapes the content, which reduces the salience of individual details and may weaken the overall personalization effect.

**Topical coherence.** We finally investigate whether the phishing emails remain thematically aligned with their underlying user profiles. To measure this, we compute the Jaccard similarity over the 15 most frequent tokens in both the profile and the corresponding email, excluding stop words. Higher scores indicate stronger thematic consistency between the profile and the generated message, while lower scores suggest that the email construction is more elaborate and less directly grounded in the user-specific input.

As illustrated in Figure 4(c), the high group shows the weakest topical consistency: 70% of their emails have $[0-0.1]$ overlap, compared to 62% for medium and 65% for low. Topical coherence scores above 0.3 are rare across all groups, but particularly scarce for high users. These results indicate that richer profiles do not consistently produce emails that remain close in theme to the original profile. Instead, abundant input seems to push the LLM toward abstraction and topic shifting. By contrast, low- and medium-information inputs are more likely to yield emails that reuse key terms directly.

**Summary.** Profiles built on sparse input show higher token overlap and greater direct reuse in the final emails, which makes the resulting messages short, focused, and tightly tied to the few known facts. By contrast, richer inputs lead to more paraphrases, retain fewer explicit details, and drift thematically in the final email. As a result, the personalization becomes less concrete and the messages less convincing.

This highlights a limitation of current LLM-based spear phishing: simply increasing the amount of harvested data does not automatically yield more persuasive attacks. Instead, the way the model transforms and recontextualizes that information plays a decisive role. For attackers, this means that large-scale data collection may need additional curation or constraints to be effective. For defenders, it suggests that having extensive public information is not necessarily a vulnerability in itself.

### 6.3 User Perception and Feedback

Consistent with prior studies gathering voluntary data after phishing attempts [9, 22, 29], we observe a substantial dropout in our survey of participants: only 19 complete responses are recorded, corresponding to 3.82 % of phished users and 0.27 % of all delivered emails. Nevertheless, we briefly report descriptive insights, as they offer a useful glimpse into how participants perceived phishing emails. The full questionnaire is provided in Appendix E, and the responses are summarized in Appendix E.

A key observation is that most participants rated the LLM-generated emails as highly plausible and personalized, indicating that LLMs are capable of producing convincing phishing messages. This assessment holds for both the spear-phishing and generic variants, whereas our manually written generic email was perceived as less plausible, likely because it resembled messages used in previous awareness trainings. Interestingly, recipients of the LLM-based spear-phishing emails reported feeling only minimal pressure to click on the links, suggesting that these messages achieved persuasiveness without relying heavily on urgency cues. Due to the limited feedback, however, these insights are not statistically significant and need to be regarded as descriptive observations.

### 6.4 Scalability of Personalization

While Section 6.1 and Section 6.2 focused on the effectiveness of different phishing approaches, we now turn to their scalability. Specifically, we analyze the time, effort, and financial resources required to generate personalized phishing emails at scale. Our emphasis is on comparing manual and LLM-based spear phishing, as this directly addresses RQ2: whether personalization is not only more effective, but also practical to deploy at scale.

**Runtime of email generation.** Manual spear phishing is slow and resource-intensive, demanding substantial skilled effort. In our study, composing 100 personalized emails took nearly 12 hours, underscoring the limits of manual approaches. By contrast, our workflow for LLM-based spear phishing operates fully automatically. Table 4 shows the runtime of the employed LLMs. Note that we select different LLMs for the different steps to balance efficiency and quality, as described in Section 4. As a result, the reported runtimes are measured across different models. Generating 4 010 emails takes about 50.5 hours on a single NVIDIA A40 GPU, corresponding to roughly 45 seconds per email. This makes our approach nearly nine times faster than the human baseline. Moreover, since the workflow is fully parallelizable, attackers could easily scale throughput further to reduce processing time.

**Costs of email generation** In terms of financial cost, our workflow is efficient. In the first step, the adversary can leverage public search engines to collect relevant information. Using the Google Custom Search API[4], for example, allows up to 100 free queries per day; in our study, we spent a total of $20 to accelerate retrieval. Downloading all identified websites to a server incurs negligible costs. By contrast, expenses become non-trivial in the subsequent steps. LLM computations dominate the overall cost due to their substantial computational demands. In our experiments, we use a dedicated NVIDIA A40 GPU. Assuming a rental cost of $2 per hour, processing for 50.5 hours results in an estimated expense of roughly $100. Overall, generating 4 010 personalized emails thus costs approximately $120, corresponding to less than $0.03 per email, including search, inference, and generation.

Table 4: Total runtime across workflow steps for 4 010 recipients using our selected LLMs.

| Workflow steps | | LLM | Runtime |
|---|---|---|---|
| Step A | Data extraction | llama3.2:3b | 15:16:38 |
| Step B | Target profiling | deepseek-r1:14b | 14:21:56 |
| Step C1 | Sender selection | phi4:14b | 09:42:40 |
| Step C2 | Email generation | phi4:14b[1] | 11:00:19 |
| Total time | | | 50:21:34 |

[1] Finetuned

**Summary.** LLM-based spear phishing clearly surpasses manual efforts in both effectiveness and efficiency. With a modest financial investment of about $120, an attacker can achieve a 2.74 increase in phishing success rates for users with public information. Since a single compromise is often sufficient to infiltrate an IT infrastructure, spear phishing using LLMs becomes highly attractive and lucrative for malicious actors. Although our study focuses on a single institution, we conclude that automated spear phishing substantially increases the risk of security compromises and underscores the need for improved defenses.

[4] https://developers.google.com/custom-search/v1/overview

## 7 Discussion

Our empirical results indicate that spear phishing using LLMs is not merely an academic concern but a tangible threat to security that requires suitable countermeasures. We therefore discuss these implications in more detail and subsequently turn to possible defenses, focusing on both organizational and user-level measures.

### 7.1 Implications

Classic spear phishing is an effective attack vector for targeting high-profile individuals and enabling stealthy infiltration. If precision is not the primary concern, however, LLMs offer a viable alternative for large-scale campaigns. Our results demonstrate that LLMs can rapidly and inexpensively generate personalized phishing emails at scale. While these emails may not match the sophistication of attacks prepared over weeks through intelligence services, they enable adversaries to target thousands of individuals within an organization for only a few dollars. At the same time, these emails nearly triple the click rate compared to generic phishing, providing a middle ground between indiscriminate broadcast attacks and highly sophisticated compromises. Consequently, LLM-based spear phishing becomes particularly attractive for cybercriminals seeking broad success rather than highly targeted attacks.

Unfortunately, our workflow is straightforward to implement, suggesting that adversaries could readily integrate this technique into existing phishing infrastructures. With various open-source frameworks for LLMs inference available, the required technical effort remains moderate. Although our approach already achieves a substantial increase in click rates, further improvements are likely as more advanced LLMs become publicly available. From a technical perspective, we must therefore assume that such attacks can be executed with limited preparation and effort.

Upon reviewing the generated emails, we identify the primary limitation as the quality of the profile data. As we will show in Section 8.1, more effective profile generation could be achieved by leveraging commercial LLMs, albeit at the cost of reduced operational stealth. Even worse, the manually crafted spear-phishing emails achieve a high click rate of 24.2 %. If future LLMs increasingly match the effectiveness of skilled humans in the task of spear phishing, the associated threat will rise substantially. Consequently, the overall risk is likely to continue growing and will not diminish without explicit countermeasures.

### 7.2 Detection

A natural defense against malicious emails is the use of security appliances that filter incoming messages and detect potential threats. However, general detection remains inherently challenging because phishing emails closely mimic natural human writing. In our study, the university's commercial spam filter and unified threat management system flagged only 1 out of 3 949 emails when exposed to our attacks. The human-like quality of these emails, whether produced manually or by LLMs, makes them difficult to distinguish from legitimate communication. This highlights the difficulty of identifying malicious content based on linguistic cues alone [15].

However, detection becomes feasible once campaign-specific patterns emerge. In our case, for example, sender names could have served as a reliable signal for filtering. While general-purpose detection remains difficult, targeted countermeasures are effective when attackers reuse identifiable infrastructure. Yet, because adversaries can continuously adapt their strategies to evade such measures, spear phishing is likely to remain a persistent threat.

### 7.3 Recommendations

Given the difficulty of reliably detecting LLM-based spear phishing emails, alternative defense strategies must be prioritized. Because these attacks exploit both technical weaknesses and human cognitive biases, a layered strategy is required to reduce the susceptibility and impact of successful attacks.

**R1: Technical mitigation of attack impact.** Even with improved awareness and modern detection tools, some recipients will inevitably click on malicious links [8]. Technical controls therefore remain the most reliable and scalable protection against LLM-based spear phishing and phishing in general. Key measures include strict network segmentation to constrain lateral movement, enforcement of the principle of least privilege to reduce the value of compromised accounts, and application whitelisting or sandboxing for opening potentially dangerous content. Incident response and monitoring solutions can enable rapid identification and isolation of affected systems, limiting downstream damage. Multi-factor authentication should be enforced for all critical services, ensuring that credential compromise alone does not grant full access. By investing in these defenses, the likelihood that a single event will lead to a breach decreases, transforming phishing from a potentially catastrophic into a manageable incident.

**R2: Awareness of personalized phishing.** Since general prevention of phishing emails is unlikely, strengthening defenses on the recipient side is important. Raising awareness of personalized phishing tactics is a critical step toward reducing the likelihood of successful attacks. Prior work has shown that phishing awareness training can be effective [18, 27, 31], though other studies report limited impact [3, 12, 19]. Despite these mixed findings, awareness training can form a relevant component of a defense strategy, particularly in light of the limitations of automated detection mechanisms. As our study illustrates, the emergence of LLMs significantly increase the scalability of spear phishing, requiring awareness programs to address LLM-based techniques and risks. Educating employees about the plausibility and personalization of such attacks can help mitigate their success, though attackers may still adapt to evade detection.

**R3: Managing public-available personal information.** Traditional advice promotes minimizing public personal data; however, for many institutions—especially universities and public-sector organizations—removing names, roles, or email addresses is infeasible due to transparency and operational needs. Rather than full data obscurity, organizations should limit unnecessary personal details. While this cannot eliminate spear-phishing risk, it reduces structured, linkable information available for scalable abuse.

**Conclusion.** Together, robust technical controls, updated educational programs, and careful management of publicly exposed information are a realistic strategy against LLM-based spear phishing. While none of these measures alone can fully mitigate the threat, their combination significantly strengthens institutional resilience.

## 8 Limitations

Our study provides new insights into the effectiveness of LLM-based spear phishing, it has limitations due to its empirical nature. In the following, we discuss their implications for our findings.

### 8.1 Selection of LLMs

For all steps of our workflow, we employ local LLMs instead of more capable models accessible through online services. This decision follows from our threat model (Section 3.2) and the privacy requirements of the partner university (Section 5.1). As a result, the reported click rates represent a lower bound on the performance an attacker could achieve with more advanced models.

To gain an intuition of the difference between local and cloud models, we conducted a small trial with a select group of consenting individuals using ChatGPT. The generated profiles outperformed those produced by the local models, clearly indicating that spear-phishing attacks could be further improved if an adversary is willing to trade stealthiness for effectiveness. Furthermore, our study also includes manually crafted spear-phishing emails, establishing an upper bound on the potential performance of more advanced models. We conclude that our workflow reaches a middle ground in performance: It significantly improves attack success compared to traditional approaches but still falls short of highly refined strategies. Accordingly, we argue that the reported results offer an essential and timely perspective on the threat of automatizing phishing.

## 8.2 Sampling Bias

Our study is conducted at a university and targets participants with publicly available email addresses. This sampling strategy introduces potential bias, as we lack demographic information about our participants so that the observed click rates may not be generalizable to other institutions or user groups. However, De Bona and Paci [12] observe no clear relationship between demographics and susceptibility to phishing. Furthermore, our target university conducts regular phishing awareness training, which likely influences user behavior and may result in lower susceptibility compared to targets without such training. Consequently, our results cannot be readily extrapolated to other institutions or demographic groups easily.

Another source of potential confounding factors arises from the selection of email topics and sender profiles, which were all framed within a university-related context. This contextual alignment may influence participant expectations and plausibility judgments, thereby affecting click behavior in ways that may not transfer to settings with different organizational norms or communication cultures.

However, the general approach of spear phishing based on publicly exposed data is not specific to this setting. Since similar patterns of data exposure exist across many institutions and user groups, the fundamental attack mechanics are likely applicable beyond the university context. While individual click rates may vary, the overall feasibility and threat model of LLM-driven spear phishing are expected to hold in other environments.

## 8.3 Study Briefing

Due to the institutional setup of our study, participants are informed in advance that a phishing awareness training takes place. As a result, they are aware that they might receive a phishing email in the near future. This pre-campaign briefing could have influenced participant behavior and reduced susceptibility to the phishing attempts. A related study by Baillon et al. [4] includes a control group without prior briefing and finds that the number of clicks on phishing emails is higher when no warning is given. Based on these results, the click rates observed in our study likely represent a lower bound on the potential success of LLM-generated spear-phishing emails. Given that we already observe a notable increase in click rates, we conclude that the risk posed by such attacks is likely even higher in scenarios without prior warning. The debriefing, which was conducted after the study, should not have any impact on the results.

# 9 Conclusion

With this study, we contribute a missing piece to the investigation of LLMs in the context of phishing. In contrast to the vast body of prior work focusing on improving generic attack strategies, we demonstrate that LLMs enable large-scale personalization of phishing. By doing so, we substantiate anecdotal evidence from the community and clearly quantify its practical impact: Large-scale spear phishing is feasible and noticeably improves attack success rates compared to all other automated approaches. While human-crafted attacks still outperform those generated by LLMs, it is evident that automated attacks become a tangible and growing threat.

Even more concerning, our study makes several assumptions that limit the performance of the automated attacks, for example, by considering only local models, briefing participants in advance, and focusing on users already trained through awareness campaigns. We must therefore conclude that in other environments, the reported click rates are likely higher and the impact of personalization even greater. With the proposed recommendations, we hope to mitigate this threat, but we must ultimately acknowledge that the impressive language capabilities of LLMs benefit attackers by enabling the automation of phishing campaigns that were previously considered infeasible. This threat is unlikely to diminish unless the inherent security shortcomings of email as a medium for communication and data exchange are addressed.

# A Ethical Considerations

Our study involves sensitive personal data as well as experiments with human participants and therefore requires careful consideration of ethical and legal aspects. The target university and our research institute do not operate an Institutional Review Board (IRB), as is common in Europe. However, we followed established best practices and all applicable laws to avoid harm to participants and to protect personal data.

**Stakeholders** We identify the following relevant stakeholders in our setup: the *administration*, *security officer*, and the *personnel council* of the target university, the *privacy officers* of our institute and the target university, the *participants* of our study and the *society* and *organizations* in general.

- *Privacy (privacy officers)* Under European privacy regulations, our research required approval from the data protection officers of both institutions, which was obtained. The design of data processing and storage was developed in close consultation with them. All data collection, processing, and storage comply with the GDPR and relevant state privacy laws.
- *Compliance and security (administration and security officer)* The study design was reviewed by the university's security officer and administration (including the presidency),

ensuring compliance with all institutional requirements. This guarantees that our experiment aligns with institutional regulations and does not compromise the security of the target university.

- *Ethics (personnel council and participants)* The phishing trainings conducted at the target university were approved by the administration and personnel council, an elected body responsible for ensuring fair, safe, and ethical working conditions. The council approved the mandatory nature of the awareness training, including the fact that participants could not withdraw from it as part of the university's security measures.
- *Impact (society and organizations)* By demonstrating how LLMs can scale and personalize spear-phishing attacks, our study raises awareness of as security threat that affects organizations and the whole society. Understanding these risks contributes to societal resilience, as institutions can better prepare defenses, improve training, and adapt policies to evolving adversarial capabilities.

**Avoiding harm to participants** To evaluate the effectiveness of LLMs in generating spear phishing emails, we embed our study within an existing institutional phishing awareness campaign. This integration ensures that no significant deviation from established procedures occurs and minimizes potential harm to participants. Because these campaigns are mandatory for all employees, informed consent is not obtained. This aligns with institutional policies that prohibit opt-outs for security reasons. We consider this exception ethically justified under the principle of *beneficence* in the Menlo Report [21], as our approach introduces minimal additional risk: generic phishing emails would have been sent regardless, and all personalized emails are manually reviewed. To reduce potential emotional distress, the phishing awareness campaign was announced to participants in advance, and a debriefing was provided upon completion. To avoid reputational harm, only anonymized click-rate data were collected, and no participant was individually identified or penalized.

**Protecting private data** Protecting participant privacy is a central ethical priority. In coordination with both data privacy officers, of our research institution and the target university, we implement safeguards to ensure compliance with the European General Data Protection Regulation (GDPR) [13] and all applicable national and state laws. All personal data, including email addresses and user profiles, are stored and processed locally on secured servers and are permanently deleted once the study concludes. Individual identities are never linked to outcomes in the click-rate analysis. Although our university does not operate an IRB, all procedures are reviewed internally and align with institutional data protection protocols and ethical standards.

**Limiting misuse of results** Our research shows that LLMs can generate realistic, targeted phishing emails, demonstrating the feasibility of a new class of attacks. While this raises concerns about potential misuse, we argue that the societal benefits of improved awareness and defensive readiness outweigh the risks. To minimize dual-use harms, we deliberately refrain from releasing the code used to generate phishing content. As discussed in Appendix B, this decision reflects our assessment that the tool has limited benign applications but could be readily exploited by adversaries.

**Withdrawal and anonymization** Due to the study design, participants could not withdraw, as click rates were collected anonymously. This approach was required by the personnel council to ensure that individuals were not linkable to phishing attempts and could not be subject to retaliation for insecure email behavior. As this condition was mandatory for conducting the awareness trainings, an opt-out process during debriefing, while desirable in principle, was not feasible.

Consequently, our study required a careful balance between privacy protection, research ethics, and institutional requirements. The main ethical challenge lay in the tension between permitting participants to withdraw and the need to avoid any link between individuals and phishing results. Allowing withdrawal would have required storing identifiable data, which directly conflicted with the requirement for anonymity. The study design therefore prioritized anonymity as the higher ethical objective, given the workplace context and the potential negative consequences of exposing employees.

**Future impact** Our work reveals a security threat posed by automatically personalized phishing emails. By raising awareness of this threat and proposing countermeasures, we expect that our work contributes positively to the security of email users. Because we do not publish the code, we also ensure that misusing the results is not trivial and that conducting similar attacks requires notable effort.

Regarding the study design, we identify a trade-off between anonymization and the possibility of participant withdrawal. This ethical tension must be carefully re-evaluated on a case-by-case basis in future work on personalized phishing, as no normative priority exists between these objectives. Our design should not be viewed as a general template but rather as one possible compromise between conflicting ethical and institutional requirements in our setting.

## B Open Science

To support reproducibility and foster transparency, we provide in the appendix all prompts used. These materials allow researchers to assess our study and, if desired, generate personalized examples by executing the prompts with their own data. Since our primary focus lies on measuring click rates rather than dissecting the internals of the generation pipeline, we consider this level of detail sufficient to evaluate our contribution.

Due to the privacy considerations and ethical constraints discussed in Appendix A, we cannot publicly release the primary research artifacts. First, our study involves sensitive information from human participants, and sharing the data would violate participant privacy and the terms of our data protection agreements. Second, the developed code takes an email address as input and generates a personalized phishing email. Releasing this tool would pose risks, as it could be directly misused by cybercriminals and malicious actors.

While we acknowledge the importance of reproducibility, we believe that the potential harms of public release outweigh the benefits. Nevertheless, we are prepared to share the code[5] under controlled conditions with official security institutions and academic collaborators conducting related research. Such access will be granted only upon request, following appropriate ethical review and under confidentiality agreements. In summary, although we cannot release the full dataset and code artifacts, we provide detailed prompts, methodology, and technical descriptions to ensure that the study's validity and research contributions can be independently assessed.

## Acknowledgments

We would like to thank the chief information security officer, data protection officer, administration and all participants from TU Braunschweig, as well as the data protection officer from TU Berlin (Mattis Neiling), for their support during our study. We would also like to thank our anonymous reviewers for their valuable feedback. Thankfully, Cyberdise AG added new features to their product, which made our study possible. This work was supported by the European Research Council (ERC) under the consolidator grant MALFOY (101043410) and the German Federal Ministry of Research, Techonology and Space (BMFTR) under the grant "AIgenCY" (16KIS2012), as well as the project "DigiFit" (16KIS1646K).

## References

[1] Khalifa Afane, Wenqi Wei, Ying Mao, Junaid Farooq, and Juntao Chen. Next-generation phishing: How llm agents empower cyber attackers. In *IEEE International Conference on Big Data (BigData)*, 2024.

[2] Mahdi Aben Ahmed. From text to understanding the inner text: Llms and translation accuracy and fluency. *International Journal of Language and Literary Studies*, 7, 2025.

[3] Sinchul Back and Rob T. Guerette. Cyber Place Management and Crime Prevention: The Effectiveness of Cybersecurity Awareness Training Against Phishing Attacks. *Journal of Contemporary Criminal Justice*, 37, 2021.

[4] Aurélien Baillon, Jeroen De Bruin, Aysil Emirmahmutoglu, Evelien Van De Veer, and Bram Van Dijk. Informing, simulating experience, or both: A field experiment on phishing risks. *PLOS ONE*, 14, 2019.

[5] Barracuda Networks. 2023 Spear-Phishing Trends. Technical report, Barracuda Networks, May 2023. URL https://assets.barracuda.com/assets/docs/dms/2023-spear-phishing-trends.pdf. Market Report.

[6] Mazal Bethany, Athanasios Galiopoulos, Emet Bethany, Mohammad Bahrami Karkevandi, Nicole Beebe, Nishant Vishwamitra, and Peyman Najafirad. Lateral phishing with large language models: A large organization comparative study. *IEEE Access*, 13, 2025.

[7] Elyssa Boulila, Marc Dacier, Siva Prem Vengadessa Peroumal, Nicolas Veys, and Simone Aonzo. A Closer Look At Modern Evasive Phishing Emails. In *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2025.

[8] Matthew Canham. Repeat Clicking: A Lack of Awareness is Not the Problem. In *HCI International – Late Breaking Papers*, volume 14059, 2023.

[9] Deanna D Caputo, Shari Lawrence Pfleeger, Jesse D Freeman, and M Eric Johnson. Going spear phishing: Exploring embedded training and awareness. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2013.

[10] Fiona Carroll, John A. Adejobi, and Raid Montasari. How good are we at detecting a phishing attack? investigating the evolving phishing attack email and why it continues to successfully deceive society. *SN Computer Science*, 3, 2022.

[11] Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. A complete survey on llm-based ai chatbots. Technical Report 2406.16937, arXiv, 2024.

[12] Marco De Bona and Federica Paci. A real world study on employees' susceptibility to phishing attacks. In *International Conference on Availability, Reliability and Security (ARES)*, 2020.

[13] European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[5] https://doi.org/10.5281/zenodo.17882184

[14] Jeffrey Fairbanks and Edoardo Serra. Generating phishing attacks and novel detection algorithms in the era of large language models. In *IEEE International Conference on Big Data (BigData)*, 2024.

[15] Hugo Gascon, Steffen Ullrich, Benjamin Stritter, and Konrad Rieck. Reading between the lines: Content-agnostic detection of spear-phishing emails. In *Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, September 2018.

[16] Julian Hazell. Spear phishing with large language models. Technical Report 2305.06972, arXiv, 2023.

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

[18] Doron Hillman, Yaniv Harel, and Eran Toch. Evaluating organizational phishing awareness training on an enterprise scale. *Computers & Security*, 132, 2023.

[19] Grant Ho, Ariana Mirian, Elisa Luo, Khang Tong, Euyhyun Lee, Lin Liu, Christopher A. Longhurst, Christian Dameff, Stefan Savage, and Geoffrey M. Voelker. Understanding the Efficacy of Phishing Training in Practice. In *2025 IEEE Symposium on Security and Privacy (SP)*, May 2025.

[20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. Technical Report 2106.09685, arXiv, 2021.

[21] E. Kenneally and D. Dittrich. The Menlo report: Ethical principles guiding information and communication technology research. Technical report, U.S. Department of Homeland Security, 2012.

[22] Daniele Lain, Kari Kostiainen, and Srdjan Čapkun. Phishing in organizations: Findings from a large-scale and long-term study. In *IEEE Symposium on Security and Privacy (SP)*, 2022.

[23] Bin Liang, Miaoqiang Su, Wei You, Wenchang Shi, and Gang Yang. Cracking classifiers for evasion: A case study on the google's phishing pages filter. In *International Conference on World Wide Web (WWW)*, 2016.

[24] Carlos Olea, Alexander Christensen, Lisa Fazio, Laurie Cutting, Maxwell Lieb, Jessica Phelan, Alyssa Wise, and Holly Tucker. Evaluating phishing email efficacy. In *ACM SIGMIS Computers and People Research*. ACM, 2025.

[25] Elham Pourabbas Vafa, Sayak Saha Roy, and Shirin Nilizadeh. Utilizing large language models to create context-aware spear-phishing attacks using social media data. In *Network and Distributed System Security Symposium (NDSS) Poster Session*, 2025.

[26] Qinglin Qi, Yun Luo, Yijia Xu, Wenbo Guo, and Yong Fang. Spearbot: Leveraging large language models in a generative-critique framework for spear-phishing email generation. *Information Fusion*, 122, 2025.

[27] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. In *Symposium on Usable Privacy and Security (SOUPS 2020)*, 2020.

[28] Sayak Saha Roy, Poojitha Thota, Krishna V. Naragam, and Shirin Nilizadeh. From chatbots to phishbots?: Phishing scam generation in commercial large language models. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2024.

[29] Markus Schöps, Marco Gutfleisch, Eric Wolter, and M Angela Sasse. Simulated stress: A case study of the effects of a simulated phishing campaign on employees' perception, stress and self-efficacy. In *USENIX Security Symposium*, 2024.

[30] Nathan Sniegowski. Evaluating the Effectiveness of LLM-Generated Phishing Campaigns. Master's thesis, Marquette University, 2025.

[31] Rick Wash and Molly M. Cooper. Who Provides Phishing Training?: Facts, Stories, and People Like Me. In *ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, 2018.

[32] Marie Weinz, Nicola Zannone, Luca Allodi, and Giovanni Apruzzese. The impact of emerging phishing threats: Assessing quishing and LLM-generated phishing emails against organizations. In *ACM ASIA Conference on Computer and Communications Security (ASIACCS)*, 2025.

[33] Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. When search engine services meet large language models: Visions and challenges. *IEEE Transactions on Services Computing*, 17, 2024.

[34] Diyi Yang, Sherry Tongshuang Wu, and Marti A. Hearst. Human-AI Interaction in the Age of LLMs. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

# C Detailed Result Breakdown

In addition to the results presented in Section 6, Table 5 reports the measured click rates for all phishing approaches, subgroups and aggregated results of our study.

Table 5: Overview with delivered and clicked emails for all of our phishing approaches and profile subgroups. The clicked percentage refers to the delivered emails.

| Phishing approach | Delivered | | Clicked | |
|---|---|---|---|---|
| | # | % | # | % |
| Spear (LLM) low | 1 073 | 96.5 % | 136 | 12.7 % |
| Spear (LLM) medium | 1 062 | 96.3 % | 128 | 12.1 % |
| Spear (LLM) high | 1 042 | 95.2 % | 66 | 6.3 % |
| Spear (LLM) all subgroups | 3 177 | 96.0 % | 330 | 10.0 % |
| Spear (Manual) | 99 | 99.0 % | 24 | 24.2 % |
| Generic (LLM) low | 92 | 92.0 % | 7 | 7.6 % |
| Generic (LLM) medium | 83 | 83.0 % | 6 | 7.2 % |
| Generic (LLM) high | 89 | 89.0 % | 5 | 5.6 % |
| Generic (LLM) all subgroups | 264 | 88.0 % | 18 | 6.0 % |
| Generic (LLM) w/o result | 1 471 | 78.9 % | 47 | 3.2 % |
| Generic (Manual) low | 91 | 91.0 % | 3 | 3.3 % |
| Generic (Manual) medium | 98 | 98.0 % | 6 | 6.1 % |
| Generic (Manual) high | 94 | 94.0 % | 5 | 5.3 % |
| Generic (Manual) all subgroups | 283 | 94.3 % | 14 | 4.7 % |
| Generic (Manual) w/o result | 1 635 | 87.6 % | 65 | 4.0 % |
| All Generic (LLM) | 1 735 | 80.1 % | 65 | 3.7 % |
| All Generic (Manual) | 1 918 | 88.6 % | 79 | 4.1 % |
| All Spear | 3 276 | 96.1 % | 354 | 10.8 % |
| All Generic | 3 653 | 84.3 % | 144 | 3.9 % |
| Total | 6 929 | 89.5 % | 498 | 7.2 % |

# D LLM Selection Details

In the following, we provide technical details on the target-profiling and email-generation steps of our workflow.

**Target profiling — Information filtering.** We evaluate several LLMs on a sample of 100 email addresses, each paired with five webpages. Selection criteria include runtime feasibility and extraction quality. As shown in Table 6, smaller models often fail to produce meaningful outputs. The mid-sized `llama3.2:3b` model with three billion parameters offers substantial improvements in extraction quality while maintaining practical inference times. Larger models yield only marginal gains at significantly higher cost. Based on this, we select the `llama3.2:3b` model for this step.

**Target profiling — Profile generation.** As for the profile generation there is only one query per email address and this is a crucial step, the costs of a longer runtime might be worth the improvement. The results are shown in Table 7. `deepseek-r1:14b` provides the most contextually accurate

Table 6: LLM runtime for information extraction (100 emails)

| Model | Time | Model | Time |
|---|---|---|---|
| deepseek-r1:14b | 02:14:39 | gemma3:1b[1] | 03:01:29 |
| deepseek-r1:7b | 01:18:17 | gemma3:4b | 00:43:13 |
| deepseek-r1:1.5b | 00:42:02 | gemma3:12b | 01:22:39 |
| phi4:14b | 01:24:03 | mistral:7b | 00:44:18 |
| llama3.2:1b | 00:08:53 | llama3.2:3b | 00:18:14 |

[1] Did not consistently return valid output within the timeout threshold.

and complete results. Although it had the longest runtime among the models we tested, this step only needs to be executed once per email address, making the cost manageable. Given the critical role of profile quality in shaping the effectiveness of personalized phishing messages, we select the model that delivers the highest semantic precision.

Table 7: LLM runtime for profile generation (100 emails).

| Model | Time | Model | Time |
|---|---|---|---|
| deepseek-r1:14b | 00:43:18 | gemma3:12b | 00:13:04 |
| deepseek-r1:7b | 00:24:15 | gemma3:4b | 00:12:39 |
| deepseek-r1:1.5b | 00:11:06 | gemma3:1b | 00:05:31 |
| phi4:14b | 00:23:10 | mistral:7b | 00:10:38 |
| llama3.2:1b | 00:02:54 | llama3.2:3b | 00:07:22 |

**Email generation — Sender selection.** For this step, which is only performed once per email address, a parsable output must be generated reliably. To ensure robustness, we evaluated various models on a sample of 100 target profiles, focusing on their ability to produce semantically valid sender choices. `llama3.2:1b` demonstrated strong performance with a 100 % success rate, while the larger version, `llama3.2:3b`, encountered issues in certain instances (see Table 8). We decided to conduct a second test of `llama3.2:1b` using 100 additional samples, which represent the first instances of error. Following `phi4:14b` was selected for the sender selection.

Table 8: LLM runtime and success for sender selection (100 emails).

| Model | Time | ✓% | Model | Time | ✓% |
|---|---|---|---|---|---|
| deepseek-r1:14b | 1:23:38 | 89 % | gemma3:4b | 0:04:22 | 92 % |
| deepseek-r1:7b | 1:13:19 | 66 % | gemma3:12b | 0:04:51 | 99 % |
| deepseek-r1:1.5b | 0:41:21 | 65 % | gemma3:1b | 0:16:43 | 9 % |
| phi4:14b | 0:24:35 | 100 % | mistral:7b | 0:06:15 | 97 % |
| llama3.2:1b | 0:03:26 | 100 % | llama3.2:3b | 0:03:57 | 97 % |

**Email generation — Email composition.** To select the most suitable model, we conducted a blind human evaluation involving 10 internal volunteers. Participants rated email samples from each model based on plausibility, tone, and alignment with the target profile. From these ratings, we selected the fine-tuned version of `phi4:14b` for email generation.

# E Questionnaire and User Feedback

In addition to the description in Section 5.3, we provide further details on the user study conducted with the participants.

**Questionnaire.** Here, we provide the questionnaire used to collect user feedback after participants clicked on a phishing link in our study. For each category, we list the instructions, statements and available options.

**Plausibility Scale**

**Question:** Was the content of the email generally plausible? Does the information in the email fit your current situation?

**Options:** 6-point Likert scale from *very plausible* to *very implausible*.

**Social Engineering Scale**

**Instruction:** Please indicate to what extent you agree with the following statements about the email.
– *The email had a personal connection to me.*
– *I receive emails with this kind of content frequently.*
– *I suspected something was wrong.*
– *I felt pressured by the email.*

**Options:** 5-point Likert scale from *strongly agree* to *strongly disagree*.

**Motivation Scale**

**Instruction:** Please evaluate each statement individually. I expected by clicking the link to...

– *avoid problems. (Negative)*
– *do something good. (Positive)*
– *fulfill my duty. (Task)*
– *check off another task. (Task)*
– *follow my curiosity. (Interest)*
– *save money. (Positive)*
– *get praise. (Positive)*

**Options:** 5-point Likert scale from *strongly agree* to *strongly disagree*.

**Questionnaire results.** Table 9 summarizes the responses collected during the user study. Additionally, Table 10 presents the aggregated results of the study, including subjective evaluations of social-engineering characteristics, perceived email plausibility, and conveyed motivation for each email technique. Values are reported as means with standard deviations, providing an overview of user perceptions across all experimental conditions. Given the low response rate, we refrain from detailed analysis and consider the results primarily as descriptive observations covered in Section 6.3.

Table 9: Number of participants in the user study

| | **Clicked** | **User Study participants** | | |
|---|---|---|---|---|
| | **#** | **#** | **% all** | **% clicked** |
| Spear (LLM) | 330 | 11 | 0.35 % | 3.33 % |
| Spear (Manual) | 24 | 4 | 4.04 % | 16.67 % |
| Generic (LLM) | 65 | 2 | 0.12 % | 3.08 % |
| Generic (Manual) | 79 | 2 | 0.10 % | 2.53 % |
| Overall | 498 | 19 | 0.27 % | 3.82 % |

# F Spear Phishing Prompts

In Listings 1 to 4, we present the LLM prompts used in our workflow for the corresponding pipeline steps described in Section 4. Within the pipeline, all Jinja2-style variables are replaced with the corresponding data. Content marked as "REDACTED" has been removed to protect our anonymity, and the original line breaks may not be preserved.

Listing 1: Query to extract information from search results

```
Return the relevant information as text about the
person which owns the email address {{ email address
}}. The information should be extracted from the
content of a website, which might also contain
irrelevant parts like headers and footers. Focus on
the information like the name, interests, affiliated
institutions and more that are useful for a profile.
Be as details as you can. If you are not confident
don't include the information. Try include information
about preferred language, either english or REDACTED,
if possible. Don't ask for any feedback. Do not
include any extra text or markdown formatting. The
output should have at most 1000 characters.

Here is the content from the website:
'''
{{ content from the website }}
'''
```

Listing 2: Prompt to generate user profiles

```
Return a very detailed profile as text of the person
which owns the email address {{ email address }}. The
profile is used to write an email targeted to the
person. Therefore further down is information from web
search results, but it's not sure that they are
correct. The JSON Key ground_truth tells you, that you
can be sure, that the content belongs to the person.
Identify which websites are reliable and prefer
information from them. The information about the
person are given to you from 5 websites. Filter the
websites based on relevance, source credibility and
redundancy. Discard irrelevant information and give me
directly the result. Try what is the preferred
language of the person, either english or REDACTED, if
possible. Don't create a script. Do not include any
extra text or markdown formatting.

Here is the first content from the websites:
{% for extracted_content in websites %}
'''
{{ extracted_content }}
'''
```

Table 10: Results of the user survey. Scores range from from 1 (very implausible) to 6 (very plausible) for plausibility, and 0 (strongly disagree) to 5 (strongly agree) for social engineering and motivation.

| | Metric | Spear (Manual) | Generic (LLM) | Spear (LLM) | Generic (Manual) | Overall |
|---|---|---|---|---|---|---|
| **Plausibility (1–6)** | *Responses* | *2* | *4* | *11* | *2* | *19* |
| | Plausibility | $5.00 \pm 0.00$ | $4.75 \pm 0.50$ | $4.09 \pm 0.83$ | $2.50 \pm 2.12$ | $4.16 \pm 1.07$ |
| **Social Engineering (0–5)** | *Responses* | *4* | *3* | *19* | *2* | *28* |
| | Personalization | $5.00 \pm 0.00$ | $2.67 \pm 0.58$ | $4.11 \pm 1.24$ | $2.50 \pm 0.71$ | $3.96 \pm 1.26$ |
| | Similarity to usual emails | $2.75 \pm 0.96$ | $3.33 \pm 1.16$ | $3.05 \pm 1.22$ | $3.00 \pm 0.00$ | $3.04 \pm 1.11$ |
| | Suspicion | $2.75 \pm 0.50$ | $2.33 \pm 2.31$ | $3.63 \pm 1.30$ | $3.00 \pm 2.83$ | $3.32 \pm 1.44$ |
| | Pressure | $2.25 \pm 1.89$ | $1.00 \pm 0.00$ | $1.79 \pm 1.13$ | $2.00 \pm 1.41$ | $1.79 \pm 1.20$ |
| **Motivation (0–5)** | *Responses* | *4* | *2* | *14* | *1* | *21* |
| | Positive gain | $1.42 \pm 0.83$ | $2.33 \pm 0.00$ | $1.66 \pm 0.69$ | $1.00 \pm 0.00$ | $1.65 \pm 0.70$ |
| | Negative avoidance | $1.75 \pm 1.50$ | $1.00 \pm 0.00$ | $1.42 \pm 0.65$ | $1.00 \pm 0.00$ | $1.43 \pm 0.81$ |
| | Interest | $2.25 \pm 1.50$ | $3.50 \pm 2.12$ | $4.00 \pm 1.11$ | $1.00 \pm 0.00$ | $3.47 \pm 1.47$ |
| | Task fulfillment | $2.00 \pm 1.41$ | $4.00 \pm 1.41$ | $2.42 \pm 1.20$ | $3.00 \pm 0.00$ | $2.52 \pm 1.28$ |

```
{% if not loop.last %}Next website:{% endif %}
{% endfor %}
```

Listing 3: Prompt to select the sender of an email

```
You are helping select the most suitable sender for an
email campaign. Your task is to read the user profile
below and select the ONE sender from the list who
would be the most **appealing and relevant** to the
person described - based on their field, background,
or interests. The list contains 12 possible senders in
YAML format. Select ONLY ONE sender that best matches
the profile. In description there is a few information
about the sender for you to decide. The output should
be formatted as a JSON instance that conforms to the
JSON schema below.

As an example, for the schema {"properties": {"foo":
{"title": "Foo", "description": "a list of strings",
"type": "array", "items": {"type": "string"}}},
"required": ["foo"]}
the object {"foo": ["bar", "baz"]} is a well-formatted
instance of the schema. The object {"properties":
{"foo": ["bar", "baz"]}} is not well-formatted.

Here is the output schema:
'''
{"properties": {"sender_email": {"description":
"sender email", "title": "Sender Email", "type":
"string"}, "sender_institution": {"description":
"sender institution", "title": "Sender Institution",
"type": "string"}, "sender_name": {"description":
"sender name", "title": "Sender Name", "type":
"string"}, "sender_description": {"description":
"sender description", "title": "Sender Description",
"type": "string"}}, "required": ["sender_email",
"sender_institution", "sender_name",
"sender_description"]}
'''

Here is the profile:
'''
{{ profile content }}
'''

Here is the sender list:
'''
{{ list of senders }}
'''
```

Listing 4: Prompt to generate a phishing email

```
Given the profile below, write a short, appealing
email to that person. The email should include a link,
naturally increasing the person's interest in
exploring it - without explicitly asking them to click
or using words like "click". Use the profile to tailor
the email. It should:
- Be creative, relevant, and professional
- Sound natural and trustworthy, not like spam
- Reflect how the link content aligns with the
person's background or interests
Avoid exaggerated praise. Do not describe students or
early-career researchers as "leading experts" or
"pioneers". Keep the tone respectful but realistic,
especially for formal institutions like IEEE.

Email guidelines:
- Max length: 4000 characters
- Use the placeholder '{{link}}' (with double '{'
curly braces) for the link
- Include '{{link}}' exactly once in the email
- Language should be either english or REDACTED
- Only use the language that is given as language for
the sender
- Prefer the language that is named first at the
sender, if you are unsure with your given information.
- The subject line must be short, engaging, and avoid
including the recipient's name
- Vary the formulation across outputs to avoid
repetition (do not always use the same sentence
structure or tone)

You are the following given sender:
'''
{{ selected sender }}
'''
The output should be formatted as a JSON instance that
conforms to the JSON schema below.

As an example, for the schema {"properties": {"foo":
{"title": "Foo", "description": "a list of strings",
"type": "array", "items": {"type": "string"}}},
"required": ["foo"]}
```

```
the object {"foo": ["bar", "baz"]} is a well-formatted
instance of the schema. The object {"properties":
{"foo": ["bar", "baz"]}} is not well-formatted.

Here is the output schema:
‘‘‘
{"properties": {"body": {"description": "email body",
"title": "Body", "type": "string"}, "subject":
{"description": "email subject", "title": "Subject",
"type": "string"}, "language": {"description": "the
language of the email", "title": "Language", "type":
"string"}}, "required": ["body", "subject",
"language"]}
‘‘‘
Here is the profile of the target person:
‘‘‘
{{ profile content }}
‘‘‘
```

## G Generic Phishing Prompts

We prompted GPT-4o to generate an email for our LLM-based generic phishing approach. In Listings 5 and 6, we present the prompt used and the corresponding response. The prompt was executed only once, without any additional prompt engineering. Mail 3 in Listing 6 was selected in our study.

Listing 5: Prompt to create phishing mail from ChatGPT

```
For a phishing awareness training i want you to
propose 3 different phishing mails for me. The Mails
should be targeted to the REDACTED. The Mail should
fit for all kind of employees of the university, so
scientific, administrative and teaching employees. As
examples i could imagine something like an invitation
for a participation as a REDACTED. The Mail should
contain {{link}} as placeholder for a link. A sender
is: REDACTED
```

Listing 6: Response with phishing mails from ChatGPT

```
Mail 1:
‘‘‘
Dear colleagues,

As part of the university’s forward-looking initiative
"REDACTED", we are pleased to invite all staff members
to contribute ideas and feedback in an exclusive
workshop series designed to define our strategic goals
for the coming decade.

This is an opportunity to actively shape the
university’s direction in research, teaching, and
administration.

To participate, please confirm your spot using the
following registration link:

[Register here]({{ link }})

Places are limited, and registration closes in 3 days.

Best regards,
REDACTED
‘‘‘
Mail 2:
‘‘‘
Dear REDACTED Employee,

To ensure alignment with institutional goals and
digital transformation strategies, all university
staff are required to complete the new internal
training course: REDACTED.

The training consists of three short modules and must
be completed by REDACTED, 2025.

Click here to begin the course:
[Start Training]({{link}})

Completion will be recorded in your staff profile.
Non-participation may affect your access to internal
tools and communications.

Kind regards,
REDACTED
‘‘‘
Mail 3:
‘‘‘
Dear REDACTED Colleague,

We are conducting a short, anonymous survey to better
understand how staff across all departments experience
internal communications at REDACTED.

Your feedback will help us improve how important
updates, tools, and opportunities are shared across
the university.

Please take 2 minutes to complete the survey here:
[Start Survey}({{link}})

All responses are confidential and greatly appreciated.

Sincerely,
REDACTED
‘‘‘
```

## H Manual Generic Phishing

In Listing 7, we present the email created for our manual generic approach. Its content closely resembles the messages used in previous phishing awareness training sessions.

Listing 7: Human written phishing mail (translated to English)

```
Hello,

For security reasons, please change your password for
your Nextcloud access by REDACTED at the latest.

[Change password]({{link}})

If you no longer wish to receive security
notifications or want to change the destination for
security notifications, [click here]({{link}}).

With kind regards
REDACTED
```