Dharmi Patel
Pandit Deendayal Energy University
dharmip060@gmail.com
+91 9624951017
31st March 2025

Professor Daniela Rus
Director, Computer Science and Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology (MIT)
32 Vassar St, Cambridge, MA 02139, USA

Subject: Application for Research on AI-Generated Phishing Emails and Security Measures

Dear Professor Rus,

I am writing to express my strong interest in conducting research on how AI-generated phishing emails ~~by phishing~~ bypass traditional security measures and manipulate employees into revealing sensitive data. Given the increasing sophistication of cyber threats, particularly those powered by artificial intelligence, my study aims to analyze the effectiveness of AI-driven phishing attacks through penetration testing and assess corporate employees' response rates and vulnerabilities.

With a background in Computer Science, my academic and professional experience has prepared me to investigate cybersecurity threats critically. I have developed expertise in penetration testing, which will be ~~insutted~~ instrumental in conducting this research. My methodology involves simulating real-world AI-generated

phishing attacks, collecting response data, and evaluating security weaknesses to propose more effective countermeasures.

I am particularly interested in working under your guidance at MIT, as your expertise in cybersecurity and AI security aligns closely with my research goals. I am confident that this study will contribute valuable insights to enhancing cybersecurity awareness and defense mechanisms against AI-driven phishing attacks.

I appreciate your time and consideration and would welcome the opportunity to discuss how my research with the ongoing work at MIT. Please find my CV attached for your reference. I look forward to your positive response.

Sincerely,
Dharmi Patel

# RESEARCH PROPOSAL

## Analyzing AI-Generated Phishing Threats in Corporate Environments


**Prepared By**
Dharmi Patel

**Prepared For**
Professor Daniela Rus
Director, CSAIL
MITT

# Table of Contents

## ABSTRACT

The integration of artificial intelligence (AI) into cybercrime tactics has significantly increased the complexity and effectiveness of phishing attacks, posing a serious challenge to traditional cybersecurity measures. AI generated phishing emails leverage advanced natural language processing (NLP) and machine learning models to craft highly persuasive messages that can evade rule-based spam filters and deceive even trained employees. Unlike conventional phishing attempts, AI-powered attacks can dynamically adapt content based on context, making them more difficult to detect and more likely to succeed.

This study aims to investigate how AI-generated phishing emails bypass traditional security defenses and exploit human vulnerabilities to extract sensitive information. The research will employ penetration testing, where AI driven phishing emails will be deployed in controlled corporate environments to assess employee response rates and identify key behavioral and technical vulnerabilities. By analyzing email content, engagement patterns, and the psycological triggers exploited in these attacks, we will determine the most effective tactics used by AI-generated phishing campaigns. Additionally, we will evaluate the performance of existing security measures, such as spam filters, AI based email security solutions, and employee awareness training to understand their limitations against AI driven threats.

The findings from this study will provide critical insights into the evolving nature of phishing attacks and their implications for corporate security. Based on our analysis, we will propose improved countermeasures, including AI-enhanced phishing detection systems, adaptive cybersecurity training tailored to AI-generated threats and the integration of behavioral analytics into security frameworks. This research aims to bridge the gap between emerging AI-driven cyber threats and existing defense mechanisms, contributing to the development of more resilient cybersecurity strategies in the face of increasingly sophisticated phishing attacks.

Keywords

AI generated phishing, Cybersecurity threats, Social engineering attacks, Penetration testing, Phishing detection, Email security, AI in cybersecurity, Cyber threat intelligence, Security awareness training, Behavioral cybersecurity.

Date: YOUVA

## Introduction

The rapid advancement of Artificial Intelligence (AI) has revolutionized various industries, including cybersecurity. While AI has been leveraged to enhance security measures, it has also introduced new and more sophisticated cyber threats. One of the most alarming developments in the use of AI in phishing attacks, where machine learning models and Natural Language Processing (NLP) generate deceptive emails that closely mimic human communication. These AI-generated phishing emails can evade traditional security defenses such as rule-based spam filters and employee awareness training, posing a significant risk to organizations.

Phishing remains one of the most effective cyberattack techniques, responsible for a large proportion of data breaches and financial fraud worldwide. Unlike conventio-nal phishing emails, AI-powered phishing attacks can dynamically adapt their content, improve grammatical correctness and generate personalized messages at scale, making them harder to detect and more convincing to targets. This increasing level of sophistication has exposed vulnerabilities in traditional email security systems and human awareness programs, requiring new defense strategies to mitigate the risks.

This research aims to investigate how AI-generated phishing emails bypass traditional security measures and manipulate corporate employees into revealing sensitive information. The study will conduct penetration testing using AI-generated phishing emails to assess employee

response rates and identify key behavioral and technical vulnerabilities. By analyzing the tactics used in AI-powered phishing campaigns and evaluating the effectiveness of existing security defenses, the study seeks to provide insights into the weakness of current anti-phishing mechanisms and propose advanced mitigation strategies.

## Literature Review

Phishing has long been a significant cybersecurity threat, with attackers continually refining their techniques to bypass security defenses. The emergence of artificial intelligence (AI) and Natural Language Processing (NLP) has transformed phishing attacks, enabling the creation of highly convincing, adaptive, and context aware phishing emails. Unlike traditional phishing emails, which often contain grammatical errors or suspicious content, AI-generated phishing messages leverage machine learning (ML) models to mimic legitimate communication patterns, making them significantly harder to detect.

Despite advancements in cybersecurity, traditional email security measures such as spam filters, blacklists, and rule-based detection struggle to counter AI-generated phishing attacks. These systems rely on detecting known patterns, suspicious keywords, or previously flagged URLs, but AI-generated phishing emails can modify their content dynamically to evade detection.

To combat AI-powered phishing, researchers have developed AI-driven detection systems and adaptive security training to counter AI-powered phishing. Deep learning models analyze linguistic patterns and sender behaviour to identify phishing emails. AI-powered training programs using personalized simulations and real-time feedback outperform traditional methods. However, these solutions remain vulnerable to attacks, emphasizing the need for continuous improvements in detection and security.

Despite advancements in mitigation AI generated phishing threats, gaps remain in real-world application and long term adaption.

## Research Gap

With the rise of AI-generated phishing attacks, traditional cybersecurity defenses are becoming increasingly ineffective. AI-powered phishing emails leverage natural language processing (NLP) and machine learning (ML) to create highly convincing and adaptive phishing messages, making them harder to detect. While researchers have explored AI-generated phishing threats, particularly in real-world corporate environments. Key gaps in existing research include:

- Ineffectiveness of Traditional Security Measures: AI-generated phishing emails can bypass rule-based spam filters, blacklists and even AI powered detection systems.
- Lack of Real World Empirical Studies: Most studies on AI-generated phishing focus on controlled lab settings rather than actual corporate environments.
- Limited Research on Employee Behavioral Adaption: There is insufficient data on how employees react to AI-generated phishing attacks over time, raising concerns about the long term effectiveness of security awareness training.

Addressing these research gaps is critical for strengthening corporate cybersecurity against AI-powered phishing attacks. This study aims to bridge these gaps by conducting real-world penetration testing, analyzing employee response by evaluating AI-driven security defenses. The insights gained will help in developing more adaptive and effective cybersecurity strategies against AI-generated phishing threats.

## Research Questions

As AI-generated phishing attacks become more sophisticated, it is essential to understand how these threats bypass existing security measures and manipulate human behavior. This study aims to explore key aspects of AI-powered phishing attacks, security vulnerabilities, and potential defense mechanisms through the following research questions:

- How do AI-generated phishing emails bypass traditional email security measures?
- What psychological and behavioural factors influence employees' susceptibility to AI-driven phishing attacks?
- How effective are current corporate security measures in preventing AI-powered phishing threats?
- What AI-driven phishing detection and prevention mechanisms can be developed to enhance cybersecurity defenses?
- How can adaptive security awareness training improve employee resilience against AI-generated phishing attacks?

This study will provide empirical insights into how AI-powered phishing emails evade detection and influence human behaviours.

## Research Objectives

The primary objectives of this study are to:

- Analyze AI-generated phishing techniques and their ability to evade detection.
- Assess employee response rates to AI-powered phishing attacks in a corporate environment.
- Evaluate the effectiveness of current security measures against AI-generated phishing threats.
- Develop AI-driven detection and mitigation strategies to counter evolving phishing attacks.
- Enhance cybersecurity awareness programs by integrating AI-powered adaptive training.

## Expected Outcomes

- Deeper understanding of how AI-generated phishing emails evade traditional security measures.
- Identification of psychological triggers that make employees vulnerable to AI-driven phishing attacks
- Empirical insights on the effectiveness of current corporate security protocols.
- Development of AI driven detection frameworks to counter AI generated phishing threats.

## Methodology

This study aims to investigate how AI-generated emails bypass traditional security measures and manipulate employees into revealing sensitive information. To achieve this, a penetration testing approach will be used, where AI-generated phishing emails are deployed in controlled corporate environment to analyze employee response rate, security vulnerabilities, and the effectiveness of existing security

- **Research Design**

A mixed methods approach will be used, combining quantative analysis and qualitative analysis. The study will be Conducted in three phases:

**Phase-1** – Phishing Email Generation

AI-driven phishing emails will be Created using Natural Language Processing (NLP) models such as GPT-4 and BERT

**Phase-2** – Penetration Testing & Data Collection

AI-generated phishing emails will be sent to participants under controlled conditions to measure Click rates, responses, and detection rates.

**Phase 3** – Evaluation & Security Analysis

Data will be analysed to assess employee susceptibility, the effectiveness of security defenses and the gaps in existing cybersecurity measures.

- AI-Generated Phishing Email Development

1. AI Models Used

The phishing emails will be generated using machine learning models trained in natural language processing. The study will leverage:

- GPT-4 (for generating realistic phishing messages)
- BERT and T5 (for contextual understanding and variation in email text)
- Reinforcement Learning (to optimize phishing email effectiveness based on engagement metrics.

2. Types of Phishing Emails

Three categories of AI-generated phishing emails will be tested:

→ Generic Phishing - Basic deceptive emails (eg fake login)
→ Spear Phishing - Personalized emails using employee information
→ Business Email Compromise (BEC) - Emails impersonating executive or IT departments.

3. Data Collection

3.1 Participants Study.

Participants will include corporate employees from different departments. The sample size will be 50-100 employees, ensuring diversity in roles and cybersecurity knowledge levels.

### 3.2 Deployment of Phishing Emails

- AI generated phishing emails will be sent over 4 weeks in a controlled corporate setting.
- Employee interactions (click rate, response data, and reporting rate) will be tracked.
- No real credential theft will occur - dummy login page will be used for tracking without data collection.

### 3.3 Measuring Employee Responses

The following metrics will be recorded:

- Click rate (percentage of employees who clicked phishing links)
- Response rate (employees who provided sensitive information)
- Report rate (employees who identified and reported phishing attempts)

## 4. Evaluation of Security Measures

The study will assess the effectiveness of existing cybersecurity defenses, including:

- Spam filters & AI based Email Security (How well do existing filters detect AI-generated phishing emails?)
- Email Authentication Protocols (DMARC, SPF, DKIM) (How easily can AI-generated phishing emails bypass security protocols?)

- Employee Awareness Training (Comparison of responses from trained vs. untrained employees).

5. Ethical Considerations

- Informed Consent: Employees will be informed that a cybersecurity study is being conducted that will not be told exactly when phishing simulations will occur to ensure natural responses.

- Data Privacy & Anonymization: No real credentials will be collected; all responses will be anonymized.

- Institutional Review Board (IRB) Approval: The study will adhere to ethical guidelines for cybersecurity research and obtain IRB approval before optimization.

This study will use AI-driven penetration testing to analyze how AI-generated phishing emails bypass security measures and manipulate employees into responding.

# Research Timeline

The research will be conducted over six months, progressing through key phases from planning and execution to analysis and reporting.

- Month 1: Literature Review, Proposal Approval and IRB Clearance

→ Conduct in depth literature review on AI-generated phishing and cybersecurity defenses.

→ Finalize the research proposal and submit it for Institutional Review Board (IRB) approval to ensure ethical compliance.

- Month 2: AI-Generated Phishing Email Development

→ Develop phishing emails using AI models to create generic, Spear phishing and BEC attacks.

→ Configure email tracking tools to monitor employee responses.

- Month 3-4: Deployment & Data Collection

→ Send AI generated phishing emails in controlled batches over eight weeks to measure click rates, response rates and reporting rates.

→ Assess the effectiveness of corporate email security systems in blocking phishing attempts.

Month 5: Data Analysis & Evaluation

→ Analyze employee susceptibility to AI-generated phishing and compare it with traditional phishing success rates.

→ Identify weak points in current phishing detection mechanisms.

Month 6: Report Writing & Recommendations

→ Compile findings into a final research report detailing phishing success rates and security.

→ Present results to cybersecurity professionals, IT teams, and academic researchers.

Research Budget

The estimated budget for this research is $4,000 covering AI model access, phishing tools, cybersecurity software. Breakdown is as follows

| Category | Estimated Cost (USD) |
|---|---|
| AI Model Usage | $500 |
| Phishing tools | $300 |
| Cybersecurity tools | $700 |
| Cloud Computing | $400 |
| Participant Incentives | $1,000 |
| Training & Awareness | $600 |
| Miscellaneous | $500 |
| Total | $4,000 |

## Findings

This study is expected to reveal significant weakness in traditional cybersecurity measures when confronted with AI-generated phishing attacks. AI-powered phishing emails will likely bypass conventional spam filters and authentication protocols due to their ability to generate contextually accurate and highly persuasive messages. It is anticipated that employees will be more susceptible to AI-generated phishing emails compared to traditional phishing attempts, especially when the messages incorporate personalized content, authority cues, and urgency.

Furthermore, existing security awareness training programs may prove insufficient in preparing employees to detect AI-driven phishing attacks, with trained employees still falling victim at a notable rate. The study is also expected to highlight the inadequacies of current AI-based phishing strategies. Overall, the findings will emphasize the urgent need for advanced AI-driven phishing detection mechanisms and adaptive security training programs to counter the growing threat of AI-powered cyberattacks.

## Results and Discussion

This research will present empirical evidence on the effectiveness of AI-generated phishing emails by bypassing traditional security measures and manipulating human behavior. The findings will highlight key vulnerabilities in email security systems, employee awareness levels, and existing cybersecurity defenses.

The result will show that AI-generated phishing emails outperform traditional phishing attempts in terms of bypassing spam filters, evading detection mechanisms, and deceiving employees. It is expected that a significant percentage of AI-generated phishing email will go undetected, demonstrating the limitations of current email security protocols such as DMARC, SPF and DKM. The study will also reveal that employees are more likely to engage with AI-generated phishing emails, particularly when the messages include personalized content, authority cues, or urgency. Higher success rates are expected in spear phishing and business email compromise (BEC) attacks, indicating that AI-generated phishing is most effective when customized to its target.

Overall, the study will emphasize the growing threat of AI-generated phishing and the urgent need for more sophisticated AI-driven cybersecurity solution. The discussions will provide practical recommendations for improving phishing detection systems, enhancing adaptive security awareness training, and integrating AI-powered defense mechanisms to counter evolving phishing threats.

# Future Scope

As AI generated phishing attacks continue to evolve, future research must focus on developing more advanced cybersecurity solutions to combat these threats. This study highlights several areas that require further exploration:

1. AI-Driven Phishing Detection Systems
- Future studies can improve machine learning-based phishing detection models that dynamically adapt to new generated phishing techniques.
- Research should explore adversial AI defense mechanisms, where AI learns to counteract evolving phishing tactics.

2. Adaptive Security Awareness Training
- Future work should develop personalized, AI-driven training programs that evolve based on employee susceptibility patterns.
- Real-time phishing simulations and gamified learning environments can be integrated to enhance cybersecurity awareness.

3. AI for Cyber Defense and Threat Intelligence
- AI can be leveraged not only for phishing detection but also for predictive threat intelligence, helping organizations anticipate new attack patterns before they occur.
- Future research should integrate AI-powered threat hunting tools with existing cyber defense frameworks.

As AI generated phishing evolves, future research must develop adaptive cybersecurity solutions. Integrating AI-driven detection, personalized training and behavioral analytics will be key to strengthening cyber defenses.