AI in Social Engineering and Phishing Campaigns

Team Members: Prabhav Pranay Nerurkar, Omkar Abhay Bhambid,

Nikhil Ganesh Damse

Organized by: Digisuraksha Parhari Foundation

Powered by: Infinisec Technologies Pvt. Ltd.

Abstract

The rapid advancement of generative AI has created a new frontier in social engineering and phishing. AI tools can craft highly persuasive messages, deepfake media, and targeted campaigns that exploit human trust. This paper examines how AI amplifies phishing attacks and proposes a hypothetical AI-based defense tool. We systematically review recent research and industry reports on AI-driven phishing, highlighting techniques such as personalized spear-phishing using large language models and deepfake impersonation. Our methodology combines literature analysis with a conceptual tool design to simulate real-world scenarios. We describe “PhishGuardAI,” an imagined AI-enhanced email filtering system, and present hypothetical evaluation results showing high detection rates. We discuss ethical concerns of AI misuse and the tool’s market relevance, especially given India’s booming digital sector and rising phishing incidents. Our findings underscore the urgent need for advanced countermeasures to keep pace with malicious AI.

Problem Statement & Objective

Social engineering – the art of manipulating individuals into revealing information or performing actions – remains a leading cybersecurity threat. Phishing attacks exploit human trust through deceptive emails, messages, or calls. Traditional defenses (spam filters, user training) are being challenged by advances in AI/ML. Generative models (e.g. GPT-4) and deepfake technologies now allow attackers to produce context-aware, personalized phishing content at scale. Reports confirm a dramatic rise in phishing: for example, recent surveys show over 80% of organizations faced phishing incidents, and firms have seen phishing volumes surge 1265% since late 2022. In India’s financial sector alone, phishing rose by 175% in H1 2024. The core

problem is that AI greatly magnifies the reach and realism of social engineering, making attacks harder to detect and more damaging. The objective of this research is to analyze AI's role in enabling new phishing tactics, review existing literature and cases, and outline a conceptual AI-driven countermeasure. We aim to (1) understand how AI/ML are weaponized in phishing, (2) propose a design for an AI-based defense tool, and (3) assess its ethical impact and relevance, particularly for India's cybersecurity landscape.

Literature Review

Research on AI-enabled social engineering highlights several key areas. Generative AI allows creation of highly realistic content: phishing emails and messages can now mimic corporate communication styles with minimal effort. A systematic review identifies three "pillars" of AI-driven social engineering: (a) Realistic Content Generation – AI models craft human-like text or synthetic media (voices, faces) for impersonation; (b) Advanced Targeting and Personalization – LLMs and data scraping enable tailored messages to specific individuals or groups; (c) Automated Attack Infrastructure – AI automates campaign logistics (e.g. scheduling, multiple channels). For example, LLM-based agents (using GPT-4 and Claude) have been shown to generate spear-phishing emails achieving a 54% click-through rate—on par with human-crafted attacks but at far lower cost. These results confirm that AI tools can produce convincing spear-phishing with minimal human input. Other studies note the rise of malicious "LLM-powered" toolkits (e.g. WormGPT, FraudGPT) which enable non-experts to easily generate phishing messages or code.

Deepfakes are another advancing threat: AI-driven voice synthesis and face-swapping can mimic trusted individuals in calls or video meetings. Attackers have already exploited deepfake audio to trick employees into transferring millions of dollars, and voice-cloning tools are used in "vishing" attacks. In India, reports warn that fraudsters may use deepfakes during virtual meetings to coerce finance teams or capture MFA codes.

Defensive research has focused on AI for phishing detection. Machine learning classifiers (using NLP features or deep learning) can identify phishing emails with high accuracy. For instance, a recent ML model achieved an F1-score of 0.99 on a large phishing dataset by integrating explainable AI techniques for user trust. While detection models exist, the arms race continues, as traditional cues (bad grammar, generic wording) are now absent in AI-generated phishing. Industry reports have observed that as AI-generated phishing bypasses conventional filters, organizations must adopt advanced analytics.

In summary, the literature confirms that generative AI is reshaping social engineering. Researchers predict that AI will soon match or surpass human quality in crafting phishing content. Ethical analyses stress the human impact: AI-enhanced

attacks exploit psychological vulnerabilities like authority and trust, making them harder to counter. The consensus is clear: AI can “weaponize” social engineering and create a more hostile threat landscape, necessitating new countermeasures and awareness.

Research Methodology

This study is conducted through a qualitative analysis of current research, coupled with a conceptual design of an AI-based tool. First, we performed a systematic review of academic papers, industry reports, and news articles on AI in phishing and social engineering (as cited above). This review identified key techniques (LLM-generated text, deepfakes) and trends (phishing volume increases, AI toolkit proliferation). Second, we propose a hypothetical AI tool and outline its intended implementation and evaluation. Since actual coding is beyond scope, our methodology focuses on designing the tool’s architecture and predicting its behavior. We assume standard development practices: using open-source AI models (e.g. GPT-4, BERT embeddings) and public phishing datasets. For evaluation, we simulate a scenario where synthetic phishing emails (generated by LLMs) are fed to the tool to measure detection accuracy. Key metrics include false-positive rate and detection rate. This approach allows us to reason about expected results and limitations without a live deployment.

Tool Implementation

We propose PhishGuardAI, an AI-enhanced email phishing defense platform. The tool’s architecture combines machine learning, NLP, and knowledge-based analysis to flag suspicious messages. Key components include:

- Data Collection: Inputs are email content (text, headers) and attached media. The system retrieves external context (e.g. sender metadata, known domain reputation).

- Preprocessing: Emails are tokenized; NLP techniques extract semantic features. Phishing indicators (links, attachments, urgent phrases) are parsed.

- AI Analysis: A large language model (e.g. GPT-4) is used in a reverse-processing mode: the email content is fed as a prompt to gauge unnatural language patterns. Simultaneously, a transformer-based classifier (such as BERT fine-tuned on phishing data) evaluates the text.

- Detection Engine: Outputs from AI models are combined in an ensemble classifier (e.g. gradient-boosted trees) that computes a phishing likelihood score. Explainable AI methods (LIME/SHAP) highlight which phrases or links

contributed to the score.

- Response Module: If the score exceeds a threshold, the email is quarantined or flagged.

Implementation Tools: In a real-world build, we would use Python with libraries like scikit-learn and TensorFlow, APIs for LLM (OpenAI/GPT API or similar), and email processing frameworks. A sample pipeline might involve: extracting features via spaCy or nltk, generating embeddings with Sentence-BERT, and training a Random Forest classifier.

Results & Observations

In our hypothetical evaluation, PhishGuardAI is tested on a simulated phishing campaign. We generate 1,000 test emails: a mix of benign business messages and sophisticated phishing attempts created by GPT-4 using real public data. The tool flags 970 emails correctly as benign or phishing, yielding a detection accuracy of 97%. The false-positive rate is 2%, while the false-negative rate is 3%. These figures are comparable to recent research.

Notably, AI-generated phishing emails that include deepfake media or unusual calling patterns were all correctly identified. The LLM component helped catch emails where social cues were subtly crafted. The Explainable AI overlay revealed that phishing URLs and mismatched sender domains were key indicators in most detections.

Observations: Our simulated results suggest that an AI-driven tool can significantly outperform traditional keyword or rule-based filters against generative phishing. However, some limitations emerge: highly contextual spear-phishing occasionally bypasses the classifier, indicating the need for continual retraining. Additionally, adversarial tactics require supplementary modules.

Ethical Impact & Market Relevance

The dual use of AI in social engineering presents a significant ethical dilemma. On one hand, AI-generated phishing campaigns exploit personal trust and privacy. The ease of automating attacks lowers entry barriers for criminals, potentially increasing the volume of cybercrime and eroding trust in digital communications. There are privacy concerns even in defense: an AI scanning employee emails must be carefully governed to avoid undue surveillance.

On the positive side, ethical AI tools can bolster security. PhishGuardAI is conceived as a white-hat measure to protect organizations. By simulating advanced phishing

for training and filtering attacks, it empowers users without misusing data. Developers must ensure transparency and fairness to maintain trust.

The market relevance of AI in phishing is high. Globally, phishing attacks account for a large share of breaches, costing organizations millions per incident. In India, the need is acute: the financial sector saw phishing escalate by 175% recently, and the average breach cost in India is reported around USD 2.18 million. The Indian cybersecurity market is rapidly expanding, with growth driven by digitalization. Government initiatives are raising security standards, and Indian firms seek AI-enhanced solutions.

Future Scope

AI-driven social engineering is an evolving field, and future research is crucial. Potential extensions of this work include:

(1) Adversarial AI Analysis: Studying how attackers might adapt and developing robust models resilient to adversarial inputs.

(2) Multi-Modal Integration: Expanding the tool to analyze voice and video messages using emerging biometric analysis and anomaly detection.

(3) Explainable & User-Friendly AI: Integrating transparent AI so that end-users and security teams understand why an email is flagged.

(4) Cross-Organizational Data Sharing: Designing privacy-preserving ways to share threat intelligence between companies.

(5) Policy and Education: Research on how to educate users about AI threats and on crafting policies/regulations that govern the use of generative AI responsibly.

References

- Schmitt, M. and Flechais, I., *"Digital deception: generative artificial intelligence in social engineering and phishing."* Artificial Intelligence Review, vol. 57, article no. 324 (2024).link.springer.comlink.springer.com

- Bharati, R.K., *"AI-Enhanced Social Engineering: Evolving Tactics in Cyber Fraud and Manipulation."* The Academic, Vol. 2 Issue 7, July 2024.

- Heiding, F., Schneier, B., and Vishwanath, A., *"AI Will Increase the Quantity — and Quality — of Phishing Scams."* Harvard Business Review (May 30, 2024).

- Perception Point, *"Detecting and Preventing AI-Based Phishing Attacks: 2024 Guide."* (Security blog post).

- Arntz, P., *"AI-supported spear phishing fools more than 50% of targets."* Malwarebytes Labs (Jan. 7, 2025).malwarebytes.commalwarebytes.com

- SlashNext Threat Labs, *"2023 State of Phishing Report."* Press Release (Oct. 30, 2023).slashnext.com

- Times of India, *"Govt-backed report warns of threats to BFSI from deep fakes and AI-generated content."* (Apr. 7, 2025).timesofindia.indiatimes.com

- Business Standard, *"India sees 135,173 financial phishing attacks in H1 2024, says study."* (Nov. 18, 2024).business-standard.com

- 63SATS Cybertech, *"CERT-In Flags Rising Cyber Threats in India's BFSI Sector."* (Apr. 11, 2025).63sats.com63sats.com

- Al-Subaiey, A. et al., *"Novel Interpretable and Robust Web-based AI Platform for Phishing Email Detection."* arXiv:2405.11619 (May 2024).arxiv.org

- IBM Security Intelligence, *"Click-through rate for AI-generated phishing is 11% vs 14% for humans."* (Oct. 2023).ibm.com

- Data Security Council of India (DSCI), *"India Cybersecurity Domestic Market 2023 Report."* (2023) – India's market grown to USD 6.06B in 2023.dsci.in