# PHISHING WITH LARGE-LANGUAGE MODELS

**Felicity Huang**
Department of Computer Science
Stanford University
huangfe@stanford.edu

## ABSTRACT

This paper will examine the relationship between Large Language Models (LLMs) and text-based Phishing attacks. It will briefly summarize the background for LLMs and phishing, a common and costly cyberattack, analyze offensive and defensive interactions the two, as well as methods for risk mitigation on the government, corporate, and consumer level.

## 1 BACKGROUND

### 1.1 SECURITY TOPIC: PHISHING

Phishing is a type of social engineering cyberattack where the attacker masquerades as a trustworthy entity through emails or other communications to trick victims into actions like visiting a malicious website, unknowingly downloading malware, or revealing sensitive user data like login credentials and credit card numbers.

Phishing attacks can come from average cybercriminals seeking monetary gain, individuals or groups with political agendas, or government-sponsored actors seeking intelligence or disruption of specific entities. These attacks can be carefully engineered to appear legitimate and convincing and exploits people's trust, "fear, curiosity, a sense of urgency, and greed to compel recipients to open attachments or click on links" (Cisco) with significant impact. According to the Cost of a Data Breach Report 2023 by IBM Security, "the global average cost of a data breach in 2023 was USD 4.45 million, a 15% increase over 3 years" (IBM Security (2023)). Phishing can lead to financial loss, fraud, reputation damage, operational disturbance, and legal penalties for failing to ensure security.

**Types of Phishing include:** (Cisco)

1. Business email compromise (BEC) attacks impersonate or hack trusted employees, vendors, companies asking for payment or sensitive information.

2. Email account takeover (ATO) attacks commonly come as a fake security email from the email provider (like Google or Microsoft) to reset their password or fix other account problems that needs their attention, along with a URL which leads to a fake login page.

3. Spear phishing targets a specific individual or group, which allows for customized communications that appear more authentic. According to the SANS Institute, 95 percent of all attacks on enterprise networks are the result of successful spear phishing.

**Real-world Cases:**

1. In 2013, Target experienced one of the largest data breaches in history resulting from a phishing attack on one of its third-party vendors that compromised over 40 million credit cards. Attackers obtained login credentials through employees of Fazio Mechanical Services and installed malware on Target's point-of-sale (POS) systems (Committee on Commerce, Science, and Transportation (2014)).

2. In 2014, Sony Pictures Entertainment suffered a high-profile breach resulting in over $100 million in losses. The attackers forged Apple ID verification emails and discovered matching credentials to Sony network, stealing over 100 Terabytes of confidential data, including unreleased films and executive emails (Town of Hempstead).

3. Even tech leaders can fall prey to phishing. Google and Facebook were scammed over $100 million from 2013 to 2015 by attackers impersonating a partner company, Quanta Computer, a hardware manufacturer based in Taiwan. They set up a company in Latvia with the same name and forged invoices, contracts, letters, corporate stamps to extract payments over the years (Town of Hempstead).

4. This year, a Hong Kong finance worker transferred more than $25 million to attackers impersonating his chief financial officer and other colleagues over video conference calls using deepfake technology (Murphy (2024)).

### 1.2 Emerging Technology: Large Language Models

Large Language Models are AI models for natural language processing and text generation. Extensive training on vast datasets of human text from the internet, books, and other sources allows them to mimic human intelligence. Although LLMs still encounter problems such as hallucination, bias, and misinformation, their capabilities have rapidly advanced in the past few years. LLMs can be prompted to generate contextually relevant text and fine-tuned to specialize in tasks, making them versatile and applicable in many areas like content creation, chatbots, virtual assistants, and programming.

In 2024, LLMs are at a stage of burgeoning growth and prevalence. Tech companies have developed intelligent, high-performing models like ChatGPT which is widely-used for all sorts of tasks. LLMs have soaring adoption across industries, from customer service, education, healthcare, entertainment, to finance, and have transformative potential across society. Ongoing research continues to improve the accuracy and performance of LLMs while addressing security and ethical considerations.

## 2 Interaction Risk Analysis

**Offensive Risks:**

LLMs increase the effectiveness (convincing messages) and efficiency (faster and less effort) of phishing. The National Cyber Security Centre assesses that the continuing commoditization of AI will exacerbate cyberthreats as AI lowers the barrier for hackers and elevates effectiveness and efficiency of reconnaissance and social engineering (National Cyber Security Centre). LLMs can easily generate convincing replicas of legitimate websites, personalized interactions by scouring a target's social media activities, contacts, and internet data, and set up automated correspondence (MalwareBytes). LLMs increase the chance of success of traditional phishing attacks and the scale by gathering intel and creating content at a faster pace (Schafer (2024)).

The development of AI led to the White House's Executive Order on AI safety in 2023, requiring rigorous safety testing and government supervision. Microsoft identifies multiple LLM-themed tactics, techniques, and procedures (TTPs) (Microsoft Threat Intelligence (2024)). Ones relevant to phishing attacks include LLM-enhanced reconnaissance, scripting, development of tools and programs, and social engineering. Microsoft has identified and disabled accounts of multiple threat actors employing LLMs for malicious operations, including Forest Blizzard (Russian military intelligence), Emerald Sleet (North Korean), Crimson Sandstrom (Iranian), Charcoal Typhoon (Chinese state-affiliated) (Microsoft Threat Intelligence (2024)). But while AI companies and regulation push for responsible AI and have taken defensive actions to forbid malicious usage, unrestricted counterparts can still be found on the dark web or created by cybercriminals, such as WormGPT or FraudGPT (Schafer (2024)).

**Defensive Mitigation:**

LLMs can potentially provide defensive assistance to security in areas like proof-of-concept attacks generation, malware recognition, and phishing detection (MalwareBytes). However, these tools require quality data and training, and the effectiveness of AI-backed security tools are currently unknown. A 2024 study on Vulnerability Detection with Code Language Models found "considerable gap between current capabilities and the practical requirements for deploying code LMs in security roles, highlighting the need for more innovative research in this domain" (Ding et al. (2024)). AI security solutions can strengthen with further research and enable advanced detection and prevention

techniques as threats evolve. Tech companies are already developing AI-backed products against phishing. Cisco Secure Email Threat Defense, for example, uses unique AI models with NLP to efficiently detect threats and enforce policy to reduce phishing response times (Cisco).

## 3 MITIGATIONS

The security risks posed by LLMs can be mitigated with action on multiple levels, government regulation, company level, to individuals.

Government regulations for security tend to be vague and inactionable, but research must be done to enforce direct security guidelines for companies.

On the company level, first, AI service providers must enforce security policies to prevent malicious use potentials and their services from abuse, data poisoning, model theft (MalwareBytes). For example, Microsoft and OpenAI's AI and Cybersecurity principles include detecting, tracking, and terminating AI service usage of threat actors, notifying and collaborating with other AI service providers and stakeholders, and remaining transparent about the nature and extent of threat and actions taken (Microsoft Threat Intelligence (2024)). This would mitigate risks to a degree but does not prevent attackers from using LLMs they obtain elsewhere or build themselves.

Secondly, mailbox providers must implement strong phishing detection and content filtering to stop malicious messages from even reaching inboxes. Nick Schafer, Senior Manager of Deliverability and Compliance at Sinch Mailgun, proposes DMARC (Domain-based Message Authentication, Reporting, and Conformance) as the primary defense (Schafer (2024)). DMARC is an email authentication protocol for email domain owners to protect their domain from unauthorized use like phishing and email spoofing. When fake content becomes virtually indistinguishable from real, automatic verification of senders becomes important. Mailbox providers like Gmail and Outlook should enforce bulk sender requirement, while reputable domains should specify DMARC framework and policies for receivers to authenticate messages and reject, quarantine, or report spoofed domains (Schafer (2024)).

Lastly, for general organizational defense against phishing, all companies must invest in multi-layered security, mitigate human vulnerabilities, and reduce damage from third-party breaches. Phishing attacks involve multiple stages, from the initial message which exploits human vulnerabilities, malicious links or malware, to the ultimate infiltration and exfiltration or sensitive actions like money transfers (Cisco).

1. Technical protections must include Multi-factor authentication (via passKeys, hardware security keys, authenticators, etc), strong anti-malware detection/endpoint protection, robust firewalls and antivirus software, and isolation and encryption of sensitive data. Transactions must have authentication and approval process (Murphy (2024)).

2. Human protections include employee training and phishing drills, timely-report guidelines, and monitoring the web for credential theft. Employees should take care to use unique passwords for work-related services and devices.

3. Companies should also ensure security of third-party vendors and partners with vetting audits and security ratings and reduce third-party access and power on the company's network.

On the consumer level, mitigation responsibility falls heavily on communication providers to incorporate stronger security and spam detection and user education on AI phishing risks. People can check hyperlinks and senders, monitor accounts regularly, avoid clicking links from unknown and seemingly legitimate sources, never send personal information over email, and be wary of phishing persuasion strategies. However, the average user, especially the elderly or less technically inclined, will struggle to avoid phishing on their own.

## REFERENCES

Cisco. What is phishing? URL https://www.cisco.com/c/en/us/products/security/email-security/what-is-phishing.html.

Committee on Commerce, Science, and Transportation. A "kill chain" analysis of the 2013 target data breach, 2014. URL https://www.commerce.senate.gov/services/files/24d3c229-4f2f-405d-b8db-a3a67f183883.

Yangruibo Ding, Yanjun Fu, Omniyyah Ibrahim, Chawin Sitawarin, Xinyun Chen, Basel Alomair, David Wagner, Baishakhi Ray, and Yizheng Chen. Vulnerability detection with code language models: How far are we?, 2024. URL https://arxiv.org/abs/2403.18624.

IBM Security. Cost of a data breach report 2023, 2023. URL https://www.ibm.com/reports/data-breach.

MalwareBytes. Ai in cyber security: Risks of ai. URL https://www.malwarebytes.com/cybersecurity/basics/risks-of-ai-in-cyber-security.

Microsoft Threat Intelligence. Staying ahead of threat actors in the age of ai, 2024. URL https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

Shannon Murphy. A deepfake scammed a bank out of 25m — now what?, 2024. URL https://www.trendmicro.com/en_us/research/24/b/deepfake-video-calls.html.

National Cyber Security Centre. The near-term impact of ai on the cyber threat. URL https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.

Nick Schafer. The golden age of scammers: Ai-powered phishing, 2024. URL https://www.mailgun.com/blog/email/ai-phishing/.

Town of Hempstead. Famous phishing incidents from history. URL https://hempsteadny.gov/635/Famous-Phishing-Incidents-from-History.