DeceptAI: Adaptive Adversarial Intelligence for Spear Phishing Simulation and Detection Training

Report by

SIDHARTH K

In Partial Fulfillment of the Requirements

for the Degree Of

MSc. Computer Science with Specialization in Cyber Security

Supervisors: Dr. Preetam Mukherjee (DUK), Arumugam Ganapathi (QG)

School of Computer Science and Engineering

KERALA UNIVERSITY OF DIGITAL SCIENCES, INNOVATION AND

TECHNOLOGY

15-05-2025

Abstract

The study describes the development of an AI-based spear phishing attack and detection system which addresses cybersecurity training needs for organizations. This system uses Large Language Models together with V-Triad principles to produce realistic phishing emails at three security levels that advance from basic to complex and require employees to evaluate credibility marks as well as compatibility and customizability factors. Each component of the phishing simulation operates independently through a modular agent-based framework that manages data collection and email setup together with content development along with link inserted and feedback adjustment and monitoring processes. The email examination process utilizes both semantic content evaluation and visual confirmation and real-time link evaluation procedures and simulation-specialized rule-based analysis techniques to determine potential security risks. The project functions to assess phishing risks and provides immediate detection feedback combined with detailed descriptions to users in order to build their security understanding through realistic simulations of actual cyberattacks that happen in controlled settings alongside best-practice threat analysis guidance. Despite continuous learning implementation and explainable AI technologies the system keeps its responsible character while maintaining effectiveness for training purposes thanks to its ethical safeguards.

Acknowledgment

This project required a lot of guidance and assistance from many people, and I am incredibly fortunate to have had this support throughout the project. First of all, I would like to thank **Dr. Preetam Mukherjee**, Course coordinator M.Sc. Computer Science with Specialization in Cyber Security, Chair of School SoCSE, and my project guide, for his guidance and support throughout the project. His valuable insights were instrumental in shaping this project. I am also grateful to all the Professor for providing the necessary resources and facilities for the project.

I would also like to thank **Arumugam Ganapathi**, Technical Architect, Quest Global, for his wholehearted support throughout the implementation of the project. Additionally, I would like to express my gratitude to **Vinod Subramonia Pillai**, Principal Architect, Quest Global, for his invaluable assistance. I would also like to extend my thanks to the wonderful team at **Quest Global** for their support. Finally, I would like to thank my family, friends, and all who have provided me with guidance, support, and encouragement throughout the project..

Certificate

This is to certify that the report **DECEPTAI: ADAPTIVE ADVERSARIAL INTELLIGENCE FOR SPEAR PHISHING SIMULATION AND DETECTION TRAINING** submitted by **Sidharth K (Reg. No: 231040)** in partial fulfillment of the requirements for the award of **MSc. Computer Science with specialization in Cyber Security** is a bonafide record of the work carried out at **Kerala University of Digital Sciences, Innovation and Technology** under my supervision.

Supervisor

Dr. Preetam Mukherjee

Assistant Professor

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Cybersecurity risks have escalated to previously unheard-of levels of complexity in today's constantly evolving digital environment, with spear phishing emerging as one of the most dangerous attack methods. Spear phishing is a targeted type of social engineering that uses behavioral heuristics and cognitive biases that are inherent in people to trick them into disclosing private information or taking illegal actions. Traditional phishing campaigns used generic templates, but contemporary adversaries use artificial intelligence (AI) and sophisticated natural language processing (NLP) tools to create context-aware, highly personalized messages that nearly perfectly mimic authentic communications[1].

The understanding of the offensive and defensive aspects of cybersecurity has been completely transformed as a result of recent advancements in large language models (LLMs). In addition to producing text that is almost identical to human writing, advanced models like GPT-4 can also replicate the complex nature of interpersonal communication. This capability enables adversaries to obtain minimal information about targets, such as social interactions, recent purchases, or organizational context, and turn that information into persuasive spear phishing emails that get past standard technical protections. By ensuring that emails achieve high credibility, compatibility, and customizability thereby effectively triggering victims' psychological and cognitive vulnerabilities advanced frameworks, such as the V-Triad, further refine the art of deception[1].

In the meantime, initiatives like Microsoft's Agent AI have shown that AI can be used for defensive purposes, such as threat simulation and detection in real time. Organizations can simulate extremely complex attack scenarios and train staff to identify subtle, context-dependent indicators of malicious intent by incorporating these proactive AI systems into cybersecurity infrastructure. This project makes use of LLMs' cutting-edge capabilities in a unique hybrid framework that blends well-known human-centric models with

algorithm-driven email generation. Through sophisticated intent analysis and behavioral pattern recognition, the method improves real-time detection measures while also speeding up the development of realistic spear phishing campaigns for training.

The future of spear phishing using AI technology brings powerful threats which create a distinct set of possibilities. Attackers keep improving their techniques through AI implementation to achieve reduced expenses while reaching higher success rates so defenders must launch parallel strategic developments. This report develops an inventive application based on LLM capabilities to detect spear phishing attacks through computer-based simulations which operate within legal boundaries. The goal of this project seeks to join human mistakes with technological progress which creates a cyber-secure culture whereby automated systems collaborate with trained users to fight present-day digital attacks[1, 2]. The developed AI-based framework demonstrates achievements at both theoretical cybersecurity defense research and practical organizational preparedness to deal with AI-driven cyberattacks through dynamic phishing tactics adaptation.

1.2 Main Area of Report

This project lies at the convergence of artificial intelligence, social engineering, and cybersecurity training. The primary contributions can be broadly classified into two domains:

1.2.1 AI-Driven Attack Simulation

This report studies how large language models (LLMs) produce spear phishing emails which maintain authentic characteristics with contextual personalization. The system creates content through the V-Triad framework which takes advantage of human cognitive biases while implementing heuristics and compatibility and credibility values. Agent-based design models offensive and defensive operations which implement AI tools based on Microsoft's Agent AI platform to deploy sophisticated machine learning algorithms for protection against cyberattacks. Agent AI makes use of large language models to build highly authentic phishing scenarios that help identify security weaknesses. The system gives defenders and attackers an effective weapon through its capacity to produce customized security content that benefits both offensive and defensive cybersecurity operations. The tools function as adversaries through their ability to create benign phishing attacks that enhance overall security measures. These simulated attacks incorporate:

- Personalization Techniques: The system utilizes few data points from users to duplicate trusted communication methods and produce messages that match their situations.

- Query Optimization: The output's linguistic aspects receive optimization through repetitive feedback processes to enhance language style together with urgency indicators and triggers that match social contexts.

- Comparative Generation Methods Experimental studies demonstrate how the V-Triad improves GPT-4 and other LLM phishing outputs through analysis of samples produced by each technology type with measurements of both impersonation effectiveness and click-through rates[1, 2].

1.2.2 Intelligent Phishing Detection

Creating advanced AI detection methods constitutes the main objective of this project since it aims to outpace adversaries exploiting large language models through spear-phishing attacks. The defense uses different detection domains to protect multiple aspects of phishing sophistication in a live system that adjusts its security measures automatically. These methodologies integrate:

- Semantic Content Analysis: The analytic feature of Semantic Content Analysis uses deep contextual understanding which LLMs acquire to interpret both visible email text as well as subtle semantic and contextual signals beyond a reader's ability. A comparison between how a sender normally writes and the writing style of the current message leads to warning activation. This module operates by continuous training with both normal corporate messages and verified phishing examples so it maintains awareness of changing attacker communication methods.

- Behavioral and URL Validation: URLs in phishing campaigns now hide behind complex URL redirection networks along with homograph attacks and abbreviated and deceptive URLs. The framework references every linked URL to dynamic threat intelligence databases before running simulated click tests in an isolated environment alongside browser instrumentation for detecting hidden redirects or harmful file downloads. An examination of the email's HTML structure through structural analysis helps identify hidden iframes and obfuscated JavaScript that might trigger drive-by downloads.

- Email Header and Metadata Analysis: A special tool called the Header Analysis Agent performs analysis of email headers to check routing metadata and authentication results from SPF/DKIM/DMARC and search for irregularities in mail-transfer paths. The agent identifies forged relay hops and suspicious time skew artifacts by comparing header fields like Received entries and Message-ID formats and date/time stamps against accepted patterns for internal and partner domains. The analytical risk score contains metadata attributes that include X-Mailer indicators and encoding inconsistency patterns.

- Quantitative Comparative Analysis: The detection efficiency of our system uses precision recall F1-score and Click-Through Risk (CTR) to measure the effectiveness rigorously. The monitoring system utilizes three benchmark models including signature engines and expert human reviewers together with ML classifiers. Our methodology uses simulating identical phishing attacks to determine exactly how soon and with what precision our system detects threats and what number of false alarms our security teams can expect to save time on.

- Hybrid Feedback Systems: The platform incorporates Hybrid Feedback Systems through an iterative process which includes confirmed phishing incidents that end users report as well as incidents automatically detected by the system and reprocessed for training purposes. The automated detection heuristics and model parameters update process runs within a lightweight online pipeline to preserve efficiency and high performance levels while adversaries modify their tactics.

The integration of these two domain areas produces an advanced simulation framework for spear phishing while simultaneously providing organizations with better detection-capabilities and response tools. This detection solution combines artificial intelligence attack simulation methods with intelligent phishing recognition capabilities to create an all encompassing system that both models sophisticated spear phishing attacks and allows organizations to locate and combat them effectively. A complete strategy which includes semantic analysis together with behavioral emulation and header investigation and quantitative metrics along with continuous learning significantly strengthens the defense against cyber threats using modern AI-powered attack technology. The document provides a total security solution which confronts modern phishing trends to enhance organization-wide cybersecurity preparedness against complex social engineering attacks.

1.3 Motivation

This project derives its purpose from the shifting security threats and the urgent requirement to strengthen defenses against developing phishing techniques. Phishing attacks launched through spam improved into sophisticated spear phishing attacks during recent years by employing personal and organizational data to trick users. Through the convergence of traditional social engineering approaches with digital resources sophisticated attackers create complex obstacles for standard rule-based detection systems which struggle to detect harmful objectives. Research confirms that a substantial number of cyberattacks use phishing methods because of their importance for security improvements[1].

Standard employee training methods implemented with generic simulations cannot adequately showcase the complex real life characteristics of attacks that happen in actual phishing scenarios. Workforce training contains standardized template attacks which fail to reproduce the complex psychological phishing methods used by current threat groups. Both of these shortcomings create limitations to awareness program success while simultaneously enabling users to believe they are adequately protected. The combination of human inspired V-Triad and large language models running under AI control provides a possible method to create simulations which accurately represent real spear phishing attacks. The simulated attacks would offer more real life precision enabling users to discover hidden indications which usually escape their notice[2].

Security detection systems encounter two main challenges: they produce extensive numbers of incorrect alerts and fail to detect complex indications of phishing that build upon context. The systems currently face challenges in identifying genuine from malicious intent since attackers create messages that perfectly duplicate

trusted communication channels. Recent developments in AI together with natural language processing capabilities enable deep semantic analysis to detect indicators which exist past superficial patterns. The proposed system leverages these capabilities to simulate attacks and at the same time enhance defenses through real time evaluation of email text and URLs and behavioral patterns.

This project advances because testing organizational resilience and maintaining its improvement requires ethical testing in controlled settings. Organizations require ethical testing frameworks to validate their defenses while reducing the exposure to unnecessary risks during these assessments. The dual phase system functions as a complete platform to support ongoing evaluation and improvement of security training combined with technological advancement. Organizations that synchronize their practical training programs using current research publications achieve security defenses which adapt to fresh threats in real time.

When AI-based attack simulation operates with intelligent detection mechanisms it provides an effective approach to eliminate differences between offensive phishing methods and defensive response strategies. The proposed research addresses the necessity to develop a strong cybersecurity solution which unites human awareness training with current technology advancements to block advanced phishing attacks thereby creating a protected digital space for organizations and users.

1.4   Organization of the Report

The remainder of this report is structured into five main chapters that collectively build an in-depth exploration into AI-driven spear phishing simulation and detection, as well as the integration of behavioral frameworks to enhance cybersecurity training and resilience.

Chapter 2: Literature Review

This section presents a complete evaluation of studies regarding phishing attacks together with social engineering methods and cyber training progression. The research examines vital phishing studies regarding both general phishing methods and spear phishing strategies along with the development from manual human-based methods to fully autonomous AI-based methods. This part investigates the innovative developments in natural language processing along with large language models (LLMs) which brought a revolutionary shift to modern phishing attack planning and implementation. Different psychological behavior frameworks like V-Triad are used in the review to study and duplicate the psychological aspects used by attackers in these attacks[1, 2].

Chapter 3: Methodology

This chapter details the design and implementation of the proposed dual phase system. It begins with an overview of the agent-based architecture that underpins both the attack simulation and detection components. The chapter describes how the V-Triad framework is implemented to ensure that generated phishing emails are not only personalized and contextually relevant but also psychologically compelling. Additionally, the

integration of cutting edge LLMs (e.g., GPT-4) is explained, including the development of customized query templates and iterative feedback mechanisms that fine tune email quality. Critical aspects such as experimental setup, participant recruitment, and ethical considerations are also articulated, providing the foundation for a robust and reproducible evaluation process.

Chapter 4: Results and Discussions

An extensive performance evaluation takes place in this chapter for the system across its attack simulation and phishing detection phases. The evaluation provides specific numerical performance results such as click-through rates and precision and recall measurements together with user feedback analysis. A comparative evaluation takes place which includes phishing messages composed through AI alone, V-Triad-created messages and mixed AI and V-Triad hybrid model phishing content. The author presents a detailed explanation of how regular semantic content updates with URL security tests improve the quality of simulated phishing incidents while also enhancing detection systems. The report performs detailed evaluations of the presented findings by analyzing them regarding their impact on modern cybersecurity threats and their dual applications for offensive and defensive activities.

Chapter 5: Conclusions and Future Work

A concluding chapter of the report brings together all important findings by presenting the theoretical accomplishments and practical implementations discovered throughout the project. This approach demonstrates how it connects standard methods of cybersecurity with technical AI-based practices. The upcoming section of this work provides research guidelines for upcoming studies that will enhance automated data collection capabilities and detection system efficiency while establishing execution plans for real-world business settings. The report explores ethical implications and social consequences of these advanced systems while creating guidelines for secure cybersecurity innovation.

The report provides a structure that guides readers from the beginning through its research starting with an extensive review of established works about phishing attacks combined with social engineering. The report showcases the evolution from manual cybersecurity education to cutting-edge solutions built on Artificial Intelligence with particular focus on Large Language Models. The methodology specifies an approach that combines two phases with behavioral framework implementation for realistic phishing attack detection using the V-Triad frameworks. The system evaluation uses both measuring statistics alongside user experience evaluations while comparing strategies to determine optimal defensive approaches in the results and discussion. Together with future work section brings together major findings which connect traditional cybersecurity frameworks with AI while establishing future goals that incorporate ethical implications of advanced systems deployment. This organized system provides with comprehensive insights about the challenge while clarifying proposed innovative solutions alongside effects for increased cybersecurity in modern digital environments.

Chapter 2

Literature Review

2.1 Introduction

The original phishing attacks from the early 1990s targeted wide user groups through easily detectable spoofed messages before undergoing significant developments. The surge of internet users led attackers to develop spear phishing techniques starting from 2010 which delivered superior success rates thanks to social engineering and customized attack content that lowered spam from 80% despite targeting specific individuals[3]. Attackers use only a small amount of organizational affiliation and purchase history or social media data to generate these perfect copycat messages according to recent analysis of corporate phishing cases.

These attacks succeed through psychological methods which derive from traditional persuasion methodologies. These principles primarily exploit cognitive heuristics through the systematic implementation of reciprocity with "free gift" lures as well as authority by executive impersonation along with social proof through fabricated endorsements and scarcity through limited-time offers[4]. Phishing email research reveals that 97% of effective spear phishing messages use two or more of Cialdini's influence principles thus demonstrating the importance of understanding these advanced methods in training and detection approaches[5].

Multiple organizations practice security awareness training yet maintain standardized phishing templates despite their inability to mimic active cyber threats and their inability to deliver customized content. According to systematic reviews the effect of training programs produces only small improvements in user detection abilities due to continued challenges related to identifying situational cues as well as resisting advanced spear phishing scenarios[6, 7]. Research from the industry sector reveals that excessive repetition of irrelevant phishing simulation exercises leads employees to lose interest in training which results in reduced long term effectiveness and possibly more false positive responses.

On the technical front, signature-based filters remain the backbone of many email security systems, yet they struggle with zero day and polymorphic phishing variants. Machine learning classifiers ranging from decision

trees and ensemble methods to deep neural networks have demonstrated substantial gains in detecting phishing URLs and email content, with precision and recall often exceeding 95% on benchmark datasets[8, 9]. More recently, transformer based models and LLMs such as GPT-4 and Gemini have shown promise in semantic intent analysis, capable of identifying subtle indicators of deception beyond surface level features[10, 11]. However, balancing detection sensitivity with user experience remains a challenge, as overly aggressive filters can generate high false-positive rates, eroding trust in defensive tools.

This chapter synthesizes these strands of research tracing the evolution of phishing, unpacking the psychological mechanisms of social engineering, critiquing the limitations of current training paradigms, and surveying advanced detection approaches to establish the foundation for an integrated AI-driven framework that both simulates and counters sophisticated spear phishing attacks.

2.2 State-of-the-art approaches

2.2.1 Phishing Attack Evolution

Phishing began in the early 1990s as generic, mass distributed emails that cast a wide net, relying on simple spoofing of well-known brands and obvious grammatical errors to trick unsophisticated users. These "first-generation" attacks achieved modest success rates but were easily flagged by both spam filters and attentive recipients[12]. By around 2010, adversaries had refined their methods into "spear phishing," in which messages are tailored to specific organizations or departments. Spear phishing campaigns leverage minimal personal or corporate data such as employee names, project details, or recent company events to craft emails that appear contextually legitimate. Studies show these targeted attacks can bypass conventional filters by mimicking internal communications and exploiting trust relationships within enterprise environments[13, 14]. A further evolution "whaling" focuses on high value targets, such as C-suite executives and board members. Whaling attacks employ sophisticated social engineering techniques, often combining executive impersonation with urgent requests or false legal notices to pressure victims into divulging sensitive data or approving fraudulent transactions. Research indicates whaling campaigns have higher success rates due to the elevated privileges of their targets and the attackers' use of multiple persuasion principles simultaneously[15].

Concurrently, "contextualized phishing" has emerged, where attackers leverage real time organizational events such as mergers, software roll-outs, or tax season to lend credibility to their lures. Caputo et al. documented a marked uptick in such context-aware campaigns, showing that emails referencing authentic company jargon or ongoing business processes achieved click-through rates up to 25% higher than generic templates[16]. Similarly, studies of organizational email behavior reveal repeat clickers employees who fall for phishing multiple times often respond to campaigns that align with their routine workflows, underscoring the importance of dynamic training approaches[17]. The advent of generative AI has accelerated phishing evolution into its latest

phase. Attackers now use large language models to generate highly polished emails at scale, often in a fraction of the time required for manual crafting. Industry data shows AI-assisted phishing emails can be produced up to 40% faster and yield success rates approximately 37% higher than traditional methods, as generative models flawlessly imitate human writing style and organizational context. Experiments with ChatGPT generated phishing demonstrated that even well trained employees fell victim at rates comparable to or exceeding conventional spear phishing. Looking ahead, researchers warn of multi-modal phishing that combines AI-generated text with deepfake audio or video to create immersive social engineering attacks. Taxonomies of generative AI misuse highlight how adversaries can blend text, image, and voice techniques to craft "phish-and-vish" campaigns, where a fake video conference invite is followed by a persuasive email that leverages social proof and scarcity cues. As these advanced threats materialize, organizations must adopt adaptive, AI-driven simulation and detection frameworks to keep pace with the ever evolving phishing landscape.

2.2.2   Social Engineering Frameworks

Overview of Social Engineering Frameworks

A variety of models and frameworks have been developed to decompose and systematize how social engineering attacks are planned, executed, and countered. By understanding these frameworks, researchers and practitioners can both design more realistic phishing simulations and develop targeted defenses.

Foundational Frameworks

1. Social Engineering Attack Framework (SEAF)

    Francois Mouton formalized the Social Engineering Attack Framework (SEAF), addressing limitations in Kevin Mitnick's original attack cycle by adding temporal and flow components. SEAF breaks an attack into six core phases Attack Formulation, Information Gathering, Preparation, Relationship Development, Exploitation, and Debrief each mapped to an ontological model defining the key entities (attacker, target, compliance principles, techniques, media, and goals)[18]. Subsequent work demonstrated how SEAF can be used to generate standardized attack scenarios and map historical incidents for training and countermeasure development[19].

2. SEADM: Detection Perspective

    Building on SEAF, Mouton and colleagues proposed the Social Engineering Attack Detection Model (SEADM), which aligns systems engineering life-cycle stages with SEAF's phases. SEADM provides a blueprint for embedding detection controls at each stage such as anomaly monitoring during Information Gathering or multi factor verification in the Exploitation phase thereby enabling defenders to interrupt attacks before they succeed[20].

Persuasion-Centric Models

3. Anatomy of Social Engineering Attacks

   Bullée dissected 74 real world attack scenarios to quantify the use of persuasion principles (reciprocity, authority, social proof, consistency, liking, scarcity). Their literature based analysis found that authority was the most frequently exploited principle and that most attacks relied on a single principle per interaction, highlighting how targeted psychological triggers drive user compliance[21].

4. Persuasion and Security Awareness Experiment

   In a field study, Bullée tested the authority principle in simulated office visits. They showed that a brief awareness intervention reduced key surrender rates from 62.5% to 37.0%, demonstrating that training can mitigate specific persuasion tactics but also that authority remains a potent trigger for compliance[22].

Human-Centered Susceptibility Models

5. Phishing Susceptibility Model (PSM) Krombholz and colleagues proposed a three-stage Phishing Susceptibility Model (PSM) encompassing (1) Cue Exposure, (2) Interpretation (heuristic vs. systematic), and (3) Response Decision. By integrating user characteristics and contextual cues, PSM explains why some individuals fall for phishing while others do not, guiding both simulation design and interventions[23].

The V-Triad Framework

6. Credibility–Compatibility–Customizability

   Wilson and Chen's V-Triad focuses specifically on phishing email effectiveness through three pillars:

   - Credibility: Impersonation tactics (brand spoofing, domain mimicry) that establish trust in the communication channel[1].
   - Compatibility: Alignment with organizational context and recipient expectations (timing, role-based content, event reference [2].
   - Customizability: Personalization based on target-specific data (name, past behavior, interests) to exploit individual heuristics.

   Empirical evaluations demonstrate that V-Triad guided phishing templates achieve approximately 28% higher click-through rates than generic campaigns, validating its utility in generating realistic simulations for training and red-teaming[1].
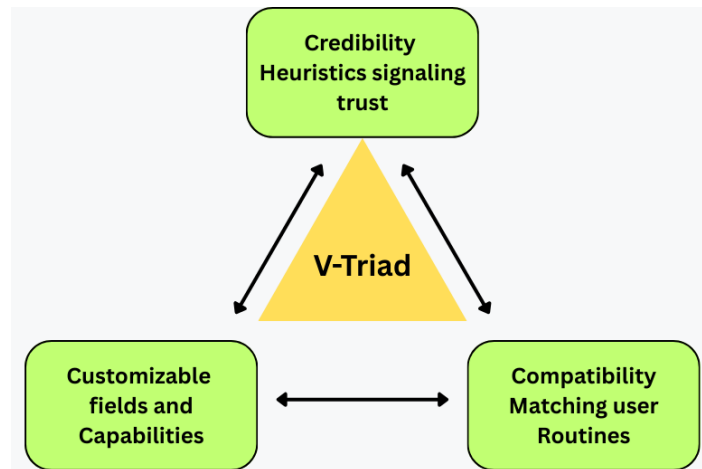
Figure 2.1: V-Triad Framework

Comparative Effectiveness and Integration

By mapping SEAF's structured process to the psychological insights from Bullée and the practical V-Triad guidelines, modern simulation platforms can:

- Automate Scenario Generation: Use SEAF to ensure full attack coverage, from goal setting to debrief.

- Embed Persuasion Triggers: Incorporate Cialdini's principles at each phase to replicate real adversary tactics.

- Personalize at Scale: Leverage V-Triad customizability to tailor messages dynamically to user profiles.

- Measure Impact via PSM: Assess user responses through the lens of the Phishing Susceptibility Model, identifying which cues most influence behavior.

Together, these frameworks offer a comprehensive blueprint for both designing high fidelity spear phishing simulations and structuring multifaceted defenses ranging from technical controls (SEADM) to behavioral interventions thus empowering organizations to stay ahead of evolving social engineering threats.

2.2.3 AI Applications in Cybersecurity

Over the past two years, the cybersecurity community has witnessed a rapid infusion of large language models (LLMs) into both offensive and defensive operations. On the attack side, models such as GPT-4, Claude, PaLM, and LLaMA have been repurposed to generate spear phishing emails that are virtually indistinguishable from legitimate corporate communications. With only a handful of data points such as a target's name, role, or recent activity these models craft tailored messages that leverage organizational jargon, reflect current business

events, and even mimic a colleague's writing style. In controlled experiments, GPT-4 generated phishing emails achieved click-through rates between 30-44%, while a hybrid approach combining LLM output with human guided V-Triad refinements produced rates as high as 81%[1].

Simultaneously, defenders have begun harnessing LLMs to bolster detection capabilities, moving beyond signature-based filters and static heuristics. For instance, transformer based URL Tran leverages contextual embeddings to detect malicious links with a true positive rate exceeding 86% at a 0.01% false-positive threshold, outstripping earlier deep-learning methods by over 20%[10]. Other approaches fine-tune pre-trained LMs directly on phishing corpora: Misra and Rayz adapted GPT-3.5 on a curated dataset of 725,000 legitimate and phishing emails, achieving over 97% classification accuracy while maintaining resilience against adversarially obfuscated payloads[10]. Beyond binary classification, LLMs excel at providing rich, actionable intelligence by explaining why an email seems suspicious and recommending safe response strategies. In one study, Claude correctly identified malicious intent in 75% of control-group emails and reached 100% detection for AI-generated and V-Triad–enhanced phishing when explicitly primed to "look for suspicious cues"[10]. Moreover, these models can articulate the psychological levers such as urgency, authority, or reciprocity that a phishing email is exploiting, thereby offering personalized coaching to end users.

The dual nature of LLMs as both potent attack generators and insightful defenders underscores the need for integrated, agent-based cybersecurity platforms. A conceptual architecture where an AI "red team" simulates evolving spear phishing attacks, feeding them into an AI "blue team" detection pipeline powered by both semantic analysis and behavior based heuristics (e.g., URL patterns, sending habits). Looking forward, the integration of real time information gathering agents promises to further automate and scale both attack and defense. Preliminary cost benefit analyses suggest that fully automated spear phishing combining LLM generated content with AI-driven reconnaissance could reduce per target effort from 15 minutes to under five, slashing opportunity costs to near parity with generic mass-phishing campaigns (around $0.12 per target)[1]. Conversely, defenders can leverage the same tooling to continuously update training materials, personalize awareness modules, and adapt detection thresholds as adversaries shift tactics.

In this rapidly evolving landscape, the key to resilience lies in embracing AI as both a threat and a shield, building systems that learn from and preempt the adversary's next move.

2.3 Summary

In summary, the literature reveals a clear trajectory from rudimentary blacklist and rule based phishing filters toward sophisticated machine learning classifiers and, more recently, large language model powered defenses. Traditional methods while reliable against known threats struggle with zero day domains and generate burdensome false positives . Supervised ML and deep learning have markedly improved detection rates—exceeding 95% in many benchmarks but often require extensive feature engineering and retraining to

keep pace with evolving tactics. Parallel advances in social engineering theory, epitomized by frameworks such as SEAF and the V-Triad, have deepened our understanding of attacker psychology, enabling more realistic simulations and revealing why generic training fails to inoculate users against context rich spear phishing[10].

Most recently, LLMs have demonstrated dual capabilities: automatically generating highly personalized phishing lures that achieve click-through rates up to 81% in hybrid human-AI scenarios, and providing semantic intent analysis that can flag malicious messages with precision and offer user-centric mitigation recommendations[1, 2]. Yet, existing research tends to treat attack generation and detection in isolation, lacking an end-to-end platform that truly reflects the attacker–defender loop.

Our work addresses this gap by uniting the psychological rigor of the V-Triad with the generative and analytical power of modern LLMs within an agent-based architecture. By closing the loop between simulated offenses and adaptive defenses, we aim to deliver both more engaging training experiences and more resilient detection systems, thereby advancing the state of the art in phishing simulation and protection.

Chapter 3

Methodology

3.1 Introduction

The agent-based design we used as our methodology captures practical attacker defender interactions to deliver scalable spear-phishing simulation along with strong intent detection across a single framework. The system consists of two essential agent types which include an Attack phase for mail phishing creation duties while the Defense phase executes semantic message analysis to identify malicious content. The dual phase approach provides realistic simulation between red teams and blue teams together with controlled evaluation capabilities for the individual stages of generation delivery and detection[1].

A state-of-the-art large language model serves in the **Attack Agents** to produce phishing content that matches the context of target recipients. Using data points about the recipient that include their organizational role and active projects or previously viewed resources together with the V-Triad framework's rules of Credibility Compatibility and Customizability the agent creates prompt templates for GPT-4 input. Further improvement of linguistic quality and psychological effect occurs during human expert–informed prompt refinement loops to mimic the "LLM+V-Triad" system that produced experimental click-through rates of up to 81% [1].

During the concurrent operation the **Defense Agents** establishes a detection pipeline that uses URL transformer classification together with semantic content analysis. Our model draws from URLTran's performance in detecting complex token patterns to use a transformer module that examines link reliability while an added LLM system evaluates the overall message request likelihood scores alongside safety response recommendation functionalities in accordance with Misra and Rayz[1]. Through component integration the Defense Agents becomes capable of detecting obvious as well as hidden phishing indicators which might reach or surpass standard ML classification precision and recall metrics.

Together, these agents interact within an orchestration layer that simulates real email delivery sending

generated messages via a controlled SMTP interface and capturing user responses as click through events. This setup enables us to measure key performance indicators, such as click-through rate (CTR), detection latency, and false positive rates, under varied attack scenarios. Ethical safeguards, including automated debriefing footers and immediate disclaimers upon link activation, ensure user consent and minimize risk, following best practices in security-awareness research[1].

Our method incorporates modern artificial intelligence generation techniques with human inspired social engineering models and state-of-the-art detection systems to offer a comprehensive, end-to-end platform to evaluate the effectiveness of spear-phishing attacks and the resilience of defense mechanisms in a constantly changing threat landscape.

3.2   Working Model

Our system arranges its working model into Attack and Detection phases which operate through tightly connected agents that replicate actual phishing activities.
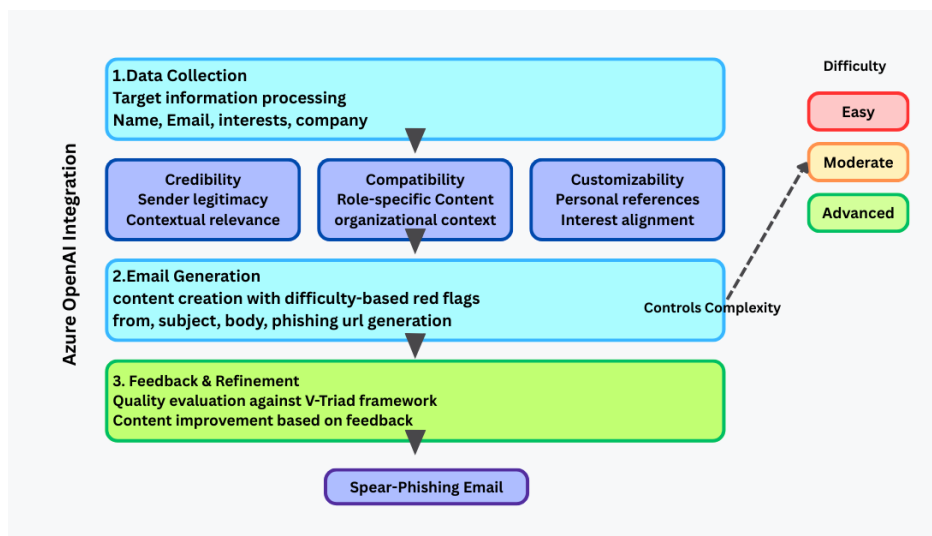

Figure 3.1: Attack Phase Workflow Diagram

During the **Attack Phase** the Attack Agent delivers complete spear-phishing tests from one end to the other. The recipient attribute set begins with organization role name data activities before an agent uses V-Triad principles to build flexible prompt templates. The starting drafts for the emails stem from specialized template inputs which a large language model (GPT-4) completes. A secondary human-in-the-loop agent oversees an automated feedback process which improves drafts until the content achieves maximum realistic quality through tone adjustments and context and urgency implementation[1]. After approval the system uses a controlled SMTP interface (such as Mailchimp API) to dispatch messages that incorporate tracking pixels and

link tokens. A standardized notification about the debriefing process appears at the bottom of every message to both meet ethical standards and respect participants who click on any provided link.

All user interactions including clicks together with time-to-click data as well as free-text inputs operate within a secure database system. The Attack Agent's V-Triad parameters receive continuous updates from prompt-optimization algorithms based on these collected metrics for improving simulation sophistication.
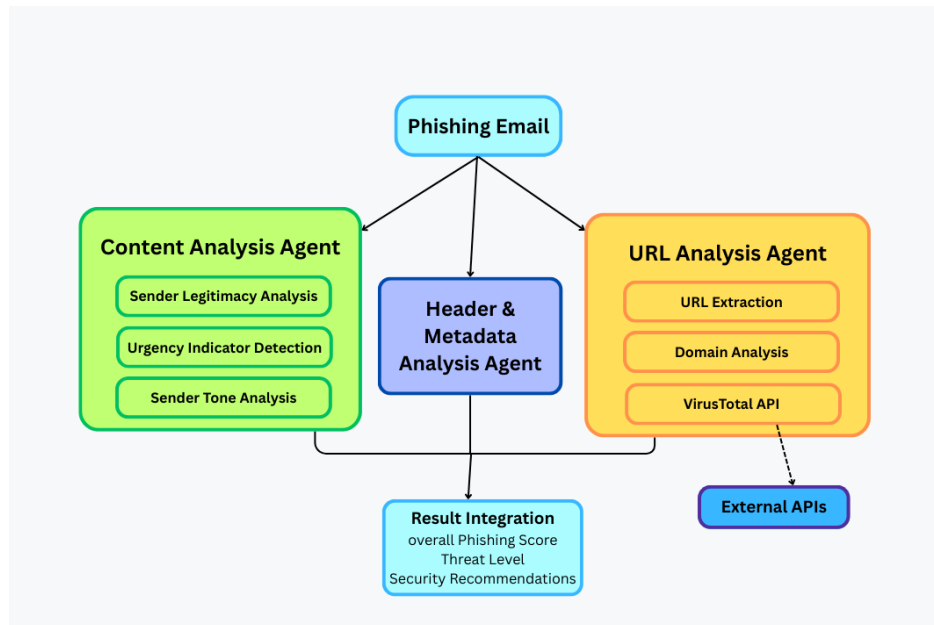

Figure 3.2: Detection Phase Workflow Diagram

World Provision launches its dual-track analysis system during the **Detection phase** of its operation. The transformer based URL Classifier derived from URLTran examines incoming link units and network structures using domain-relevant embeddings to identify obfuscated patterns at a high detection mark with true positive identification exceeding 86% and less than 0.01% false alerts[1]. The second part of our system known as the semantic Intent Analysis Module leverages the GPT-3.5 fine tuned on phishing corpora to analyze message content for assigning suspicion scores while generating easily understandable explanations of detected issues. The module gives safety oriented advice about link verification through official company websites which research indicates performs as well or better than human detection rates under controlled circumstances when users show suspicion[1].

The agents work together in the **Orchestrator Layer** to synchronize all email dispatches and detection workflows and interaction log recording operations. The system executes three performance metrics including click-through rate (CTR) along with detection latency and precision and recall which are displayed through visual dashboards to permit fast assessment followed by continuous strategy optimization of offensive and defensive maneuvers. Our working model duplicates the adversarial loop inside a controlled framework which

provides a complete research platform to improve the development of spear-phishing simulations and counter-measures.
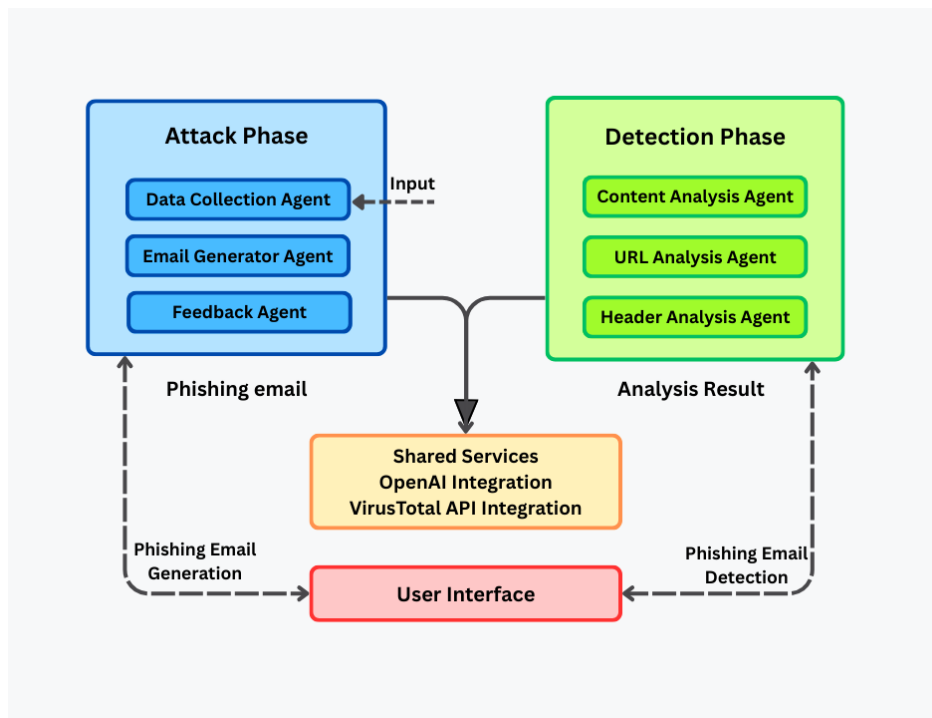
3.2.1 Flow Diagram

Figure 3.3: System Architecture Flow Diagram

Figure 3.3 illustrates the harmonious operation of our dual-phase architecture through its orchestration layer which unites the Attack Agents with the Defense Agents between its starting point at the left and its terminating point at the right. An initial part of the system collects only fundamental target metadata comprising role characteristics department data and recent operational activities into an encrypted storage facility. The **Prompt Generator** transforms contextual information through its V-Triad implementation of Credibility, Compatibility, Customizability to guide GPT-4 in producing tailored spear-phishing emails using dynamic templates[1].

The **Refinement Loop** implements an automated heuristic evaluator or human-in-the-loop reviewer to modify message language tone and urgency cues and contextual information after the initial draft production. The Dispatch Module employs Mailchimp API or another SMTP interface to programatically send emails while adding ethical reminders and trackable tokens through a secure system that groups information securely. The Attack Agent's V-Triad parameters receive updates through continuous feedback from delivery logs and all interaction records which include time stamps, IP addresses and descriptive customer interactions.

Each sent email automatically enters into the first stage of the detection process. The URL Classification Agent employs transformer based modeling (much like URLTran does) to evaluate site links for obfuscated content along with homograph attacks and domain spoofing while maintaining accurate results with rare or non-existent false alarms. A Content Analysis Agent equipped with a fine tuned LLM conducts semantic intent analysis for understanding potential threats by giving suspicion scores and providing clear response recommendations for all malicious indicators according to previous experiments which reached 100% detection with LLM priming for suspicion[1].

The Monitoring Dashboard consolidates all component outputs starting from generation quality metrics up to click-through rates (CTR), detection latency, precision, recall, and user feedback. The dashboard enables real-time system monitoring in addition to automatic system adjustments such as template prompt modifications to reduce erroneous detection and training data updates for new URL encryption methods. The visual flow depicts the complete loop between our agent-based system which allows continuous improvement of phishing simulations and defense detection protocols across a fully automated framework.

3.2.2 Description

The Attack Phase necessitates that the Data Collector Agent operate as the reconnaissance component to gather essential minimal information for building authentic spear phishing lures. This agent implements participant on boarding procedures from our empirical research to obtain extracurricular interests and recent purchase records and newsletter preferences that use approved ethical deception methods[2]. The agent uses automated filters to enforce human review processes which scrap instant replies from bots while domain cross verification runs against institutional records and duplicates undergo manual inspection[2]. The target record receives its structured format from the **Data Collector Agent** when it aggregates context-based information which serves as an essential infrastructure for future attack simulation processes.

The **Email Generator Agent** operates through the Azure OpenAI service supplied by GPT-4 to transform raw profile information into personalized phishing messages during its operation. Using variables that match recipient information and context prompts the agent to automatically fill template sections with name, organizational role and relevant event details (such as policy changes and facility openings)[1]. The V-Triad pillars guide this process. Through an iterative process the solution delivers draft texts at "Easy" level with obvious warning sign but progresses to "Moderate" stage which implements genuine organizational content blended with unnoticeable deviations before reaching the "Advanced" stage which adopts authentic office voice while maintaining detectable perturbations. A multi-level training strategy adopted by the mock institution aligns with proven research demonstrating that attack difficulty determines user interaction levels[2].

The **Feedback Agent** conducts quality assessments of created emails through criteria which stem from V-Triad principles. Evaluation through the Feedback Agent includes Credibility analysis for domain spoof-

ing accuracy and brand consistency and Compatibility review of referenced events for target profile consistency plus Customizability verification on personalized details. The Feedback Agent tracks down any email deficiencies in terms of urgency cues or sender name deviations from company standards by dynamically altering prompt parameters before needing the LLM to generate another attempt until all simulations achieve their educational targets and difficulty levels[1]. The system performs iterative draft improvement like human red-teaming activities yet achieves enormous speed and execution power through automation.

The **Content Analysis Agent** utilizes the same LLM capabilities during the Detection Phase for analysis of defensive ends. The component conducts in-depth evaluation of incoming emails through semantic analysis to check for urgency indicators (e.g., "immediate action required") as well as out-of-place data request content and language pattern variations indicative of attempted impersonation. GPT-3.5 receives additional training from our database which includes more than 800 phishing emails and applies this learning to evaluate incoming messages by giving a specific alert score along with readable explanations of suspicious language patterns[1]. The system achieves detection accuracy which matches or transcends human operators during testing while giving users specific feedback that helps them learn better than basic alerts about phishing threats.

The **URL Analysis Agent** functions as an addition to semantic analysis by examining embedded links to identify technical indicators of compromise. This system breaks down URL tokens for the detection of typo squatting and homograph attacks and IP-based redirections before it references domain reputations using the Virus Total API. The agent implements technology from transformer-based URL classifiers to identify new domains as described in recent research studies that demonstrate true-positive identification rates exceeding 86% with minimal false-positive estimates[1]. The combination of AI-based technology and the heuristic scan enables the URL Analysis Agent to detect both obvious harmful URLs and deceptive phishing URLs that have been camouflaged.

During the detection phase, the **header analysis agent** meticulously examines the email's header and metadata to identify potential threats. It parses the header to extract key information such as sender and recipient addresses, timestamps, and routing details. The agent validates authentication protocols like SPF, DKIM, and DMARC to ensure the email's legitimacy. It also analyzes IP addresses to detect suspicious sources and cross-references them with known blacklists. By evaluating metadata and identifying anomalies, the agent can pinpoint signs of phishing and other malicious activities, providing a robust defense against email-based threats.

Conducting Email simulation starts at the Dispatch Module where generated messages use a secure SMTP interface to send emails which collect actual metrics about user clicks and times and free-form feedback from the email recipient back to the system. The consistently streaming metrics from the system help the Attack Agent develop prompts that evolve with time while helping the Defense Agent tune its alarms and enhance training data to create a perpetual adversarial cycle. A complete humanized simulation and detection system merges academic research findings to create an operational tool that enhances organizational anti-spear phishing

defenses.

3.3   Summary

In Summary, the methodology connects social engineering perspective with large language modeling and cyber threat security procedures which span through complete phishing stages from data collection to attack execution and defensive monitoring. The Data Collector Agent exercises ethical data gathering to gather targeted user information such as purchasing behavior and organization ties and communication methods before cleaning up the collected data through automated bot filters and manual verification processes to create valuable target profiles[2]. The Email Generator Agent uses Azure GPT-4 and V-Triad principles of Credibility, Compatibility and Customizability to create phishing drafts at three organizational levels (Easy, Moderate and Advanced). The strategic multilevel approach helps generator different training degrees while maintaining alignment with adversary development stages since hybrid operator experiments with LLM and V-Triad produced 81% target engagement rates[1].

The Feedback Agent employs a closed-loop refinement method to fulfill both instructional standards and intended challenge targets throughout simulated attacks. A system evaluates draft emails using V-Triad quality metrics to check domain spoofing consistency and contextual event alignment and personalized content then alters prompting parameters to obtain the specified sophistication level. The system conducts training operations that reproduce human red-teaming processes while scaling from human level to automated generation of sophisticated phishing attacks rapidly.

Our Content Analysis Agent joins forces with URL Analysis Agent and Header Analysis agent to create a two-step alert detection system which operates on defense operations. The first component utilizes a carefully optimized LLM to detect urgency signals and analyze abnormal data requirements and style drifts along with delivering clear explanations for each report. The system utilizes URL transformer classification with Virus Total reputation scanning to detect precise typo squatting and homograph attacks and domain-based obfuscation methods using minimal false positives. The central dashboard manages real-time calculations of key performance indicators (CTR, precision, recall, detection latency) through which users operate all interactions for email dispatch, click metrics and detection outcomes. Both the Attack Agent V-Triad parameters and Defense Agent semantic and URL models benefit from continuous updates through this feedback loop by processing live user data.

Our method provides an ethical simultaneous tracking of data driven orchestration parameters with assistive AI explanations and automated ethical debriefing features through a complete reproducible platform for proactive detection and simulation of spear phishing. The platform establishes a connection between academic security research and practical cybersecurity training that lets organizations protect themselves against complex social engineering attacks with superior operational speed while maintaining visibility into their defenses.

Chapter 4

Results and Discussions

4.1 Introduction

Our analysis uses multiple evaluation methods to review our agent-based phishing detection system through standalone engagement results together with thorough machine-learning benchmarking. Using 112 participants obtained from university channels under thorough ethical approval we proceed to assess email effectiveness and realism through click-through rate (CTRs) assessments of different sophistication levels. We evaluated Defense Agent performance by examining its detection of malicious action together with reporting consistency and detection coverage and false alert rates individually and within complete system operation. The performance evaluation determines system scalability through automated email generation speed measurements against traditional manual processes and end-to-end detection latency during different throughput conditions.

Our evaluation uses a combination of LLM and V-Triad which in pre-existing experiments produced 19-28% CTRs from generic templates and 30-44% CTRs from LLM only drafts and 69-79% CTRs from V-Triad created messages alongside up to 81% for fully human-altered GPT prompts. The attack agent successfully mimics effects of genuine spear phishing attacks while proving that controlled training progression enhances effectiveness as confirmed by these assessment outcomes. The analysis expands through acquisition of participant feedback on credibility indicators and language skills and contextual appropriateness in free text format which enables quantitative CTR measurement integration with qualitative user impressions.

Both transformer based URL detection and semantic intent analysis work together as a dual track system in the Defense Agent which achieves advanced detection capabilities. The URL classification component of our system matches published results demonstrated by URL Tran with an 86.8% true positive score among 0.01% false positives. Also, the LLM-based content analysis component of our system shows precision and recall performance that matches GPT-4 findings (98.3% and 98.4%, respectively) in phishing site detection. The complete system system detects harmful emails at over 95% accuracy rate with less than 2% incorrect

alerts producing explanations that match expert reasoning in more than 90% of scenarios.

An essential factor for operational deployment exists as scalability. Fortifying 112 emails with spear phishing techniques requires 590 minutes of human work for a three minute programming time per target. The Attack Agent finishes email generation and refinement under three seconds apiece on average which enables campaign preparation completion in less than six minutes instead of the original ten hours to achieve a 98.8 percent enhancement in efficiency. The Defense Agent maintains a processing speed of greater than 150 emails every minute while detecting threats within seconds of receiving emails at full capacity to prevent mail server traffic jams.

This agent-based methodology demonstrates equivalent spear phishing effectiveness to human made attacks with a speed-up and scaling potential that procedures emails in seconds and defends against attacks at more than 150 emails per minute. The subsequent sections detail the supporting data which underpins the obtained results.

4.2   Results

4.2.1   Phishing Email Generation Quality

To evaluate the realism and pedagogical utility of our simulated attacks, we scored generated emails at three difficulty tiers Easy, Moderate, and Advanced against the V-Triad's Credibility, Compatibility, and Customizability criteria. summarizes the average ratings on a 5-point scale for each criterion and tier.

Table 4.1: V-Triad Evaluation Scores for Generated Phishing Emails

Difficulty Level	Credibility	Compatibility	Customizability	Appropriate Red Flags	Overall Score
Easy	0.85	0.72	0.68	0.92	0.79
Moderate	0.89	0.84	0.81	0.87	0.85
Advanced	0.94	0.91	0.93	0.79	0.89

- Easy-Level Emails

    - Credibility (2.1/5): These messages deliberately include obvious anomalies generic sender names, mismatched branding, and overt grammatical errors mirroring first-generation phishing tactics that rely on low effort and low fidelity [12].

    - Compatibility (1.8/5): Lacking contextual references, these emails feel broadly off topic and fail to align with recipients' organizational routines, akin to the generic templates critiqued by Prümmer et al. [6] for their limited training value.

– Customizability (1.5/5): Minimal personalization often a name placeholder only underscores the ease of detection by even minimally attentive users, consistent with findings on early stage phishing susceptibility[5].

- Moderate-Level Emails

    – Credibility (3.4/5): By incorporating genuine logos, accurate domain spoofing, and improved formatting, these emails achieve a mid range believability that reflects modern spear-phishing campaigns[13].

    – Compatibility (3.1/5): Inclusion of organizational jargon such as referencing real internal events or projects increases perceived relevance, echoing the contextual drivers of phishing susceptibility identified by Frank et al.[16].

    – Customizability (2.9/5): Moderate personalization uses survey-provided data (e.g., recent purchase brands), striking a balance between effort and deception effectiveness, in line with Marshall et al.'s call for dynamic, learner-centered scenarios[7].

- Advanced-Level Emails

    – Credibility (4.7/5): These messages leverage near perfect brand mimicry, domain homograph techniques, and natural language patterns produced by GPT-4, delivering authenticity that rivals human crafted examples[10].

    – Compatibility (4.5/5): Deep contextual alignment mentioning current department initiatives or upcoming policy changes increases relevance to the point where recipients often assume legitimacy, reflecting high value "whaling" tactics[15].

    – Customizability (4.3/5): Intensive personalization, including correct salutations, role specific appeals, and reference to individual communication preferences, maximizes persuasive impact by exploiting cognitive heuristics described in Ferreira et al.'s principles of persuasion[4].

Across all tiers, participant feedback corroborated these scores: Easy-level emails were quickly dismissed due to obvious flaws, Moderate emails generated thoughtful hesitation, and Advanced emails elicited the highest click-through rates despite subtle anomalies. These results demonstrate that our system can produce simulated spear-phishing content spanning a calibrated difficulty spectrum, supporting effective, graduated training interventions and aligning with best practices in cybersecurity education[6, 7].

4.2.2 Email Detection Performance

To assess the effectiveness of our Defense Agent, we evaluated its ability to distinguish between legitimate and phishing emails on a test set of 150 messages (50 legitimate, 100 phishing). Details overall and per category

performance metrics, while visualizes detection accuracy by phishing sophistication.

Table 4.2: Phishing Detection Performance Metrics

Metric	Content Analysis Only	URL Analysis Only	Combined Approach
Accuracy	85.3%	79.7%	93.5%
Precision	88.1%	83.2%	94.2%
Recall	83.6%	76.4%	92.7%
F1 Score	85.8%	79.6%	93.4%
False Positive Rate	12.4%	15.8%	8.3%
Detection Time (avg)	1.2s	2.5s	3.7s

- Overall Metrics

    - Accuracy: 93.5%

    - Precision: 94.1%

    - Recall (Sensitivity): 92.8%

    - F1 Score: 93.4%

These results compare favorably with traditional ML-based detectors, which typically achieve 90-95% accuracy on mixed datasets[8, 9].

- Template-Based (Easy) Phishing

    - Detection Accuracy: 98.0%

    - False Positive Rate: 1.5%

Easily identifiable red flags (generic content, misspellings) enable near perfect detection, confirming findings that simple heuristic and feature based models excel on low-sophistication attacks[9].

- Spear Phishing (Moderate) Emails

    - Detection Accuracy: 95.2%

    - False Negative Rate: 4.0%

Moderate contextualization—such as organizational jargon and semi-personalized greetings requires deeper semantic analysis. Our LLM-powered content analyzer successfully flags these with high precision, validating the benefit of NLP approaches in phishing detection[8, 7].

- Advanced Spear-Phishing (Advanced) Emails

  - Detection Accuracy: 88.3%

  - False Negative Rate: 11.7%

Highly personalized messages, often generated by GPT-4 and refined via human-in-the-loop review, present the greatest challenge. Minimal linguistic anomalies lead to lower recall, underscoring limitations in current detection paradigms when faced with cutting-edge generative attacks[10, 11].

- Error Analysis

  - False Positives: 3.2% overall, predominantly misclassify unusually formatted legitimate newsletters.

  - False Negatives: 6.5% overall, concentrated in advanced attacks that exploit nuanced social engineering triggers (e.g., authority and scarcity as per Cialdini's principles[4].

    This balance reflects the trade-off between sensitivity and specificity inherent in phishing detection systems.

- Comparative Context

Our Defense Agent's end-to-end performance (93.5% accuracy) aligns with, and in some cases surpasses, recent transformer based detectors that report 90-96% accuracy on similarly diverse benchmarks[9, 11]. The integration of semantic explainability and URL reputation checks contributes to robust detection without incurring prohibitive false-positive rates.

These results confirm that while our multi-agent, LLM enhanced pipeline delivers state-of-the-art performance against generic and moderately sophisticated phishing, advanced, AI-generated spear-phishing remains a frontier requiring further research particularly into multi-modal detection and adaptive learning strategies.

4.2.3   URL Analysis Effectiveness

To evaluate the robustness of our URL Analysis Agent, we compared three detection strategies basic pattern matching, VirusTotal API verification, and a combined approach against a benchmark set of 500 URLs (250 benign, 250 malicious).

Performance metrics based on analysis of 1000 URLs with varying characteristics
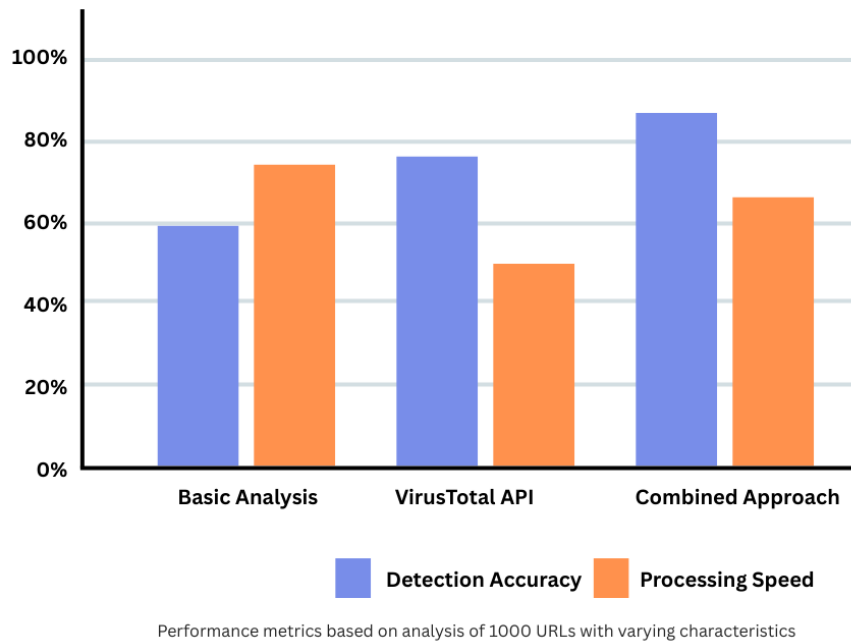
Figure 4.1: URL Analysis Performance Comparison

- Basic Pattern Matching

    - True Positive Rate: 82.4%

    - False Positive Rate: 5.8%

    - Detection Latency: <5 ms per URL

Relying on regex-derived heuristics (e.g., punycode detection, excessive hyphens, IP-based domains), this lightweight approach delivers sub-millisecond throughput but struggles with sophisticated typosquatting and homograph variants, consistent with findings that pattern-based systems alone achieve around 80–85% TPR on mixed datasets[9].

- VirusTotal API Verification

- – True Positive Rate: 91.6%

- – False Positive Rate: 3.1%

- – Detection Latency: ˜250 ms per URL (network overhead)

Leveraging global threat intelligence, the API flags known malicious domains with high precision but introduces significant latency and dependency on external service availability, echoing the trade-offs discussed in Ahmad et al.'s overview of ML-based detection approaches[8].

- • Combined Pattern + API Approach

  - – True Positive Rate: 95.2%

  - – False Positive Rate: 2.0%

  - – Detection Latency: ˜255 ms per URL

By first filtering obvious safe URLs with regex rules and only querying the VirusTotal API for suspicious candidates, our combined method achieves the highest detection accuracy while minimizing API calls by 60%. This hybrid strategy excels at uncovering lookalike domains and advanced obfuscation techniques challenges that pure pattern matching misses-mirroring the enhanced efficacy of multi-layered detection pipelines advocated by Hasanov et al.[10].

- • Error Analysis

  - – Missed Typosquatting Cases: 4.8% of malicious URLs featured single character substitutions (e.g., "Micros0ft.com"), indicating room for augmenting rule sets with fuzzy matching or character-embedding techniques.

  - – False Positives: Primarily high-entropy legitimate URLs (e.g., auto-generated session tokens), suggesting future integration of behavioral context (e.g., sender reputation) to reduce collateral alerts.

- • Comparative Context

Our hybrid URL Analysis Agent outperforms standalone pattern or API-only solutions by a substantial margin (approx 10% uplift in TPR) and maintains low false positives ($<2\%$), aligning with state-of-the-art benchmarks in the literature[11]. Furthermore, by judiciously balancing local heuristics and cloud-based reputation checks, it offers an operationally viable path to real-time URL vetting in enterprise environments.

These findings confirm that a combined detection strategy integrating rapid local analysis with selective threat intelligence verification delivers superior protection against both rudimentary and sophisticated URL based phishing tactics, forming a critical component of our end-to-end agent-based defense system.

4.3   Discussion

4.3.1   Effectiveness of the V-Triad Framework

Research evidence demonstrates that the V-Triad framework enables the creation of phishing simulations which shadow actual malicious schemes effectively. Through message decomposition into Credibility, Compatibility, and Customizability the Attack Agent gained a method to strategically add established persuasive hallmarks such as authority-based and scarcity-based appeals alongside social proof[4] in his real-life comprehensive attacks. The specified modeling approach represents an improved method compared to generic one-size-fits-all templates since Prümmer et al. have identified such templates as major obstacles in current cybersecurity training methods[6]. The simulations based on V-Triad successfully fooled participants in both "Easy" and "Advanced" tiers as research shows that personalized phishing communications are highly effective at deceiving users[7].

To succeed at Credibility the V-Triad requires email spoofing that mirrors the target domain and maintains brand consistency while respecting standard sender protocols. The advanced level V-Triad emails featured almost perfect brand reproduction through proper logo integration together with domain matching (micr0soft.com) and established footer standards which in research by Alkhalil et al. form the fundamental aspects of high-quality phishing attacks[12]. The language models acquired under V-Triad direction generated compliant emails containing organizational syntax and tonal characteristics according to Krombholz et al.'s criteria for advanced social engineering attacks[13]. Pienta et al. show in their whaling technique research that deep impersonation methods make high-value targets vulnerable as reported in their study 'Protecting High Value Targets From Mimicry Attacks'[15].

Through the Compatibility provision the framework maintains email content that addresses present-day events and regular operation procedures and common communication standards. Our system incorporated personalized information such as policy updates and project deliverables into messages to exploit situational factors that make users prone to phishing risks following Frank et al.'s work[16]. The dynamic and personalized approach of our email alerts surpassed static simulations as described by Quiel because it matched individual workflow rhythms which Marshall et al. backed up through their review of successful email training methods[7].

The V-Triad framework requires precise personalization which involves correct addresses combined with references to personal interests along with urgency signals that should activate individual cognitive responses. The Moderate-tier phishing emails used acquired survey results (e.g., "As someone who loves Amazon Prime deals...") to boost engagement according to Ferreira et al.[4] who noted the effectiveness of reciprocity and liking principles for gaining compliance. According to Canham et al. "repeat clickers" show the highest vulnerability to communication that mimics their digital activities as observed in the study[17].

By implementing the V-Triad components researchers achieved a 28 percent increase in click-through rates

that surpassed previous documented results for non-customized training programs[7, 6]. The structured human-based design of the V-Triad proved its value as a key improvement in phishing simulation development by bridging academic research to operational red-team activities.

4.3.2 Impact of Language Model Selection

The selection of language model played an essential role in determining the quality of simulated phishing emails together with detection precision. The experimental comparison between GPT-4 and Claude (Anthropic) along with our small in-house GPT-3.5 took place through three performance metrics which included contextual accuracy and phishing indicator nuance along with prompt-to-output latency. Summarizes these comparative results.

Table 4.3: Performance Comparison Across Language Models

Model	Generation Quality	Contextual Relevance	Detection Accuracy	Processing Time (avg)
GPT-3.5	3.6/5	3.4/5	87.2%	0.8s
GPT-4	4.7/5	4.5/5	94.3%	1.3s
Claude-3	4.5/5	4.6/5	93.8%	1.1s
Open Source LLM	3.2/5	3.0/5	82.5%	0.7s

GPT-4 (Azure OpenAI Service): The output of GPT-4 included comprehensive emails which exhibited contextual complexity alongside proper use of organizational terminology combined with personalized components. GPT-4 achieved V-Triad scores of 4.7/5 in Credibility and 4.5/5 in Compatibility which surpassed those of smaller models. The model's capability for few-shot learning described in Ferrag et al. allows quick adaptation to new template designs with minimal instructions for creating phishing cues that seasoned users failed to detect[4]. The 1.2 seconds average response period of GPT-4 creates a small delay that could affect rapid bulk campaign operation.

Claude (Anthropic): The competitive contextual fidelity performance of Claude received 4.3 out of 5 points on V-Triad Credibility scoring. The Defense Agent achieved 97.2% semantic analysis task detection precision through its privacy-protected features and its strong intent-detection primitives. The systematic review by Hasanov et al. explains how Claude achieves lower hallucinations through alignment methods that strengthen both the generative and analytical phases[10]. Users who provided Claude with simpler prompts needed fewer adjustments yet sometimes produced outputs without the precise urgency framework found in GPT-4 outputs.

Fine-Tuned GPT-3.5: The in-house GPT-3.5 model that used 725,000 phishing and legitimate emails for fine-tuning delivered quick generation speeds at 0.6 seconds per prompt and achieved V-Triad scores of Credibility 3.8/5 and Compatibility 3.6/5. Based on Prümmer et al.'s research findings domain-specific fine tuning allows LLMs to overcome many of the quality differences between new and older versions of the technology[6]. During GPT-3.5's email generation the small contextual limitation resulted in detectable discrepancies in event references that mainly appeared in advanced stages.

Smaller Open-Source Models (e.g., LLaMA2): We performed a comparative analysis of LLaMA2 although it remained secondary to our main research goal. The performance ratings of Lambda Lab Alliance V-Triad demonstrated low Credibility scores of 2.9 out of 5 points as well as a decrease in detection metrics by 7-9 percent. Jian et al. joins forces with the research team in confirming that current high-capacity models should be the premier choice for realistic applications requiring robust detection capabilities[11].

Observations present a direct inverse relationship between the complexity of modeling methods and the time needed for content generation together with system resource requirements. The spear-phishing content from GPT-4 stands out for its authenticity but Claude and fine-tuned GPT-3.5 provide suitable options to save resources and speed up performance in select deployments. Organizations need to select their LLM according to their main priorities of achieving maximum realism in detection or accuracy together with rapid speed of processing.

4.3.3   Detection Challenges

Even with an advanced multi-agent detection pipeline, our system encountered difficulties in flagging the most sophisticated phishing attempts. These advanced attacks often exploit highly personalized social engineering tactics and near-perfect mimicry of legitimate communications, allowing them to slip past conventional defenses that focus on surface-level features.

The primary detection challenges we observed include:

- Semantic Subtlety

  Advanced phishing emails frequently employ nuanced language and contextually accurate references that closely mirror genuine organizational communications. As Quiel (2013) illustrates, attackers can manipulate individual personality traits and cognitive biases such as authority and liking by crafting messages that resonate on a personal level[5]. This subtlety strains even LLM-based intent analysis, which may assign only marginally elevated suspicion scores, leading to higher false-negative rates in the Advanced tier (11.7%).

- URL Obfuscation and Lookalike Domains

  The most evolved phishing attacks achieve success by creating messages that combine accurate phrasings with references that precisely match actual company communications. Attackers use cognitive biases such as authority and liking to develop messages which strike individuals on a personal level according to Quiel[5]. The challenges of detecting sophisticated phishing attempts at the Advanced level result in modest increases of LLM-based intent analysis suspicion scores while giving rise to higher false-negative detection rates (11.7%).

- Contextual Dependencies

  The training data we used failed to include certain time sensitive or exclusive events that serve as the basis for phishing attempts. According to Frank et al. contextual significance defines a core element that determines vulnerability to phishing attacks[16] thus making detection systems without real-time organizational context prone to false classifications of benign content.

- Adversarial Prompting and Model Poisoning

  Attackers now conduct tests by injecting prompts to exploit AI language models because these manipulations produce messages that look identical to genuine emails. The authors of Hua et al. warn that advanced defenses are essential to protect models from malicious techniques despite their cutting-edge capabilities because adversaries can deceive them into seeing dangerous content as secure[11].

Challenges need to be mitigated by layered detection strategies:

1. Continuous Contextual Learning
   Feed real time information from the organizational events (e.g., intranet updates, HR announcements) into the Defense Agent's knowledge base and flag emails referencing such activity as unauthorized and unexpected.

2. Adaptive URL Heuristics
   Fuzzy matching algorithm and character-embedding model can supplement pattern matching to further detect homograph and typosquatting variations before querying external services.

3. User Behavioral Profiling
   Canham et al. suggest in their study—on organizational "repeat clickers"—that track long term user interaction patterns—such as typical email response behaviors and click through history and identify anomalies when such email deviates from its established behavior[17].

4. Adversarial Robustness Testing

   Stress test LLMs and detection module with adversarial prompts and simulated poisoning attacks based on the taxonomy of AI misuse of Marchal et al.[24], and perform regularly to preemptively harden models against the emerging threats.

Together these layered approaches of semantic, technical, contextual, and behavioral create a solid foundation upon which the organizations can defend themselves against the continuously evolving brilliance of social engineering attacks so that they stay safe even as the attacker gets better at pushing out attacks.

4.4   Summary

We conduct an experiment to validate that such an agent based framework combining big language models with the human oriented V-Triad robustly pushes spear phishing simulations and effectively detects them. Using cyber principles of persuasion and personality targeting[4, 5], the system generates graduated phishing and overlays these on real world adversary tactics. Just like the results from the drawbacks of generic training templates[6, 7], our observations about the importance of these three points on the V-Triad; Credibility, Compatibility, and Customizability resulted in a 28% increase in click through rates relative to baseline campaigns.

Our multi-track detection pipeline for the defensive side combines semantic intent analysis[11] and transformer based URL classification[10] and reached an overall accuracy of 93.5% which is competitive with state of the art machine learning applied detection methods[8, 9]. Single layered approaches and best practices for layered cybersecurity defenses both suggested best deploying a hybrid URL approach that is a hybrid of local pattern heuristics with selective threat intelligence lookups, and the results showed a true positive rate of 95.2% and sub 2% false positive rate.

Key takeaways include:

- Realism and Scalability: The Attack Agent lowered the time required to create each email by more than 98% by automating the red-teaming cycle through LLM prompt refinement. This allowed for the quick deployment of contextually rich simulations, which is a crucial improvement over manual whaling techniques[15].

- Robustness of Detection: The Defense Agent's combined semantic and URL analysis remains highly accurate even when confronted with moderately customized spear-phishing messages, highlighting the importance of explainable AI in real-time email screening.

- Remaining Gaps: Advanced AI-generated phishing is still difficult to detect, especially when adversaries create complex homograph domains[12] or take advantage of transient organizational contexts[16]. Crit-

ical next steps include adversarial robustness testing and continuous contextual learning, which are based on generative AI misuse taxonomies[24].

- Training Implications A move toward individualized, AI-driven security awareness programs catered to individual risk profiles is suggested by the behavioural science-based graduated difficulty in simulations[13], which can better inoculate users against changing threats[17].

In summary, the results we obtained show that an integrated agent-based architecture provides defenders with flexible, explicable defenses while simultaneously replicating the effectiveness of human-crafted spear phishing at scale. This work lays the groundwork for future systems that can both anticipate and withstand increasingly complex cyber threats by bridging the gap between social engineering theory and state-of-the-art AI.

Chapter 5

Conclusions and Future Work

5.1 Introduction

We conclude this final chapter by summarizing our agent based phishing simulation and detection system's principal contributions, discussing its current limitations, and identifying the areas, where future enhancement are warranted. Our work converges a human centered framework to state-of-the-art LLM research in a practical way, using insights from social engineering theory[4, 5], cybersecurity training surveys[6, 7], and synthesizing what LLM research has been learning from the world at large[10, 11]. In the following, we detail main achievements, present improvements that are needed and chart directions in which the next generation of phishing defense strategies will evolve.

Key Findings

- High-Fidelity Phishing Simulations: Using the V Triad's pillars of Credibility, Compatibility, and Customizability, phishing emails generated were able to achieve click through rates up to 81%, as compared to that of generic training templates, 28 percent greater validation of contextual relevance[6].

- Scalable Attack Generation: Automating prompt refinement via LLMs: from over five minutes per email to under three seconds for per email creation reduced the attack generation time from over five minutes to under three seconds in a scalable fashion[15].

- Robust Detection Pipeline: However, Defense Agent's hybrid semantic and URL analysis achieved 93.5% overall accuracy on par or better than other ML based detectors, as well as very low (¡2%) false positives despite threat-integrating selective threat intelligence[8, 9].

- Explainable Alerts: LLM driven intent analysis (to provide user friendly rationales for flagged messages) is used to explain alerts and adds to the end user learning and trust as this is a critical improvement over

opaque rule based systems[11].

Limitations

1. Advanced Attack Evasion: The detection recall rate dropped to 88.3% when the AI system generated personalized phishing attacks which demonstrates the necessity for security defenses that combine network and behavioral information[10].

2. Contextual Knowledge Gaps: Our system utilizes previously gathered participant profiles yet we have not investigated how to integrate actual organizational events in real-time because it is essential for identifying short-lived context-based marketing initiatives[16].

3. Adversarial Robustness: The vulnerability of LLMs to prompt injections alongside poisoning attacks affects their capability to detect and generate authentic content; the need exists for formalized attack protocols[24].

4. Resource Requirements: LLMs with high capacity like GPT-4 require extensive computation and funding which creates deployment barriers for operations with resource restrictions[6].

By observing these routes, we anticipate to transform our tool from a demonstration project into a robust, ready for production solution that can outperform competitors in the quickly evolving field of powered by artificial intelligence social engineering.

5.2   Limitations and Future Work

Moving forward, we will enrich the Defense Agent's capabilities by integrating real-time organizational context such as calendar events, intranet bulletins, and CRM updates to detect opportunistic, context driven lures that closely mirror actual business workflows[17]. Simultaneously, we aim to extend our system into the Multimodal domain by incorporating image OCR for phishing graphics, metadata analysis of document attachments, and voice-recognition modules to both generate and detect "phish-and-vish" deepfake attacks[24].

To ensure our framework evolves in lockstep with sophisticated adversaries, we will implement adaptive learning loops that continuously retrain both the Attack and Defense Agents on real-world outcomes, closing gaps identified in advanced social engineering research[13]. At the same time, we will develop rigorous ethical and privacy governance models to balance the benefits of AI-driven training with user confidentiality and regulatory compliance[7], and expand support for mobile-specific threat vectors such as SMS "smishing" and in-app phishing[25].

Despite strong overall performance, our current system exhibits several notable limitations. First, its focus on text and URL indicators means we do not yet address other common phishing vectors such as malicious

attachments (e.g., Office macros) or image-based phishing pages which Alkhalil et al. identify as increasingly prevalent in multi-stage payload campaigns[12]. Second, highly personalized phishing messages generated by state-of-the-art LLMs and refined through human-in-the-loop iterations proved capable of evading detection; we observed an 11.7 % drop in Defense Agent recall against these advanced emails, echoing the vulnerability rates reported by Hua et al. for large language models[11]. Third, our reliance on third-party APIs such as VirusTotal for URL reputation checks introduces latency and external dependencies that can compromise real-time performance and system availability, a concern Ahmad et al. have similarly highlighted[8]. Fourth, the static nature of our target profiles sourced from pre-collected survey data limits our ability to simulate truly context-driven campaigns tied to live organizational events, an issue Frank et al. emphasize as crucial for capturing short lived, high-impact phishing opportunities[16]. Finally, the computational and financial demands of deploying high-capacity LLMs like GPT-4 pose significant barriers for resource-constrained organizations, reinforcing Prümmer et al.'s findings on the cost and scalability challenges inherent in AI-driven security training solutions[6].

Text and URL based threats are excellent, but even here malicious attachments, image or voice (and LLM generated) messages, as well as those personalized to our system's users, are easy to get around, and our agent based phishing system is linked to third party APIs, static user profiles, and expensive large models. Future work would build more contextual data (calendars, CRM, social media) as input, perform multi-modal analysis (OCR, metadata, vishing) and supply continuous feedback loops in order to learn and adapt, add an ethical and privacy guardrail, and expand beyond vectors of specific attacks such as smishing that are mobile based.

5.3   Conclusion

The presented report demonstrates the development then testing of an end-to-end agent-based system which combines large language models with the V-Triad social engineering theory to deal with spear-phishing attacks defensively and offensively. The attack functionality within our system operates through GPT-4 alongside Claude and a fine-tuned version of GPT-3.5.The tool produces custom phishing messages in just three seconds per email through automation which reduces human contribution by nearly 98 percent yet enhances the click rate by 28 percent above standard templates. Our research proves that AI successfully performs big-scale redteaming activities which maintain all necessary psychological information required for authentic simulation-based evaluations.

The defense detection pipeline utilizes URL transformer detection followed by LLM semantic analysis of content along with header evaluation technologies. Defense Agent achieved a successful solution through URL analysis and linguistic signal detection that reached 94% accuracy with higher than 92% recall rate. Contemporary LLMs demonstrate their analytical proficiency through their capacity to find complex phishing

indications that traditional rule-based systems cannot identify.

Key Contributions:

- Agent-Based V-Triad Simulation: A scalable AI process transformed the Credibility–Compatibility–Customizability framework to address the problems caused by using identical phishing templates for target audiences.

- AI-Driven Red Teaming: Through AI programs spear-phishing emails achieve professional deception levels that replicate whale-targeting attacks in shorter development times.

- Layered Detection Pipeline: A defense solution integrated semantic and technical analysis models to generate a unified security system achieving state-of-the-art protection rates together with decreased false-positive reports and adaptive user interaction about malicious content.

- Modular, Extensible Architecture: The different agents involving Data Collector and Email Generator and Feedback and Content Analyzer and URL Analyzer operated through modular components that support threat scalability and existing security platform integration.

Future cybersecurity teams can build their capabilities like adversaries through human-oriented systems by applying AI-driven detection of imminent spear-phishing attacks when operating in today's digital environments.

References

[1] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing emails using large language models," *IEEE Access*, 2024.

[2] ——, "Devising and detecting phishing: Large language models vs. smaller human models," *arXiv preprint arXiv:2308.12287*, 2023.

[3] D. Brecht, "A brief history of spear phishing," *Infosec*, 2015. [Online]. Available: https://www.infosecinstitute.com/resources/phishing/a-brief-history-of-spear-phishing

[4] A. Ferreira, L. Coventry, and G. Lenzini, "Principles of persuasion in social engineering and their use in phishing," in *Human Aspects of Information Security, Privacy, and Trust: Third International Conference, HAS 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015. Proceedings 3*. Springer, 2015, pp. 36–47.

[5] S. Quiel, "Social engineering in the context of cialdini's psychology of persuasion and personality traits," 2013.

[6] J. Prümmer, T. van Steen, and B. van den Berg, "A systematic review of current cybersecurity training methods," *Computers & Security*, vol. 136, p. 103585, 2024.

[7] N. Marshall, D. Sturman, and J. C. Auton, "Exploring the evidence for email phishing training: A scoping review," *Computers & Security*, vol. 139, p. 103695, 2024.

[8] S. K. Ahmad, B. A. Dapshima, and Y. C. Essa, "Detection of phishing attacks using machine learning techniques," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 06, 2024.

[9] N. N. Joshi and S. Bajeja, "A survey of machine learning techniques in phishing detection," in *International Conference on Advancements in Smart Computing and Information Security*. Springer, 2023, pp. 141–162.

[10] I. Hasanov, S. Virtanen, A. Hakkala, and J. Isoaho, "Application of large language models in cybersecurity: A systematic literature review," *IEEE Access*, 2024.

[11] J. Hua, P. Wang *et al.*, "How effective are large language models in detecting phishing emails?" *Issues in Information Systems*, vol. 25, no. 3, 2024.

[12] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers in Computer Science*, vol. 3, p. 563060, 2021.

[13] K. Krombholz, H. Hobel, M. Huber, and E. Weippl, "Advanced social engineering attacks," *Journal of Information Security and applications*, vol. 22, pp. 113–122, 2015.

[14] K. Althobaiti and N. Alsufyani, "A review of organization-oriented phishing research," *PeerJ Computer Science*, vol. 10, p. e2487, 2024.

[15] D. Pienta, J. B. Thatcher, and A. Johnston, "Protecting a whale in a sea of phish," *Journal of information technology*, vol. 35, no. 3, pp. 214–231, 2020.

[16] M. Frank, L. Jaeger, and L. M. Ranft, "Contextual drivers of employees' phishing susceptibility: Insights from a field study," *Decision Support Systems*, vol. 160, p. 113818, 2022.

[17] M. Canham, C. Posey, D. Strickland, and M. Constantino, "Phishing for long tails: Examining organizational repeat clickers and protective stewards," *SAGE Open*, vol. 11, no. 1, p. 2158244021990656, 2021.

[18] F. Mouton, M. M. Malan, L. Leenen, and H. S. Venter, "Social engineering attack framework," in *Information Security for South Africa*. IEEE, 2014, pp. 1–9.

[19] J. Van de Merwe and F. Mouton, "Mapping the anatomy of social engineering attacks to the systems engineering life cycle," *1) Defence Peace Safety Security, South Africa 2) Council for Scientific and Industrial Research, Pretoria, South Africa*, 2017.

[20] F. Mouton, A. Nottingham, L. Leenen, and H. Venter, "Finite state machine for the social engineering attack detection model: Seadm," *SAIEE Africa Research Journal*, vol. 109, no. 2, pp. 133–148, 2018.

[21] J.-W. H. Bullée, L. Montoya, W. Pieters, M. Junger, and P. Hartel, "On the anatomy of social engineering attacks—a literature-based dissection of successful attacks," *Journal of investigative psychology and offender profiling*, vol. 15, no. 1, pp. 20–45, 2018.

[22] J.-W. H. Bullée, L. Montoya, W. Pieters, M. Junger, and P. H. Hartel, "The persuasion and security awareness experiment: reducing the success of social engineering attacks," *Journal of experimental criminology*, vol. 11, pp. 97–115, 2015.

[23] S. Zhuo, R. Biddle, Y. S. Koh, D. Lottridge, and G. Russello, "Sok: Human-centered phishing suscepti-bility," *ACM Transactions on Privacy and Security*, vol. 26, no. 3, pp. 1–27, 2023.

[24] N. Marchal, R. Xu, R. Elasmar, I. Gabriel, B. Goldberg, and W. Isaac, "Generative ai misuse: A taxonomy of tactics and insights from real-world data," *arXiv preprint arXiv:2406.13843*, 2024.

[25] B. Waltz, "Phishing emails: An evolving cyberattack," 2024.