The American Journal of Engineering and Technology
ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 111-116

Check for updates

Cognitive Vulnerabilities in the Age of LLMs: Mitigating Generative AI-Driven Social Engineering Through Context-Aware Threat Detection

Dr. Elias Thorne
Department of Computer Science and Information Systems, Institute of Advanced Cybernetics

OPEN ACCESS

SUBMITTED 07 November 2025
ACCEPTED 15 November 2025
PUBLISHED 26 November 2025
VOLUME Vol.07 Issue 11 2025

CITATION

Dr. Elias Thorne. (2025). Cognitive Vulnerabilities in the Age of LLMs: Mitigating Generative AI-Driven Social Engineering Through Context-Aware Threat Detection. The American Journal of Engineering and Technology, 7(11), 111–116. Retrieved from https://theamericanjournals.com/index.php/tajet/article/view/6945

Abstract: The advent of Large Language Models (LLMs) has fundamentally altered the cybersecurity landscape, specifically within the domain of social engineering. While LLMs facilitate productivity, they also empower threat actors to generate hyper-personalized, grammatically perfect, and contextually relevant phishing campaigns at scale. This paper explores the intersection of generative AI, cognitive psychology, and intrusion detection to propose a novel defense framework. We investigate the efficacy of current AI-driven social engineering tactics, utilizing the Five-Factor Model of personality to map cognitive vulnerabilities exploited by generative agents. Furthermore, we introduce a Context-Aware Defense System (CADS) that leverages fine-tuned LLMs to detect semantic anomalies and psychological manipulation triggers in real-time communications. Our methodology involves simulating high-fidelity spear-phishing attacks against generative agent personas representing diverse psychological profiles. Results indicate that traditional signature-based detection fails against LLM-generated content, whereas the proposed semantic analysis approach improves detection rates significantly. We find that high Agreeableness and Neuroticism correlate with higher susceptibility to AI-generated pretexts. The study concludes that effective defense against the next generation of social engineering requires a paradigm shift from static filtering to dynamic, psychological, and semantic content analysis.

Keywords:

Large Language Models, Social Engineering, Generative AI, Intrusion Detection, Cybersecurity, Cognitive Vulnerabilities, Natural Language Processing.

Introduction

The rapid proliferation of Generative Artificial Intelligence (GenAI), specifically Large Language Models (LLMs) such as GPT-4, has precipitated a paradigm shift in both information synthesis and cybersecurity threat landscapes. Historically, social engineering—the psychological manipulation of people into performing actions or divulging confidential information—relied heavily on the manual craft of the attacker. Traditional phishing campaigns were often identifiable by grammatical errors, generic greetings, and a lack of contextual awareness. However, recent advancements in natural language processing have lowered the barrier to entry for sophisticated attacks, allowing threat actors to automate the generation of highly convincing, personalized narratives at scale [1].

The emergence of tools dubbed "ThreatGPT" or malicious derivatives of foundation models suggests that the cybersecurity community is entering an era of "AI versus AI." In this landscape, attackers utilize generative models to parse Open Source Intelligence (OSINT) and craft spear-phishing emails that exploit specific cognitive biases of the target [2]. Conversely, defenders must rely on equally sophisticated models to detect these threats, as traditional signature-based Intrusion Detection Systems (IDS) are increasingly rendered obsolete by the polymorphic nature of AI-generated text.

This research addresses a critical gap in current cybersecurity literature: the intersection of cognitive psychology and automated threat detection. While technical vulnerabilities are frequently patched, the "human firewall" remains susceptible to psychological exploitation. Previous research has established that personality traits significantly influence susceptibility to phishing [3], but few studies have quantified how LLMs exploit these specific traits or how defensive systems can be engineered to recognize these psychological manipulation attempts.

The primary objective of this study is to propose and evaluate a Context-Aware Defense System (CADS). This system moves beyond metadata analysis to examine the semantic and psychological structure of incoming communications. By integrating the Five-Factor Model (FFM) of personality into our threat modeling, we aim to demonstrate that understanding the who (the target's psychological profile) is as critical as understanding the what (the malware or payload) in defending against modern social engineering.

Theoretical Framework and Literature Review

The Evolution of AI-Driven Threats

The capability of LLMs to pass the Turing Test serves as a benchmark for their utility in deception. Recent studies indicate that human participants are increasingly unable to distinguish between human-written and AI-generated text [4]. This indistinguishability is the cornerstone of modern social engineering. Attackers no longer need to be fluent in the target's language or familiar with their organizational culture; the LLM acts as a cultural and linguistic bridge.

The concept of "personalized persuasion at scale" has been highlighted as a significant risk [5]. Generative AI allows for the micro-targeting of individuals based on their digital footprints. By analyzing public social media data, an LLM can infer a target's interests, communication style, and recent activities, crafting a message that bypasses initial skepticism. This contrasts sharply with earlier data mining concepts [6], which focused on structured data patterns rather than semantic manipulation.

Intrusion Detection and Feature Engineering

Historically, feature selection for intrusion detection relied on packet headers, traffic flow, and protocol anomalies [7]. The KDD 99 dataset and its successors focused on network-layer attacks [8]. However, social engineering operates at the cognitive layer, which is undetectable by traditional packet inspection.

Later frameworks attempted to construct features for IDS based on user behavior profiles [9]. While effective for insider threats, these models often fail to detect external social engineering attacks that hijack legitimate communication channels without triggering volume-based alarms. The integration of Deep Reinforcement Learning (DRL) in cybersecurity has shown promise in adapting to dynamic threat environments [10], yet the application of DRL specifically to natural language threat detection remains an emerging field.

Cognitive Vulnerabilities and Personality

The "human factor" in security is often analyzed through the lens of the Five-Factor Model (FFM): Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [11]. Research by Parrish et al. laid the groundwork for understanding how these traits correlate with phishing susceptibility [3]. For instance,

individuals with high Agreeableness may be more compliant with requests from perceived authority figures, while those with high Neuroticism may react impulsively to fear-based appeals.

Generative agents—simulations of human behavior using LLMs—have demonstrated the ability to replicate these personality traits with high fidelity [12]. This capability allows researchers to simulate social engineering attacks against "synthetic populations," providing a safe and ethical testing ground for defense mechanisms without exposing real users to risk.

Methodology

System Architecture: Context-Aware Defense System (CADS)

The proposed CADS architecture operates as a middleware layer between the external communication gateway (e.g., email server, chat client) and the end-user. It consists of three primary modules:

1. The Semantic Analyzer: An LLM-based component fine-tuned to detect persuasive language patterns, urgency triggers, and requests for sensitive actions.

2. The Context Engine: A module that compares incoming message content against known organizational context (e.g., verifying if a request for a wire transfer aligns with standard vendor payment schedules).

3. The Personality Risk Mapper: A theoretical module that adjusts alert thresholds based on the user's role and estimated psychological susceptibility.

Simulation Environment

To evaluate the system, we employed a Generative Agent Simulation [12]. We created a virtual organization consisting of 1,000 distinct generative agents, each assigned a specific personality profile based on the FFM.

- Attacker Model: An instance of GPT-4 configured to act as a sophisticated social engineer. It was provided with varying levels of OSINT data regarding the target agents.
- Defender Model: The CADS implementation, utilizing a fine-tuned BERT model for feature extraction and a logical regression layer for threat classification.

Data Generation and Feature Extraction

The Attacker Model generated 5,000 unique phishing emails targeting the synthetic population. These emails ranged from "low sophistication" (generic mass-market phishing) to "high sophistication" (context-rich, multi-turn spear-phishing).

We extracted features not typically used in standard IDS. Instead of relying solely on header information, we focused on:

- Semantic Urgency Score: Quantification of time-pressure language.
- Authority Alignment: Measurement of how the sender attempts to establish hierarchy.
- Mismatch Index: The semantic distance between the sender's purported identity and their linguistic style.

Ethical Considerations in Methodology

While the use of generative agents mitigates direct harm to human subjects, the development of autonomous attack models poses a "dual-use" risk. To adhere to responsible AI research guidelines, the "Attacker Model" prompts were designed to be non-exportable and were terminated immediately post-simulation. No real-world PII (Personally Identifiable Information) was used; all "victims" were synthetic constructs.

Results

The simulation revealed a strong correlation between specific personality traits and the success rate of the AI-generated attacks. Agents programmed with high Agreeableness were 34% more likely to fall for authority-based pretexts (e.g., CEO fraud). Agents with high Openness were susceptible to curiosity-based lures but resilient against fear-based tactics. Conversely, agents with high Conscientiousness were the most resilient overall, often verifying policy before compliance, unless the attack specifically mimicked bureaucratic procedure.

These findings suggest that "one-size-fits-all" security awareness training is insufficient. A defense system must account for the specific cognitive blind spots of the user.

Performance of CADS vs. Traditional Filters

We compared the CADS performance against a standard keyword-based Bayesian filter (a common component in legacy email security gateways).

- Keyword Filter: Achieved a detection rate of 42% against high-sophistication AI attacks. It struggled

significantly with "clean" attacks—emails that contained no malicious links or attachments but sought to initiate a conversation (pretexting).

- CADS (Proposed): Achieved a detection rate of 89% against the same dataset. The semantic analysis capabilities allowed CADS to identify the intent of the message rather than just looking for known malicious signatures.

False Positives and Operational Latency

A critical metric for any IDS is the False Positive Rate (FPR). The initial CADS configuration showed a higher FPR (12%) compared to the keyword filter (2%). Legitimate emails containing urgent business requests were occasionally flagged as social engineering. Through reinforcement learning feedback, where the system was penalized for flagging internal communications from verified high-trust nodes, the FPR was reduced to 4.5% without significantly compromising detection accuracy.

Latency analysis showed that the additional semantic processing added an average of 400ms to email delivery time. In a high-frequency trading environment, this might be unacceptable, but for standard corporate communication, this delay is negligible relative to the security benefit.

Extended Analysis

The interaction between the Attacker Model and the Defender Model in our simulation highlights a burgeoning "cognitive arms race." As we peel back the layers of this interaction, it becomes evident that the efficacy of AI-driven social engineering lies not just in the generation of text, but in the dynamic adaptation to the target's mental state—a capability we term "Empathetic Malignance."

The Mechanism of Empathetic Malignance

In traditional social engineering, the attacker guesses the target's emotional state. In AI-driven attacks, the LLM analyzes the target's responses (in multi-turn interactions) to adjust its tone. Our simulation data showed that when a target agent responded with hesitation, the Attacker Model shifted strategies from "intimidation" to "reassurance." This dynamic pivot is where traditional security controls fail completely. A static filter sees the text; it does not see the trajectory of the conversation.

The CADS framework addresses this by maintaining a "Conversation State Tensor." This component tracks the emotional gradient of the interaction. If a conversation moves rapidly from neutral to high-anxiety or high-compliance without an external contextual justification, the system flags the interaction. This suggests that future defense mechanisms must be stateful, preserving the context of a dialogue over time rather than analyzing messages in isolation.

Semantic Camouflage and Feature Obfuscation

A significant finding in the "High Sophistication" dataset was the emergence of "Semantic Camouflage." The Attacker Model, upon realizing that certain keywords (e.g., "urgent," "transfer," "password") were triggering defenses, began to use circumlocutory language. Instead of "send the password," the model would generate, "kindly facilitate the access credentials at your earliest convenience."

This obfuscation defeats keyword filters but also challenges simple NLP models. To counter this, we implemented a layer of abstraction in the CADS utilizing "Intent embeddings." By mapping sentences to a vector space representing intent rather than syntax, the system could recognize that "facilitate access credentials" and "send password" occupy the same semantic neighborhood. This finding reinforces the necessity of using transformer-based models (like BERT or RoBERTa) in defense pipelines, as they are capable of understanding deep contextual relationships that defy simple obfuscation.

The Role of Synthetic Data in Defensive Training

One of the profound limitations in cybersecurity research is the lack of labeled datasets for zero-day social engineering attacks. By the time a dataset is compiled, the attack vector has often evolved. Our use of generative agents to create the training data offers a solution to this "Cold Start" problem.

We observed that training the CADS on purely human-generated phishing emails resulted in a model that was unprepared for the subtle intricacies of machine-generated text. Machine-generated text often exhibits lower perplexity and higher structural consistency than human text. By including AI-generated attacks in the training set, the defense model learned to identify the "fingerprint" of an LLM—the subtle statistical regularities that distinguish GPT-generated text from human writing. This implies that robust defense systems must be trained on hybrid datasets containing both human and machine-generated adversarial examples.

Regulatory and Privacy Implications of Cognitive Defense

The implementation of a system like CADS raises significant ethical questions regarding employee privacy. To effectively detect psychological manipulation, the system must analyze the tone, sentiment, and content of employee communications. This borders on "surveillance AI."

In our discussion of the architecture, we propose a "Privacy-Preserving Processing" method. The text data is tokenized and converted into vector embeddings locally. Only the numerical vectors—which cannot be easily reverse-engineered back into the original text without specific keys—are passed to the central analysis engine. This ensures that while the intent and risk are analyzed, the raw content of private conversations remains encrypted or obfuscated. Balancing the need for deep semantic analysis with the requirements of regulations like GDPR and CCPA is the next major hurdle for deployment.

Authentication Integration

While this study focuses on textual analysis, it is clear that text analysis alone is a delaying action. The ultimate mitigation for social engineering is the removal of the reliance on human judgment for authentication. The integration of CADS with biometric systems creates a "Zero Trust" environment.

In our proposed theoretical extension, a "High Risk" flag from the CADS would not simply block the email (which can disrupt business) but would instead trigger a "Step-Up Authentication" protocol. If an employee attempts to act on a flagged email, the system would require a secondary biometric verification (e.g., fingerprint or facial scan) as detailed in biometric acceptance studies [10]. This creates a safety net: even if the human mind is hacked by the social engineering narrative, the digital system intercedes to verify the identity and authorization of the user before the payload is executed.

Limitations and Future Directions

While the results of this study are promising, several limitations must be acknowledged to contextualize the findings.

First, the reliance on "Generative Agents" to simulate human victims is a proxy measure. While recent literature supports the validity of LLMs in simulating human behavior [12], synthetic agents cannot fully replicate the erratic, irrational, or fatigued states of real human workers. Real-world users are influenced by external factors—deadlines, personal stress, office politics—that are difficult to parameterize fully in a simulation. Therefore, the susceptibility rates observed in our study should be viewed as a baseline rather than an absolute predictive metric for human organizations.

Second, the "Attacker Model" used in this study was a static instance of GPT-4. In a real-world scenario, threat actors utilize "jailbroken" or fine-tuned models specifically stripped of safety filters. These adversarial models may employ more aggressive or unethical tactics (e.g., extortion, explicit threats) that our research model was safety-constrained from generating. Future research should explore the defensive requirements against uncensored open-source models (e.g., LLaMA derivatives) that may be weaponized by bad actors.

Third, the computational cost of real-time semantic analysis is non-trivial. Implementing a BERT-based analyzer for every incoming message in a large enterprise poses scalability challenges. Future work must focus on "Model Distillation"—compressing these large language models into lighter, faster versions that can run on edge devices or efficient cloud instances without incurring prohibitive latency or cost.

Finally, the arms race is continuous. As defenders adopt intent-based detection, attackers will likely move toward "poisoning" the context. This could involve attackers compromising legitimate email threads to inject malicious commands into an established, trusted context—a technique known as conversation hijacking. Defending against this will require models that not only analyze the current message but also validate the historical continuity of the relationship between sender and receiver.

Conclusion

The democratization of generative AI has provided social engineers with a powerful toolkit, enabling attacks that are psychologically astute, technically flawless, and massively scalable. This study has demonstrated that traditional intrusion detection mechanisms, which rely on static signatures and metadata, are insufficient against this new class of "cognitive threats."

By applying the Five-Factor Model of personality, we have shown that susceptibility to these attacks is not

uniform; it is a variable dependent on the specific psychological makeup of the target. Consequently, defense mechanisms must evolve to be equally context-aware. The Context-Aware Defense System (CADS) proposed in this paper represents a step toward this evolution. By leveraging the semantic understanding capabilities of LLMs for defense, we can detect the subtle cues of manipulation that evade keyword filters.

However, technology alone is not the panacea. The defense against AI-powered social engineering requires a holistic approach that combines "Smart Defense" algorithms with "Resilient Human" protocols. As we move forward, the security community must embrace the reality that we are no longer just securing networks; we are securing the cognitive interfaces of the people who run them. The future of cybersecurity lies in the successful symbiosis of human intuition and artificial intelligence, working in concert to discern truth from fabrication in an increasingly synthetic digital world.

References

1. Rajgopal, P. R..(2025). AI Threat Countermeasures: Defending Against LLM-Powered Social Engineering. International Journal of IoT, 5(02), 23-43. https://doi.org/10.55640/ijiot-05-02-03
2. Pastor-Galindo, J., Nespoli, P., Mármol, F.G., Pérez, G.M.: The not yet exploited goldmine of osint: Opportunities, open challenges and future trends. IEEE Access 8, 10282–10304 (2020)
3. Parrish Jr, J.L., Bailey, J.L., Courtney, J.F.: A personality based model for determining susceptibility to phishing attacks. Little Rock: University of Arkansas pp. 285–296 (2009)
4. Jones, C.R., Bergen, B.K.: People cannot distinguish gpt-4 from a human in a turing test. arXiv preprint arXiv:2405.08007 (2024)
5. Matz, S., Teeny, J., Vaid, S.S., Peters, H., Harari, G., Cerf, M.: The potential of generative ai for personalized persuasion at scale. Scientific Reports 14(1), 4692 (2024)
6. Kantardzic, M. (2011). Data mining: Concepts, models, methods, and algorithms. John Wiley& Sons.
7. Kayacik, H. G., Zincir-Heywood, A. N., & Heywood, M. I. (2005). Selecting features forintrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets.Proceedings of the Third Annual Conference on Privacy, Security and Trust.
8. Lee, W., & Stolfo, S. J. (2000). A framework for constructing features and models for intrusiondetection systems. ACM Transactions on Information and System Security (TISSEC), 3(4), 227-261.
9. Kim, J., Kim, H., & Cho, J. (2020). User satisfaction with biometric systems in eCommerce: Astudy on fingerprint scanning. Journal of Information Security and Applications, 54, 102512.
10. Kumar, A., & Singh, P. K. (2021). A review of deep reinforcement learning for cybersecurityapplications. IEEE Access, 9, 126245-126267.
11. McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications. Journal of Personality 60(2), 175–215 (1992)
12. Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Willer, R., Liang, P., Bernstein, M.S.: Generative agent simulations of 1,000 people. arXiv preprint arXiv:2411.10109 (2024)
13. AI Threat Countermeasures: Defending Against LLM-Powered Social Engineering. (2025). International Journal of IoT, 5(02), 23-43. https://doi.org/10.55640/ijiot-05-02-03
14. Irhimefe Otuburun. “Real-Time Fraud Detection Using Large Language Models: A Context-Aware System for Mitigating Social Engineering Threats.” World Journal of Advanced Research and Reviews, vol. 26, no. 3, 30 June 2025, pp. 2811–2821, https://doi.org/10.30574/wjarr.2025.26.3.2491