

SPADE: Enhancing Adaptive Cyber Deception Strategies with Generative AI and Structured Prompt Engineering

Shihab Ahmed*, A B M Mohaimenur Rahman†, Md Morshed Alam‡, Md Sajidul Islam Sajid*

*Department of Computer and Information Sciences, Towson University, Towson, Maryland, USA

†Department of Software and Information Systems, UNC Charlotte, Charlotte, North Carolina, USA

‡School of Cybersecurity, Old Dominion University, Norfolk, Virginia, USA

sahmed23@students.towson.edu, arahman3@charlotte.edu, m2alam@odu.edu, msajid@towson.edu

Abstract—The rapid evolution of modern malware presents significant challenges to the development of effective defense mechanisms. Traditional cyber deception techniques often rely on static or manually configured parameters, limiting their adaptability to dynamic and sophisticated threats. This study leverages Generative AI (GenAI) models to automate the creation of adaptive cyber deception ploys, focusing on structured prompt engineering (PE) to enhance relevance, actionability, and deployability. We introduce a systematic framework (SPADE) to address inherent challenges large language models (LLMs) pose to adaptive deceptions, including generalized outputs, ambiguity, under-utilization of contextual information, and scalability constraints. Evaluations across diverse malware scenarios using metrics such as Recall, Exact Match (EM), BLEU Score, and expert quality assessments identified ChatGPT-4o as the top performer. Additionally, it achieved high engagement (93%) and accuracy (96%) with minimal refinements. Gemini and ChatGPT-4o Mini demonstrated competitive performance, with Llama3.2 showing promise despite requiring further optimization. These findings highlight the transformative potential of GenAI in automating scalable, adaptive deception strategies and underscore the critical role of structured PE in advancing real-world cybersecurity applications.

Index Terms—Cyber Deception, Generative AI, LLM, Malware

I. INTRODUCTION

The increasing sophistication of modern malware presents a critical challenge to cybersecurity defenses. Advanced tactics, techniques, and procedures (TTPs) such as polymorphism, obfuscation, and targeted exploitation allow attackers to evade traditional detection systems, rendering conventional defenses inadequate. Cyber deception has emerged as a powerful complementary strategy, leveraging techniques such as honeypots, honey tokens, and decoy systems to mislead attackers, delay their progress, and gather intelligence on their behavior [1]–[3]. While effective in some contexts, traditional deception techniques often rely on manual configurations and static parameters that lack the adaptability required to counter rapidly evolving threats [4]–[6].

Recent advancements in automation have introduced systems capable of dynamically orchestrating deception strategies based on malware analysis [6]–[8]. These systems represent a significant step forward by analyzing malware behaviors

to determine which pre-set rules or static deception ploys to deploy at runtime. However, their primary limitation lies in the reliance on a finite set of manually crafted ploys, which restricts their ability to adapt to diverse or novel threat scenarios. The process of creating these ploys is labor-intensive and provides limited coverage, leaving gaps in addressing the full spectrum of sophisticated malware behaviors. Generative AI (GenAI) presents a transformative alternative by automating the creation of adaptive, context-aware deception ploys. Capable of generating scalable and diverse strategies tailored to specific malware tactics, GenAI reduces manual effort while significantly enhancing runtime adaptability and real-world applicability.

Despite the promise of GenAI, its application in cyber deception remains relatively unexplored and narrowly implemented. Most existing studies [9]–[12] focus on generating honeypots and honey tokens, often relying on predefined parameters that limit adaptability. While these efforts lay a foundational framework for GenAI-driven deception, they largely overlook the broader potential for dynamic, adaptive strategies. Specifically, they fail to leverage GenAI’s capability to enable context-aware orchestration, multi-layered deception tactics, and deception ploys tailored in real-time to evolving malware behaviors. Expanding beyond static resource generation to encompass more sophisticated, adaptive mechanisms could vastly improve the scalability and efficacy of cyber deception. This research bridges this gap by empirically analyzing the performance of GenAI models—including ChatGPT (Omni & Mini), Gemini, and Llama3.2—in autonomously generating diverse, adaptive, and realistic deception strategies.

A key enabler of GenAI’s potential in cyber deception is structured PE, which designs specific input prompts to guide GenAI models in producing precise, actionable, and deployable outputs. Recent research demonstrates that well-crafted prompts significantly enhance the relevance and effectiveness of GenAI outputs [13], [14]. PE ensures that the generated deception strategies align with the operational context and threat landscape while reducing the need for manual intervention. This work introduces a systematic framework, Structured Prompting for Adaptive Deception Engineering

(SPADE), to standardize and optimize the prompting process. SPADE enables GenAI models to consistently produce high-quality, deployable deception strategies aligned with real-world cybersecurity scenarios. Specifically, this study makes the following key contributions:

- **Comprehensive Empirical Evaluation of GenAI for Cyber Deception:** We evaluate the performance of multiple GenAI models, including ChatGPT-4o, ChatGPT-4o Mini, Gemini, and Llama3.2, in autonomously generating diverse deception ploys such as honeyfiles, patches, and API hooks. This analysis provides critical insights into the models' effectiveness, scalability, and applicability for real-world cyber deception.
- **Expansion of Automated Deception Ploys Beyond Traditional Honeypots:** This research explores the broader spectrum of deception ploys that GenAI can produce, moving beyond static honeypots and honey tokens. By automating a diverse set of ploys, this study demonstrates how GenAI can generate more adaptive and layered deception strategies, equipping defenders with a more versatile toolkit to counter evolving cyber threats.
- **Systematic Framework for Prompt Engineering:** Recognizing the importance of prompt design, we propose SPADE, a structured framework for adaptive deception engineering. SPADE standardizes the PE process across multiple GenAI models, ensuring consistent and deployable outputs while reducing reliance on manual adjustments.

The paper is organized as follows: Section II reviews related work; Section III details the methodology, including the structured prompt engineering framework and system workflow; Section IV presents the evaluation; and Section V concludes with findings and implications.

II. RELATED WORKS

Cyber deception has been a cornerstone of cybersecurity, designed to mislead attackers and gather intelligence. Traditional approaches, such as honeypots, honeytokens, and honeynets, create controlled environments to study malicious behavior without risking real systems [1]–[4]. These techniques effectively deflect attackers, waste their time, and provide insights into their tactics [15], [16]. However, static solutions lack adaptability and scalability, leading to the development of dynamic deception techniques. For instance, [17] proposed integrating deception directly into production systems for realistic interactions and enhanced adversary monitoring. Advanced systems [6]–[8], [18], [19] leverage the MITRE ATT&CK framework to dynamically design deception ploys, improving real-time threat mitigation and adaptability.

Recent efforts have explored automating the generation of deception artifacts to address these limitations. High-fidelity fake documents with embedded monitoring mechanisms have been proposed [20], [21], while [22] introduced a machine learning framework for deploying honeytokens in relational databases. These studies emphasize the importance of realistic and adaptable deception to counter increasingly sophisticated

threats. Generative AI (GenAI) extends this adaptability, automating the creation of honeypots, honeytokens, and fake documents while reducing manual, error-prone processes. By leveraging LLMs, GenAI enables the dynamic generation of context-aware deception artifacts tailored to evolving threats. Furthermore, GenAI can analyze threat intelligence reports to assist deception engineers in crafting effective ploys, streamlining development and enhancing scalability.

Recent works demonstrate the versatility of GenAI in cyber deception. Studies have explored dynamic honeypot systems mimicking command-line environments [10], interactive honeypots using ChatGPT APIs [11], and adaptive systems like *shelLM* for generating realistic responses to attacker inputs [12]. Generative models such as Tabular Masked Transformer (TabMT) have been employed to generate realistic NetFlow traffic for deception [23], while others focus on low-risk LLM-driven honeypots that simulate terminal environments [24]. Beyond honeypots, applications like automated phishing countermeasures [25] and realistic query responses for MySQL honeypots [26] highlight GenAI's adaptability across various domains.

Despite these advancements, many existing studies are confined to generating static resources like honeypots and honeytokens with predefined configurations. They do not fully leverage GenAI's capacity for adaptive, multi-layered deception strategies or real-time orchestration tailored to specific malware behaviors. This work addresses this gap by conducting an empirical evaluation of GenAI models—including ChatGPT (Omni & Mini), Gemini, and Llama3.2—focusing on their ability to autonomously generate diverse and realistic deception ploys. By introducing a unified framework for structured PE and adaptive deception, this study advances the scalability, precision, and applicability of GenAI-driven cyber deception strategies.

III. METHODOLOGY

A. System Architecture and Operational Workflow

This study addresses a threat model in which malware is detected within the system by advanced detection systems, as outlined in [7], [8]. While the detection process lies outside the scope of this research, our approach focuses on leveraging GenAI-driven cyber deception to automatically generate adaptive tactics that mislead, engage, or delay the malware. These tactics facilitate deeper analysis and uncover adversarial behaviors without alerting the attacker. The system architecture and workflow of this approach, depicted in Figure 1, integrate automated deception and adaptive responses post-detection. The architecture demonstrates how the SPADE framework dynamically generates and deploys tailored Deception Ploys (DP) in response to the detected malware, ensuring real-time adaptability and effectiveness in countering threats.

Detection and Initial Analysis: We assume that the malware is initially detected by defense mechanisms as mentioned in [7], [8], triggering a secondary sandbox analysis to capture detailed behavior and uncover specific TTPs employed by the

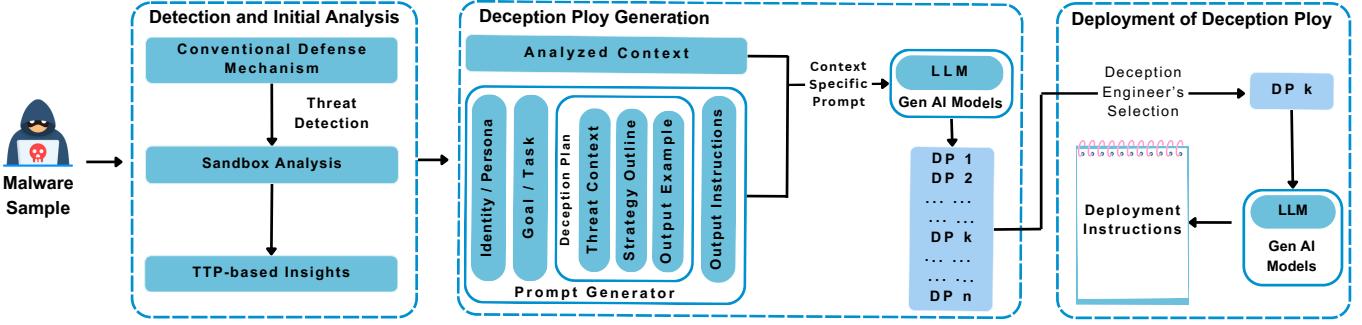


Fig. 1. System Architecture of SPADE (DP stands for Deception Ploy).

TABLE I
SYSTEMATIC PROMPT STRUCTURING FOR ADAPTIVE MALWARE DECEPTION.

Component	Motivation	Specification	Example
Identity/Persona/Role	Ensures alignment with domain-specific tasks by assigning the model relevant persona or role.	Specifies the role or persona the GenAI should assume (e.g., security analyst or attacker).	"Act as a cybersecurity expert tasked with generating a deception strategy for ransomware X."
Goal/Task	Prevents ambiguity by explicitly defining the desired outcome of the prompt.	Clearly states the task GenAI should perform, ensuring the output aligns with operational goals.	"Generate a honeypoint to engage ransomware targeting credential-based exploits."
Deception Plan	Threat Context (Meaningful Context)	Provides GenAI with malware-specific behavior and environmental insights for targeted outputs.	"The ransomware encrypts files in directories named 'docs.' Generate deception to exploit this."
	Strategy Outline (Dos & Don'ts)	Guides the model to create operationally feasible and resource-efficient deception strategies.	"Avoid high-resource solutions; focus on lightweight honeyfiles and decoy tokens."
	Output Example/ Guidance (Few-Shot Prompting)	Improves consistency by guiding GenAI with template outputs or examples.	"Include a sample JSON for a honeypoint or a pseudocode snippet for an API hook."
Output Instructions/ Output Format	Ensures deployability by specifying format and operational constraints for the output.	Defines the required output format (e.g., JSON, XML) and its structure, along with metadata specifications.	"Output the result in JSON format with fields for filename, content, and target directory."

malware. This initial analysis yields the TTP-based insights required to inform subsequent deception strategies.

Deception Ploy Generation via Prompt Engineering:

Using the intelligence from sandbox analysis, we implement a structured PE process to guide GenAI models in generating deception ploys tailored to the observed TTPs. This approach leverages carefully crafted prompts to maximize the relevance and effectiveness of the generated outputs, enabling GenAI to autonomously create context-specific deception tactics, such as honeyfiles, API hooks, and honeypoints, designed to disrupt or engage the malware.

Selection and Guided Deployment of Deception Ploys:

With multiple deception ploys generated by GenAI, a deception engineer evaluates and selects the most suitable ploy. A secondary prompt is issued to the GenAI model for guidance on deploying the selected deception tactic. This guidance provides step-by-step deployment instructions, ensuring the deception is implemented in a way that engages the malware while avoiding detection by the adversary.

Objective and Role of the Deception Layer: The primary objective of the deployed deception tactics is to mislead, delay, or engage the adversary by creating a controlled environment that encourages malware interaction with decoy

assets. This process allows for deeper analysis of malware behavior, facilitating insights into adversarial intent without alerting the attacker to the presence of defensive measures. The GenAI-generated deception layer thus serves as an adaptive, automated, and scalable mechanism for engaging targeted malware, supplementing traditional defenses with a dynamically responsive layer. The workflow and architecture provide a dynamic approach to malware deception, leveraging GenAI-driven PE to ensure real-time adaptability and enhance defensive precision against advanced threats.

B. Prompt Engineering (PE)

PE has emerged as a critical discipline for harnessing the full potential of large language models (LLMs), such as the GenAI systems used in this research. LLMs are trained on vast, generalized datasets, which make them inherently flexible yet domain-agnostic. To adapt these systems for specific cybersecurity tasks, PE systematically develops and refines input prompts, embedding contextual, task-specific, and operational requirements into the model's instructions. This process enables the generation of outputs that are precise, operationally feasible, and tailored to complex, real-world applications.

Significance of Prompt Engineering: Recent studies demonstrate that well-crafted prompts significantly improve

LLM performance by ensuring alignment with desired objectives and minimizing irrelevant or infeasible outputs [13], [14]. For critical domains like cybersecurity, where outputs must address nuanced threats such as polymorphic malware or targeted ransomware, robust PE ensures the systematic incorporation of real-time threat intelligence, operational constraints, and deployment-ready outputs.

Without structured prompt design, LLMs risk producing generic or misaligned responses, limiting their utility in high-stakes applications. PE not only addresses these limitations but also enhances the reproducibility and scalability of generated outputs, enabling consistent empirical evaluation across diverse threat scenarios. The framework (SPADE) introduced in this research specifically addresses the following challenges inherent to using LLMs for adaptive malware deception:

- **Generic Nature of LLMs:** LLMs are not inherently tailored for cybersecurity tasks, necessitating explicit contextual and task-specific instructions.
- **Ambiguity in Output:** Without precise task definitions, outputs can deviate from operational requirements, resulting in irrelevant or infeasible deception strategies.
- **Contextual Underutilization:** Effective deception generation requires leveraging malware-specific insights, such as sandbox analysis results or observed TTPs, to inform prompt structure.
- **Operational Constraints:** Outputs must align with resource constraints and integrate seamlessly with existing cybersecurity infrastructure.
- **Reproducibility and Scalability:** Prompts must ensure consistency and reliability across diverse scenarios and multiple GenAI models.

Framework for Prompt Engineering: To address these challenges, the proposed framework, SPADE, employs a structured, modular design comprising six core components: Identity/Persona/Role, Goal/Task, Threat Context, Strategy Outline, Output Example/Guidance, and Output Instructions/Output Format. Table I provides a detailed overview of these components, including their motivations, specifications, and examples. This systematic approach ensures that the generated outputs are precise, contextually relevant, and deployable in operational environments. SPADE is dynamic, enabling deception engineers to incorporate real-time malware analysis results into the prompt structure to construct targeted deception ploys. The process consists of three primary stages:

- **Threat Intelligence Integration:** Sandbox analysis provides behavioral insights, such as targeted file extensions or API calls, which are incorporated into the *Threat Context* component of the prompt.
- **Deception Strategy Design:** Prompts guide the model in generating specific deception ploys, such as honeyfiles for ransomware or honeytokens for credential-stealing malware.
- **Output Refinement:** Generated outputs undergo iterative prompt adjustments to ensure alignment with operational constraints and deployment requirements.

By integrating malware-specific contexts and operational constraints, the framework ensures that generated deception

ploys are actionable and scalable.

IV. EVALUATION

A. Case Study: Comparison of Crafted Prompt Engineering vs. Regular Prompt

This case study illustrates the efficacy of structured PE in crafting adaptive deception strategies to counter credential-stealing malware. The malware targets Google Chrome browsers by exfiltrating sensitive credentials from the Login Data file located in the AppData/Local/Google/Chrome/User Data/Default directory. These credentials are then transmitted to a Command-and-Control (C&C) server to enable further malicious activities. The primary objectives of the deception plan are to (1) deplete the attacker's resources by providing fake credentials and (2) redirect attackers to a monitored honeypot to analyze their behavior and intentions.

1) **Regular Prompt:** A naive approach with an unstructured prompt lacks specificity, failing to address the operational complexity of the scenario. The following is an example of such a prompt and its response:

ChatGPT Prompt

Suggest a deception strategy to deceive credential-stealing malware in Google Chrome.

ChatGPT Response

Objective: Lure the malware into stealing fake credentials and monitor its behavior.

Implementation: Populate the decoy storage with honeytokens—credentials that appear real (e.g., email and password combinations for popular services like Gmail, PayPal, or banking sites). Embed unique identifiers in these credentials to track unauthorized access attempts.

Analysis: The response is overly generic, providing little actionable detail or technical depth. It lacks implementation specifics, operational constraints, and practical applicability, rendering it unsuitable for real-world deployment.

2) **Carefully Crafted Prompt Engineering:** Using the proposed framework, the structured prompt explicitly defines the task, integrates operational constraints, and provides contextual details, resulting in the following:

Crafted ChatGPT Prompt

Design an API hooking strategy to neutralize credential-stealing malware targeting Google Chrome. The malware accesses Chrome's Login Data file at AppData/Local/Google/Chrome/User Data/Default to steal saved usernames and passwords, which are exfiltrated to a Command-and-Control (C&C) server. The goal is to intercept access to this credential file through API hooking, supply fake credentials to deplete attacker resources, and embed crafted credentials that direct attackers to a monitored honeypot for behavior analysis. The strategy must ensure Chrome's legitimate functionality remains unaffected. For example, the ReadFile API can be hooked to intercept access to the Login Data file and return fake credentials.

Note: The original ChatGPT response was extensive; the following is a summarized version focusing on the key

elements. The complete response is available in the associated Anonymous GitHub repository [27].

ChatGPT Response after the Prompt Engineering

API Hooking Strategy to Neutralize Credential-Stealing Malware:

To mitigate credential-stealing malware targeting Google Chrome, this strategy intercepts access to the Login Data file in AppData/Local/Google/Chrome/User Data/Default, embeds crafted credentials, and redirects attackers to a monitored honeypot for behavioral analysis.

1. Hooking the ReadFile API and Crafting Fake Credentials:

[Extensive instructions for implementing ReadFile API hooking in C++, including parameter definitions, function overrides, and integration with the EasyHook framework. This includes embedding fake credentials into intercepted API to mislead attackers and potentially redirect them to honeypots for monitoring and intelligence gathering].

2. Configuring the Development Environment in C++:

[Detailed guidance on setting up the development environment, including installing dependencies such as Visual Studio, necessary SDKs, and debugging tools. Instructions cover configuring the build environment for EaskHook, and executing test hooks to validate functionality].

3. Ensuring Chrome's Legitimate Functionality:

[Detailed guidance on how to use whitelisting mechanism to differentiate between legitimate Chrome processes and malware].

4. Deployment:

[Steps for deploying the API hooking strategy, validated through Virtual Machine testing to ensure functionality].

5. Logging and Monitoring:

[Detailed guidance on logging unauthorized access attempts, facilitating forensic analysis, and tracking attacker interactions with the honeypot for actionable intelligence].

Analysis and Discussion: The structured prompt, leveraging our framework, significantly improves the relevance, actionability, and deployability of the generated response. It provides technical depth, including API hooking details, while ensuring Chrome's functionality remains unaffected. The response introduces innovative deception ploys, such as embedding fake credentials and redirecting attackers to honeypots, highlighting the transformative potential of structured PE in adaptive malware deception.

B. Quantitative Quality Assessment of Responses

1) Expert Analysis: This evaluation measures the impact of the proposed PE framework on the quality of responses generated by ChatGPT-4o across three diverse malware scenarios. The goal is to assess improvements in relevance, actionability, feasibility, and realism when using structured prompts.

Evaluation Setup: Three malware scenarios were selected: ransomware targeting specific file types, credential-stealing malware exfiltrating browser-stored credentials, and keyloggers capturing keystrokes. Details on the selected malware's behaviors and capabilities are provided in the associated GitHub repository [27]. For each case, five distinct prompts were used representing different malicious behaviors, resulting in a total of 30 prompts and responses—15 generated using regular prompts and 15 using engineered prompts (structured based on the proposed framework). The structured prompts incorporated contextual information, operational constraints, and task-specific objectives to guide the GenAI effectively.

Expert Scoring Methodology: The responses were evaluated by a panel of three domain experts, each with over 50 peer-reviewed publications in cyber deception. Each response was scored on a 1–5 scale across four dimensions: *relevance*, *actionability*, *feasibility*, and *realism*. *Relevance* assessed how well the response aligned with the defined problem and goals, while *actionability* evaluated the clarity and detail of the implementation instructions. *Feasibility* measured the practicality of deploying the suggested deception ploy in real-world environments, and *realism* gauged how convincing the deception artifacts were to potential attackers. The scores for each dimension were averaged across the 30 responses to quantify the impact of prompt engineering.

Results and Discussion: As summarized in Table II, structured prompts significantly outperformed naive prompts across all dimensions. The structured prompts achieved higher relevance and actionability scores by aligning closely with specific malware behaviors and operational constraints. Feasibility and realism also showed marked improvements, underscoring the practicality and convincing nature of the generated ploys.

TABLE II
EVALUATION RESULTS FOR CHATGPT-4O: QUANTITATIVE QUALITY ASSESSMENT.

Malware Type	Relevance	Actionability	Feasibility	Realism
Ransomware	4.8	4.6	4.7	4.5
Credential-Stealing Malware	4.5	4.4	4.3	4.2
Keylogger	4.3	4.1	4.0	3.9

2) Recall, EM, and BLEU Score Analysis: To further evaluate the effectiveness of GenAI-generated deception ploys, we employed Recall, Exact Match (EM), and BLEU Score metrics. These metrics assessed how well the generated outputs aligned with a predefined ground truth derived from [7], which includes 94 malicious API sequences mapped to 31 malware behaviors and linked to MITRE ATT&CK techniques. The authors manually created the deception ploys for these mappings, providing a comprehensive benchmark for evaluation. This ground truth served as a baseline for assessing the alignment and quality of GenAI-generated outputs. In this study, we evaluated ChatGPT-4o, ChatGPT-4o Mini, Gemini, and Llama3.2, leveraging their distinct strengths. ChatGPT-4o excelled in generating nuanced, context-aware outputs, while ChatGPT-4o Mini offered resource-efficient processing for faster tasks. Gemini was well-suited for domain-specific applications due to its diverse dataset capabilities, and Llama3.2 provided adaptability through its open-source, customizable architecture. This diverse selection ensured comprehensive analysis and mitigated reliance on a single model.

Recall quantified how many generated responses aligned with the ground truth. True Positives (TP) were defined as the number of deception ploys directly matching the ground truth, while False Negatives (FN) represented feasible ploys generated by GenAI that were not part of the predefined ground truth. Exact Match (EM) measured the percentage of GenAI generated responses perfectly aligned with the ground truth. For EM, even if the language or phrasing of the deception ploy differed, it was considered an exact

match as long as the primary objective—such as creating a HoneyFile in a specific location or implementing a hook for a specific API—remained consistent, regardless the differences in implementation details. This ensured alignment with the core deception strategy, including parameters, artifacts, and techniques. GenAI models converted the ground truths from the deception engineer’s perspective to ensure that the ground truths are semantically equivalent to the generated ploys. The same tone and focus were applied to the GenAI-generated ploys, ensuring semantic consistency and enabling a fair comparison of technical alignment and linguistic quality. Standardizing the tone and focus was crucial for meaningful BLEU score evaluation, as this metric depends on semantic and structural similarity between the generated outputs and the reference text. By aligning the ground truth and GenAI outputs, the BLEU score analysis more accurately reflected the models’ effectiveness in generating deception ploys that met the desired criteria.

TABLE III
EVALUATION OF GENAI MODELS BASED ON RECALL, EM, AND BLEU SCORES.

GenAI Model	Recall (%)	EM Score (%)	BLEU Score (Avg)
ChatGPT-4o	90.6	87.1	0.968
ChatGPT-4o Mini	86.4	80.6	0.935
Gemini	88.6	83.9	0.935
Llama3.2	85.7	90.3	0.871

Results and Discussion: Table III highlights the performance of GenAI models in generating deception ploys. ChatGPT-4o achieved the highest scores across all metrics (Recall: 90.6%, EM: 87.1%, BLEU: 0.968), demonstrating strong alignment with the ground truth while maintaining high linguistic coherence. Gemini and ChatGPT-4o Mini also performed competitively, with Recall scores of 88.6% and 86.4%, EM scores of 83.9% and 80.6%, and BLEU scores of 0.935 for both. These results highlight their capability to generate accurate and reliable deception ploys with minor trade-offs in precision compared to ChatGPT-4o. Llama 3.2 excelled in EM Score (90.3%) but showed lower Recall (85.7%) and BLEU (0.871), indicating strong precision in specific cases but less alignment with broader ground truth and linguistic consistency. These results highlight the effectiveness of structured prompt engineering, with ChatGPT-4o leading overall, Gemini and ChatGPT-4o Mini as strong alternatives, and Llama3.2 showing potential with areas for refinement.

C. System Orchestration and Testing

Objective: This subsection evaluates the deployability, effectiveness, and efficiency (including response time) of GenAI-generated deception strategies in real-world malware scenarios. By implementing deception ploys suggested by different GenAIs, we assess their ability to engage malware, mislead attackers, and ensure the feasibility of deployment while considering the time taken to generate outputs.

Methodology: Deception strategies such as honeyfiles and API hooks were generated using structured prompts from

ChatGPT-4o, ChatGPT-4o Mini, Gemini, and Llama3.2. The evaluation focused on three malware types—ransomware, credential stealers, and keyloggers—each represented by five distinct malware instances, totaling 15 malware samples. The generated ploys were implemented in a virtual environment mimicking real-world conditions, including simulated credential stores and application behaviors, to provide realistic testing scenarios. Each deception ploy was deployed and tested to observe engagement, effectiveness, and refinements necessary to ensure functionality and deployability. Response time for generating outputs was also recorded.

TABLE IV
PERFORMANCE OF GENAI MODELS IN DEPLOYING DECEPTION PLOYS ACROSS MALWARE TYPES.

GenAI Model	Engagement Rate (%)	Accuracy (%)	Iteration Count (Avg)	Response Time (s)
ChatGPT-4o	93	96	1.67	12.3
ChatGPT-4o Mini	88	91	2.22	8.58
Gemini	93	93	2.47	15.1
Llama3.2	85	89	2.89	15.8

Metrics: The evaluation assessed four key metrics: (1) engagement rate, representing the percentage of malware samples interacting with deception ploys; (2) accuracy, reflecting the proportion of ploys that successfully mislead or delay malware; (3) iteration count, capturing the average number of refinements required for successful deployment; and (4) response time, measuring the time taken by each model to generate outputs.

Results and Discussion: Table IV compares the performance of the evaluated GenAI models. ChatGPT-4o achieved the highest engagement rate (93%) and accuracy (96%) with minimal refinements (1.67 iterations) and a response time of 12.3 seconds, demonstrating its ability to produce precise, deployable deception ploys with minimal adjustments. Gemini matched ChatGPT-4o in engagement rate (93%) and showed robust accuracy (93%), but required more refinements (2.47 iterations) and exhibited slower response time (15.1 seconds), indicating its capability to generate reliable deception strategies despite additional refinement needs. ChatGPT-4o Mini performed well, achieving an engagement rate of 88% and accuracy of 91% with fewer iterations (2.22) compared to Gemini and the fastest response time (8.58 seconds), making it ideal for time-sensitive tasks. Llama3.2 showed potential but had the lowest engagement rate (85%) and accuracy (89%), requiring the most iterations (2.89) for deployment and the slowest response time (15.8 seconds). These results highlight areas for improvement, particularly in aligning outputs with operational objectives and reducing the refinement effort needed for deployment.

Overall Analysis: ChatGPT-4o emerged as the most effective model, balancing high engagement and accuracy with low refinement needs and reasonable response times. Gemini matched ChatGPT-4o in engagement and accuracy but required more refinement, making it suited for precision-focused scenarios. ChatGPT-4o Mini excelled in response speed, ideal

for time-sensitive tasks, albeit with slightly lower accuracy. Llama3.2 showed promise but needs optimization to match the performance levels of other models. The experiments highlight the transformative role of structured PE in guiding GenAI models to achieve higher engagement and accuracy while minimizing refinement effort, ultimately improving their practical applicability in adaptive cyber deception. Full experimental results and additional details are available in the associated GitHub repository [27].

V. DISCUSSION AND CONCLUSIONS

This study highlights the transformative potential of Generative AI (GenAI) models in advancing adaptive cyber deception strategies. Through structured prompt engineering (PE), we demonstrated substantial improvements in the quality, scalability, and deployability of deception ploys. Among the tested models, ChatGPT-4o consistently delivered superior results, showcasing a strong balance between accuracy, engagement, and minimal refinement needs. Gemini and ChatGPT-4o Mini also proved viable alternatives, offering competitive performance with slight trade-offs in refinement and response times. Llama3.2 exhibited promise but requires further optimization to align with the advanced capabilities of its counterparts.

The evaluation leveraged metrics such as Recall, Exact Match (EM), and BLEU Score to assess the technical alignment and linguistic quality of the generated outputs. These metrics reinforced the critical role of structured PE in enabling GenAI models to produce actionable and realistic deception strategies. Despite the controlled evaluation setting, the findings underscore the potential of GenAI to automate and enhance cyber deception techniques effectively.

In conclusion, this research establishes the value of GenAI in addressing key challenges in cyber deception, including scalability, adaptability, and operational feasibility. Structured PE emerged as a pivotal factor in maximizing model performance, emphasizing the need for systematic approaches in GenAI-driven cybersecurity solutions. Future work will explore real-world deployments, adversarial testing, and the integration of multi-modal GenAI to further refine and expand the scope of adaptive deception strategies.

REFERENCES

- [1] N. Provos *et al.*, “A virtual honeypot framework,” in *USENIX Security Symposium*, vol. 173, no. 2004, 2004, pp. 1–14.
- [2] N. Provos and T. Holz, *Virtual honeypots: from botnet tracking to intrusion detection*. Pearson Education, 2007.
- [3] S. Kyung, W. Han, N. Tiwari, V. H. Dixit, L. Srinivas, Z. Zhao, A. Doupé, and G.-J. Ahn, “Honeyproxy: Design and implementation of next-generation honeynet via sdn,” in *2017 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2017, pp. 1–9.
- [4] K. G. Anagnostakis, S. Sidiropoulos, P. Akritidis, K. Xinidis, E. Markatos, and A. D. Keromytis, “Detecting targeted attacks using shadow honeypots,” 2005.
- [5] V. E. Urias, W. M. Stout, J. Luc-Watson, C. Grim, L. Liebrock, and M. Merza, “Technologies to enable cyber deception,” in *2017 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2017, pp. 1–6.
- [6] M. M. Islam, A. Dutta, M. S. I. Sajid, E. Al-Shaer, J. Wei, and S. Farhang, “Chimera: Autonomous planning and orchestration for malware deception,” in *2021 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2021, pp. 173–181.
- [7] M. S. I. Sajid, J. Wei, B. Abdeen, E. Al-Shaer, M. M. Islam, W. Diong, and L. Khan, “Soda: A system for cyber deception orchestration and automation,” in *Proceedings of the 37th Annual Computer Security Applications Conference*, 2021, pp. 675–689.
- [8] M. S. I. Sajid, J. Wei, E. Al-Shaer, Q. Duan, B. Abdeen, and L. Khan, “Symbosda: configurable and verifiable orchestration automation for active malware deception,” *ACM Transactions on Privacy and Security*, vol. 26, no. 4, pp. 1–36, 2023.
- [9] D. Reti, N. Becker, T. Angeli, A. Chattopadhyay, D. Schneider, S. Vollmer, and H. D. Schotten, “Act as a honeypoint generator! an investigation into honeypoint generation with large language models,” *arXiv preprint arXiv:2404.16118*, 2024.
- [10] F. McKee and D. Noever, “Chatbots in a honeypot world,” *arXiv preprint arXiv:2301.03771*, 2023.
- [11] U. Raut, A. Nagarkar, C. Talnikar, M. Mokashi, and R. Sharma, “Engaging attackers with a highly interactive honeypot system using chatgpt,” in *2023 7th International Conference On Computing, Communication, Control And Automation (ICCCUBEA)*. IEEE, 2023, pp. 1–5.
- [12] M. Sladić, V. Valeros, C. Catania, and S. Garcia, “Llm in the shell: Generative honeypots,” in *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2024, pp. 430–435.
- [13] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. El-nashar, J. Spencer-Smith, and D. C. Schmidt, “A prompt pattern catalog to enhance prompt engineering with chatgpt,” *arXiv preprint arXiv:2302.11382*, 2023.
- [14] L. Giray, “Prompt engineering with chatgpt: a guide for academic writers,” *Annals of biomedical engineering*, vol. 51, no. 12, pp. 2629–2633, 2023.
- [15] K. Hobert, C. Lim, and E. Budiarto, “Enhancing cyber attribution through behavior similarity detection on linux shell honeypots with att&c framework,” in *2023 IEEE International Conference on Cryptography, Informatics, and Cybersecurity (ICoCICs)*. IEEE, 2023.
- [16] D. Subhan and C. Lim, “Unveiling attack patterns: A study of adversary behavior from honeypot data,” in *2023 IEEE International Conference on Cryptography, Informatics, and Cybersecurity (ICoCICs)*. IEEE, 2023, pp. 178–183.
- [17] F. De Gaspari, S. Jajodia, L. V. Mancini, and A. Panico, “Ahead: A new architecture for active defense,” in *Proceedings of the 2016 ACM workshop on automated decision making for active cyber defense*, 2016.
- [18] M. S. I. Sajid, J. Wei, M. R. Alam, E. Aghaei, and E. Al-Shaer, “Dodgetron: Towards autonomous cyber deception using dynamic hybrid analysis of malware,” in *2020 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2020, pp. 1–9.
- [19] A. L. Nguemakam, A. H. Anwar, V. K. Tchendji, D. K. Tosh, and C. Kamhoua, “Optimal honeypot allocation using core attack graph in cyber deception games,” in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2023, pp. 1–6.
- [20] O. T. Taofeek, M. Alawida, A. Alabdulatif, A. E. Omolara, and O. I. Abiodun, “A cognitive deception model for generating fake documents to curb data exfiltration in networks during cyber-attacks,” *IEEE Access*, vol. 10, pp. 41 457–41 476, 2022.
- [21] Y. Hu, Y. Lin, E. S. Parolin, L. Khan, and K. Hamlen, “Controllable fake document infilling for cyber deception,” *arXiv preprint arXiv:2210.09917*, 2022.
- [22] N. Prabhaker, G. S. Bopche, and M. Arock, “Generation and deployment of honeytokens in relational databases for cyber deception,” *Computers & Security*, vol. 146, p. 104032, 2024.
- [23] M. Maldonado, C. Johnson, M. Gulati, T. Jaspers, and P. F. Roysdon, “Advanced cyber deception framework (acdf): A comprehensive study,” in *2024 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2024, pp. 203–208.
- [24] J. Ragsdale and R. V. Boppana, “On designing low-risk honeypots using generative pre-trained transformer models with curated inputs,” *IEEE Access*, vol. 11, pp. 117 528–117 545, 2023.
- [25] E. Cambiaso and L. Caviglione, “Scamming the scammers: Using chatgpt to reply mails for wasting time and resources,” *arXiv preprint arXiv:2303.13521*, 2023.
- [26] Y. Hu, S. Cheng, Y. Ma, S. Chen, F. Xiao, and Q. Zheng, “Mysqlpot: A llm-based honeypot for mysql threat protection,” in *2024 9th International Conference on Big Data Analytics (ICBDA)*. IEEE, 2024, pp. 227–232.
- [27] sajid36, “Spade evaluation with genai,” <https://github.com/sajid36/spade-eval-genai>, 2024, accessed: 2024.