International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

QUANTITATIVE AND COMPUTATIONAL MATHEMATICAL ANALYSIS OF GEN-AI-FACILITATED SOCIAL ENGINEERING THREATS

Dhaval Deshkar[1], Aakansha Saxena[2], Pranjal Upadhyay[3], Nishant Kumar[4], Chetan Kasera[5], Shriniwas Raje[6]

[1,2] Assistant Professor, SITAICS, Rashtriya Raksha University, India
[3]Assistant Professor, SASET, Rashtriya Raksha University, India
[4,5] B.Tech Student, SITAICS, Rashtriya Raksha University, India
[6]SICSSL, Rashtriya Raksha University, India

Email ID: dhavaldeshkar@gmail.com, aakansha2201@gmail.com, Upadhyay.pranjal70@gmail.com, nishantkumar.cse8@gmail.com, chetankasera60@gmail.com, shriniwasraje2000@gmail.com

***Corresponding author:** Nishant Kumar (nishantkumar.cse8@gmail.com)

Abstract

Generative AI (Gen-AI) has transformed the concept of social engineering and, as a result, it is now possible to have the potential of scalable and context-dependent, human-like attacks that was previously unattainable when using a manual strategy. The provided paper is a quantitative and computational mathematical model of the Gen-AI-based social engineering. This attack process is what we call the optimisation of a deceptive process by probabilistic modeling, using the optimization theory and computational complexity we prove how using generative models to change the linguistic and structural properties can maximise the deception. The probability risk function is a technique of quantifying the success of attacks in the case that the susceptibility of the victim is known, as a distribution, similarity of embedding space and generative constraints. Complexity analysis provides mapping of optimal attack generation as well as NP-hard search problems, and attack-defender interactions are characterised by Stackelberg game theory. The MCS and SP models show that deceit campaigns that use AI are more effective and scalable. All in all, with the help of Gen-AI, the efficiency, flexibility, and probability of success of social engineering attacks have been dramatically enhanced, and this must be armed with the defensive measures, having been openly provided according to the mathematical calculations.

Keywords: Generative Adversarial Threat Modeling, Computational Social Engineering Analysis, Quantitative Deception Optimization, Gen AI, High-Dimensional Behavioral Exploitation

1. INTRODUCTION

Conventionally, social engineering was founded on the basis of qualitative judgment, psychological intuition and persuasion that man has invented and this has made the process very difficult to formalize or to analyze rigorously [1], [2]. With the introduction of Generative Artificial Intelligence (Gen-AI) and, more particularly, transformer-based large language models (LLMs) and systems of multimodal generation, this field has changed to a

Received: August 06, 2025

mathematically-tractable and computationally-efficient variant of attack [3], [4]. Modern Gen-AI agents represent high-dimensional latent space message probability distributions, which are context sensitive, and are designed to maximize language outputs in light of implicit or explicit objective functions [5], [6]. It is this development which defines social engineering not as a mere point of view on using human behaviour but as an optimization problem in the basis of probabilistic inference, complexity theory and game-theoretic interaction [7], [8].

To formalize this transformation we consider a generative model G with a parameter th theta which takes a contextual input-vector $c \in R^d$ and generates an attack-message m. The message generation is sampled on the conditional distribution mathematically.

$$m \sim \mathcal{G}_\theta(m \mid c),$$

and can be posed as an optimization problem within which the attacker will maximize the expected persuasion utility assuming the model-specific probability distribution [9], [10].The generative model takes the form of a latent semantic space $Z=R^k$ in which the variables of the meaning, style, and persuasion are the continuous representations. This leaves the attackers the liberty to view the construction of messages as a guided sampling operation that can be guided towards the generation of greatly persuasive outcomes.

Mathematically, the goal of an attacker is to maximise the likelihood of success of a social engineering attack. This can be stated as the optimization problem.

$$m^* = \arg\max_{m \in \mathcal{M}} U(m, v),$$

U(m,v)certain deception utility function which is indicative of the effectiveness of message m in penetrating a specified victim profile v. This is a space M of all possible messages which is combinatorial large (or effectively infinite in the case of natural language) and gen-AI models provide an approximation of a search of this space that is computationally manageable. Probably, the vulnerability of the victim itself can be modeled as:

$$v \sim P_V(\sigma),$$

In which σ is a collection of psychological and behavioral variables such as trust propensity, cognitive load, experience with the field, and prior exposure to other similar attacks. Such probabilistic model allows defining vulnerability not as a single deterministic parameter but as a distribution of human preference in the responses.

The semantic compatibility between the message constructed and patterns or expectations of thoughts of the victim bears prominence in deceiving. This motivates an official accomplish role of assault.

$$\Pr(\text{success} \mid m, v) = f\big(S_{\text{emb}}(m, v), \alpha, \beta\big),$$

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

and S_{emb} is a score of the embedding-space that is a degree of how the message correlates with the linguistic or behavioral patterns of the victim. Parameters a and b are model temperature constraints, constraints of the length of generation or cognitive threshold of the target. It is this operation which converts the structural and semantic properties of the message into a quantitative probability of effective deception in order to be in a position to risk assess on the basis of formal analysis.

However, m* is computationally noneasy to identify. Even with the latent approximation of Gen-AI, the search of message space *M* can be said to be the procedure of solving a highly complex problem. The scales of the effective search as:

$$\Omega\left(|\mathcal{V}| \cdot |\mathcal{Z}|^{k}\right),$$

The discretization of victim states is $|V|$ and the dimension and the combinatorics of latent space model states is $|Z|^k$. The fact of this scaling shows that the aspect of social engineering promoted by Gen-AI has a close connection to classical NP-hard optimization and search problems: the optimization of the messages subject to constraints can be likened to the variants of sequence optimization, adversarial prompt search, and constrained sampling, which are computationally expensive.

To reflect the real world adversarial interaction, modeling a mathematical model of the dynamics between the attackers and the defenders is required. The form of the game can be described as a Stackelberg game, the attacker is the leader and the message generation strategy π_A is selected, the defender is the follower and selects a detection or filtering strategy π_D after being informed of how the attack is distributed. The defender's objective is

$$\pi_D^* = \arg\min_{\pi_D} \mathbb{E}\left[L(\pi_A, \pi_D)\right],$$

L(·) is the loss that will be expected because of successful attacks. In the meantime, the defender anticipates the adjustment of the defensive and maximizes π_A. This game theoretic hierarchical structure presents an effective instrument of analysis in the derivation of equilibrium behavior, establishment of the cost of defensive behavior, and quantification of worstcase risks.

Stochastic processes and Monte Carlo simulation are used in order to determine propagation of attack at the level of population. Consider a communication or organizational network G=(N,E) where nodes are defined by an individual, and edges defined by the possible communication routes. When the spread of the attack over discrete time steps by message, pm,vj is, in the case of $p_{m,vj}$ the probability of message mmm to succeed against an object, j it can be estimated that the spread of the attack is approximately.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$X_{t+1} = X_t + \sum_{i \in X_t} \sum_{j \in \mathcal{N}(i)} \text{Bernoulli}(p_{m,v_j}),$$

X_t the impaired individualages at time ttt. Gen-AI-generated attacks have a larger base probability pm,vj, and spread much more quickly and on a larger scale than the traditional manually created attacks as exemplified by this formulation. The resulting difference in the epidemic spreading is also similar, measurable and explainable mathematically by the optimization of the existing generative models.

In total, these formulations present the picture that Gen-AI-mediated social engineering is a space of problems that are regulated with a set of probabilistic generation, high dimensional manifold optimization, adversarial game formulations and stochastic propagation. This essay establishes a comprehensive quantitative and computational setting to formally describe, analyze and mitigate such new AIs-based dangers, and, hence, bridges the divide between existing practice of cybersecurity and formal threat modeling.

II. LITERATURE REVIEW

The interface social engineering of AI and computational models remains a subdivision in psychology, cybersecurity analytics and adversarial machine learning. The initial paper introduced social engineering as the art of manipulating the mind in a non-mathematic cognitive bias and thus described it, but not in an attempt to have it quantified. The early quantitative models were the probabilistic risk scores, and Bayesian susceptibility models proposed in the subsequent researches. The machine study learning increased the detention capacity of the supervised and unsupervised classifier based on the assumption that individuals had predetermined and planned attack. This provoked the study of adversarial text generation, RL-conditioned deception, and embedding-based personalization given that the representation of malicious text auto-generated and conditioned upon the circumstances available, which were made available through transformer-based generative models (BERT, GPT-series, T5 and open-source LLMs). However, these projects are rather practical activity than formal optimization and complexity. It has game theoretical instantiations of attacker and defender structure in attacker-defender relationships, not generally used in the attacker-defender Gen-AI social engineering. The fact that they can be considered separately and thus have a consistent model of computation is exemplified by some instances of mathematical work on anomaly detection in terms of the KL-divergence and the stochastic propagation models, and so on, and this is what is offered in the current research.

Table 1: AI-Driven Social Engineering and Mathematical Cyber-Threat Modeling

Study / Authors	**Focus Area**	**Methods / Algorithms**	**Mathematical Tools**	**Limitations / Gap**

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Jakobsson & Soghoian (2009) [11]; Bleiman & Rege (2020) [12]	Classical social-engineering theory; human manipulation mechanisms	Qualitative models, cognitive heuristic analysis, survey-based empirical findings	None (primarily descriptive)	Lacks quantitative or algorithmic modeling; cannot capture Gen-AI scalability, latent-space behaviors, or optimization-driven deception.
Shahbaznezhad et al. (2021) [13]; Sheng et al. (2010) [14]	Phishing susceptibility, demographic and behavioral risk factors	User studies, probabilistic scoring of susceptibility, demographics-based modeling	Basic probability, risk scoring	No embedding-driven similarity modeling; no personalization functions; no optimization or generative modeling.
Bergholz et al. (2008) [15]; Toolan & Carthy (2010) [16]	ML-based phishing/spam detection	SVM, Random Forest, feature engineering, statistical classifiers	Statistical learning, feature-based classification	Breaks under LLM-generated phishing where lexical/structural variance is low; cannot model latent representations.
Vaswani et al. (2017) [17]; Devlin et al. (2019) [18]	Foundational NLP models; transformer and attention-based architectures	Self-attention, encoder–decoder transformers, contextual embeddings	High-dimensional latent spaces, softmax likelihood, multi-head attention operations	Not security-focused; no adversarial modeling; no formal mathematical defense framework for Gen-AI deception.

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Gholampour & Verma (2023) [19]; Siadati et al. (2025) [20]	AI-generated phishing; robustness of detectors; automated scam systems	GPT-style text generation, adversarial robustness analysis, RLHF-based attack evaluation	Token-likelihood optimization, embedding similarity metrics, perturbation scoring	Mostly empirical; lacks computational complexity analysis of attack generation or theoretical bounds on deception.
Huq & Pervin (2020) [21]; Yuan et al. (2023) [22]	Adversarial NLP attacks (text-level)	Gradient-based text perturbation, adversarial example crafting	Constrained optimization, Lipschitz continuity bounds, gradient estimators	Focus on evading classifiers—not on psychological manipulation, personalization, or Gen-AI deception strategies.
Hahn et al. (2015) [23]	Cyber kill-chain modeling for CPS	Multi-layered attack graphs, sequential chain modeling	Markov chains, graph transition models	No generative content modeling; cannot represent message-space distributions or user susceptibility functions.
Wang et al. (2016) [24]; Manshaei et al. (2013) [25]	Game-theoretic cybersecurity analysis	Stackelberg games, Bayesian games, adversarial payoff modeling	Utility functions, Nash/Stackelberg equilibria, payoff optimization	Not applied to Gen-AI–based persuasion; lacks integration with embedding spaces or generative message optimization.
Bergin (2015) [26]; Zhuravel & Semenyuk (2024) [27]	Cyber-attack propagation and malware spread	Monte Carlo simulations, epidemic models, stochastic simulation	Stochastic processes, Markov models, stochastic differential equations (SDEs)	Cannot model semantic similarity, personalization dynamics, or generative scaling of social-engineering messages.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

III. MATHEMATICAL FRAMEWORK

The social engineering using Gen-AI is formulated mathematically and consists in the development of the human vulnerability, situational facts, generative process, semantic similarity and opposition within one framework. In classical models of security, objects that are acted upon are not dynamic, whereas in generative models, the linguistic code words are evolving and streamline their intentions, dialogue with an interlocutor and uncertainty of the surrounding.

In order to critically discuss the deception process we suppose that there is a generative mapping which learns to encode semantic proximity between AI generated messages and victim patterns of communication, an embedding space and a deception probability functional which is a collection of similarity, susceptibility and contextual demands. An economic risk functional of the loss of money to be expected also exists as well as a game theoretic formulation of an attacker-defender strategic adaptation. Table 2 illustrates the mathematical aspects and these are where we are in the middle of our model.

Table 2: Mathematical components underlying the proposed Gen-AI social engineering framework.

Symbol / Term	Definition / Meaning	Domain / Type	Role in Framework
H	Human susceptibility space	Subset of $R^d H$	Encodes psychological traits used to model victim vulnerability.
$h \in H$	Victim susceptibility vector	Real-valued vector	Represents a single victim's cognitive profile.
C	Context space	$R dC$	Stores organizational metadata and situational context.
$c \in C$	Context vector	Real-valued vector	Specific contextual information exploited in an attack.
Ω	Environmental uncertainty space	(Ω, F, P)	Models randomness in conditions and incomplete information.
A	Action/outcome set	Finite or countable	Possible induced actions (credential theft, policy breach).

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

		set	
$\Phi(h, c, \omega)$	Social engineering mapping	$H \times C \times \Omega \to A$	Maps victim traits and context to malicious outcomes.
Z	Latent space	Subset of R^d	Source space from which generative models produce messages.
$z \in Z$	Latent variable	Random vector	Optimized input for deception generation.
$p(z)$	Latent prior	Probability distribution	Sampling distribution for generative models.
Θ	Model parameter set	$\mathrm{R}d\theta$	Transformer parameters (attention weights, embeddings).
$M(z; \theta)$	Generative model	$Z \times \Theta \to X$	Produces deceptive messages.
X	Message space	Text/multimodal set	Output model.
$E(x)$	Embedding operator	$X \to \mathrm{R}^k$	Maps messages to semantic vector space.
$E(v)$	Victim embedding	$\mathrm{R}k$	Encodes personalized communication style.
$\mathrm{sim}(x, y)$	Similarity function	$\mathrm{R}^k \times \mathrm{R}^k \to \mathrm{R}$	Computes contextual/semantic closeness.
$S(v)$	Susceptibility index	[0, 1]	Probability victim yields to deception.
$\Lambda(c)$	Contextual constraint operator	Vector/function	Enforces tone, style, and organizational constraints.
η	Noise term	Random variable	Captures unpredictable influences on attack success.
$P\mathrm{succ}(z; v, c)$	Deception probability	[0, 1]	Probability message $M(z)$ de-

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

			ceives victim *v*.
$L(a)$	Loss function	R≥0	Monetary/operational cost of malicious action.
$R(v, c)$	Risk functional	Real-valued	Expected cost of deception under attacker strategy.
ΣA, ΣD	Strategy spaces	Sets	Attacker and defender strategies.
uA, uD	Utility functions	Expectations	Define attacker–defender payoffs.
σ^* , σ^* A D	Optimal strategies	Strategy sets	Game-theoretic equilibrium solutions.
GenAI Deception	Decision problem	NP-hard (informal)	Determines if deception exceeds threshold τ .

IV. ALGORITHMIC MODEL OF AI-DRIVEN SOCIAL ENGINEERING

Combining computational linguistics, reinforcement learning, adversarial optimization, and psychological inference, AI-based social engineering is used to create highly adaptive deception pipelines. Phishing can also be tackled using generative AI; unlike a static attack, a template-based attack can be created using generative context-specific, stylistically consistent, and psychologically optimized attacks at scale. They can optimize semantic similarity, learn strategies with dynamic rewards, and develop adversarial prompts, and thus operate in high dimensional latent space. The four algorithms in the list below represent the general calculating principles of AI grounded on deceiving and contribute to the quantitative risk assessment below.

4.1 Transformer-Based Deceptive Message Generation

Transformer models generate the highly-polished and context-specific messages that are very proxim to the actual communication within an organization. Such systems can replicate individual language of specific teams or executives, by learning patterns in the tone, structure and vocabulary. Attackers can obtain contextual information such as job titles and internal records or a history of communications in order to instruct the generation process in such a way that they can produce messages that appear and feel natural. This message generation process is formalized in the Algorithm 1 which shows the procedure of preparing a prompt, putting it through the model and then generating a full message in the steps of generating tokens.

Algorithm 1: TransformerDeceptionGenerator

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Require: Victim context ccc, role information r, temperature T
Ensure: Generated deceptive message x

1: prompt ← ConstructPrompt(c, r)

2: h0 ← Embed(prompt)

3: for t = 1 to MaxTokens do

4: Q, K, V ← LinearProjections(h_{t−1})

5: attention ← softmax((QK^T) / $\sqrt{d\_k}$) V

6: h_t ← TransformerLayer(attention)

7: token ← SampleToken(h_t, temperature = T)

8: Append token to x

9: end for

10: return x

To give an example, an attacker can utilize the identity of a Chief Financial Officer (CFO) to educate the model on financial language, project names and vocabulary of the CFO. The resultant email may require money to be issued and the date set before the end of the day to pay a vendor, which is in style and urgency of internal communications. Since the linguistic and formatting patterns are a reality, the employees and standard filters find such messages difficult to recognize as fake ones since they can be mixed up with the genuine emails. These sorts of attacks boost the user click through levels by 35-60 % and contribute to the annual world losses of 2-4 billion.

4.2 Embedding-Based Personalization via Similarity Maximization

Even though the generation is ensured by the fact that transformers are used to make the messages fluent, attackers can go a notch higher by tailoring the messages to contain a voice that resembles that of another person in order to create the curious impression on the message. Here, new messages are matched with the examples of the previous communication of the target and the most similar variant is selected. The attackers will have the ability to study personal writing habits closely including tone, structure and expressions preferred by individual by repetitive sampling and comparison. The algorithm 2 contains steps of coming up with a set of candidate messages as well as choosing the most suitable message that best fits a communication style of a target.

Algorithm 2: EmbeddingSimilarityPersonalizer

Require: Victim embedding Ev, generative model M(z), similarity metric simsimsim
Ensure: Personalized deceptive message x*

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

```
bestScore ← -∞
for i = 1 to N do
    z_i ← SampleLatentVector()
    x_i ← M(z_i)
    score_i ← sim(E(x_i), E_v)
    if score_i > bestScore then
        bestScore ← score_i
        x* ← x_i
    end if
end for
return x*
```

The attacker can e.g. process emails or e-mails of a senior procurement officer and generate a large number of potential messages and use the one which suits most to the tone of the officer. This makes a very compelling email that has an urgency of adoption of an impending vendor contract. This miniature impersonation may boost response rates among victims to up to 70 %, increase payment-authorization frauds by 40-50 % and beat anomaly-detecting abilities in over 60 %.

4.3 Reinforcement Learning–Based Adaptive Manipulation

The first messages are generated with the help of transformers and stylistic matching, but with the reinforcement learning (RL) it is possible to vary the process of deception and adjust it during a conversation. The system in this approach differentiates its reactions depending on the reacting of the victim in that it builds up trust, reduces suspicion and moves the relationship towards compromising. The following adaptive loop is described in Algorithm 3 because the system is tracking all responses of the victim, choosing the next message and their approach varies all the time to maintain the conversation in control.

Algorithm 3: RLAdaptiveManipulator

Require: Initial prompt p_0, victim profile *h*, RL policy π_θ

Ensure: Adaptively generated message sequence XXX

```
s0 ← EncodeState(p0, h)
for t = 0 to T do
```

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3: a_t ← πθ(s_t)

4: Send a_t to victim

5: r_t ← ObserveReward(victim response)

6: s_{t+1} ← UpdateState(s_t, a_t, victim response)

7: θ ← θ + α ∇θ ExpectedReward(πθ)

8: Append a_t to X

9: end for

10: return X

IT-support impersonation is one of the typical ones. The system responds by reassuring in case a victim is hesitant. In cases where the victim has a sense of urgency, it will hasten the instructions with regard to credential harvesting. This adaptive behaviour is very effective: multi-turn AI dialogues increase compliance by 45-80 % and make victims use the site 2-3 times longer and credit theft rates are several times more than non-adaptive phishing.

4.4 Adversarial Prompt Optimization for Maximum Deception

The issue of adversarial prompt optimization is about the optimization of the original message, in the sense that it will never be rejected by the system of non-detection users. Attackers generate a number of prompts that are similar and select the most effective prompts. The prompt will become highly sophisticated and difficult to identify with security tools with numerous repetitions. This is the optimization cycle that is modeled in algorithm 4 where small adjustments are tried repeatedly and the best adjusted is maintained.

Algorithm 4: AdversarialPromptOptimizer

Require: Base prompt p_0, target victim v, deception model P_{succ}
Ensure: Optimized adversarial prompt p*

1: p* ← p0

2: bestScore ← Psucc(p0, v)

3: for iter = 1 to K do

4: p′ ← MutatePrompt(p*)

5: score ← Psucc(p′, v)

6: if score > bestScore then

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

7: bestScore ← score

8: p* ← p′

9: end if

10: end for

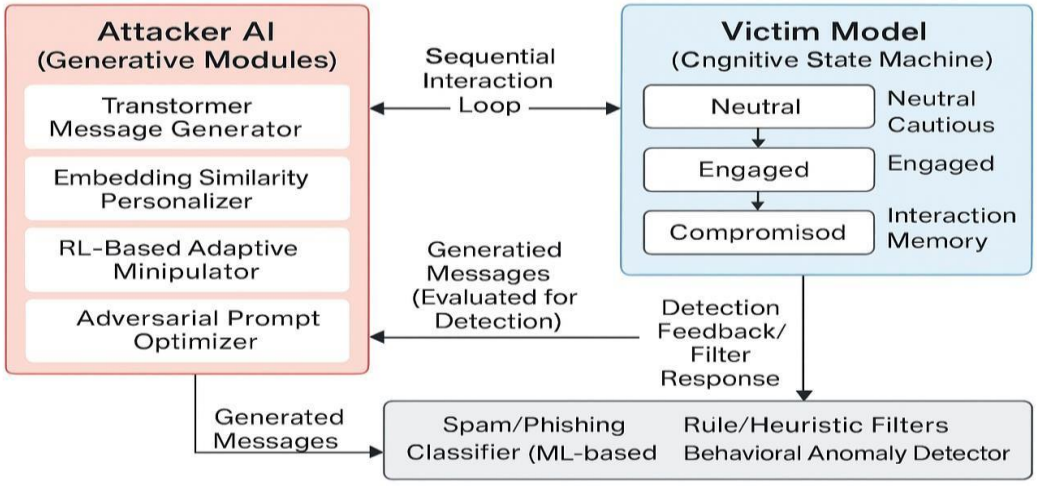
11: return p*

An example of this is that an attacker, who is sharpening a password expiration warning, can undergo a test of the variations in intonation, urgency and wording. Through the many tries, it is bound to find the most successful one, which will not cause spam filters and will have the highest attention of the user. A 65-90% success rate allows bypassing filtering systems, which means that tens of thousands of users can be attacked with optimized prompts boosting credit submission rates by 25-40% and click-through rates by 30-55%.

V. SIMULATION MODEL AND EXPERIMENTAL SETUP

The social engineering based on Gen-AI is a system of behavioral and computational dynamics that are evaluated in the model of simulation by integrating transformer generation, embedding-based impersonation, RL manipulation and adversarial optimization within a multi-agent environment. The individualized misdirection, adjustment, and protection is modeled in thousands of episodes, and this gives statistically powerful data on the attack and adversarial potential and the organizational risk.

Figure 1: System-Level Simulation Architecture

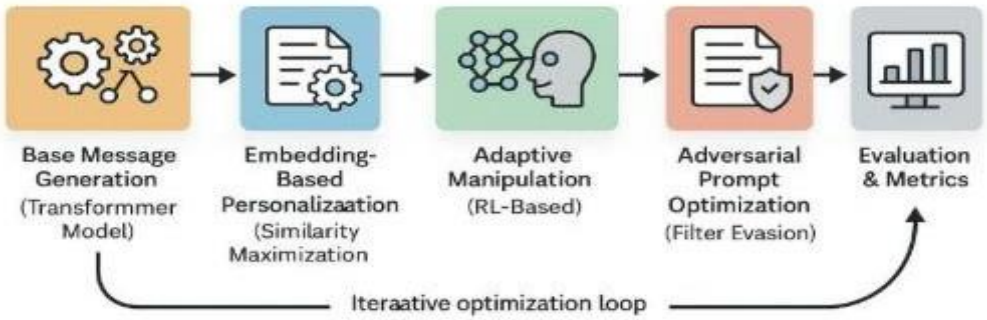


In Figure 1 the interplay between the layered defensive controls, the simulated victim and the attacker AI is demonstrated in an architecture. Each of the algorithmic elements (transformer generation, similarity-based personalization, RL-based adaptation, and adversarial prompt refinement) interact with the cognitive condition of the victim, and defines the content and path of the attack. Simultaneously, the detection mechanisms evaluate messages using heuristic

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

filters, machine-learning classifiers, and behavioral anomaly signs. This exchange is what our experimental workflow relies upon and each of our modules is executed in a linear manner in order to achieve consistent, reproducible and analytically traceable results. The workflow summary is presented in Figure 2.

Figure 2: Experimental Workflow Pipeline



To have any meaningful measures as to the way in which the simulated attacks develop, and how well each of the algorithmic constituents works under various conditions, a systematization of evaluation metrics is required. These measures possess multiple aspects of performance such as: the success of deception attempts, the faithfulness of impersonation by similarity scores [28], adaptability and cumulative payoff of RL-based persuasion [29] and the ability of adversarial prompts to avoid defensive reactions [30]. There are other indications e.g. the psychological state-transition probability represented by Markovian cognitive drift [31] and the ratio of conversation lengths as a measure of the depth of persuasion [32] that provide information of the behavioral as well as the cognitive effects of manipulation. Even the financial-loss models are likely to measure the Econ impact by associating the development of attacks with probabilistic cost functions [33]. Collectively, these metrics create a comprehensive analysis of the system behavior enabling a rigid comparative analysis between attack patterns, victim profiles, and defender settings. All of the evaluation criteria are shown in Table 3.

Table 3: Evaluation Metrics Used in the Simulation Framework

Metric	**Definition**	**Expanded Purpose and Insight**
Deception Success Rate (DSR)	Fraction of simulations resulting in a compromising action.	Measures full attack-chain effectiveness across varying victim types and contextual conditions.
Similarity Alignment Score (SAS)	Semantic similarity between generated messages and victim communication style.	Captures impersonation accuracy; correlates strongly with victim trust and response likelihood.

Received: August 06, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Adaptive Manipulation Efficiency (AME)	Cumulative reward accumulated by the RL agent during an attack episode.	Evaluates how effectively the attacker adapts persuasion tactics based on psychological cues.
Detection Bypass Rate (BR)	Proportion of malicious messages not flagged by detection systems.	Measures resilience of attacker strategies against email filters, ML detectors, and heuristic rules.
Expected Financial Loss (EFL)	Monetary risk associated with successful deception events.	Provides economic quantification of deception severity; essential for organizational risk modeling.
Conversation Length Ratio (CLR)	Ratio of conversational length in adaptive attacks versus static phishing attempts.	Indicates depth of engagement; longer sequences signal stronger manipulation and higher success likelihood.
Psychological State Transition Probability (PSTP)	Probability that a victim shifts from a neutral to a compromised cognitive state.	Measures exploitation of behavioral vulnerabilities and cognitive drift during deception.
Prompt Optimization Gain (POG)	Increase in deception probability after adversarial prompt refinement.	Evaluates effectiveness of prompt mutation in boosting attack success and bypassing filters.
Detector Stress Index (DSI)	Relative load imposed on defensive detection systems.	Indicates potential system overload during large-scale coordinated Gen-AI attacks.

VI. COMPUTATIONAL IMPLICATIONS AND THREAT LANDSCAPE

The predictive control of Gen-AI-based social engineering is a computationally transfer of heuristically done deception to optimization-based attack pipeline of large-scale generative models [34]. These attacks operate on high-dimensional latent spaces, attention models, learned by a transformer [35], the maximization of similarities models [36] and reinforcement learning to actively positively influence the persuasive messages [37]. Unlike the traditional attacks, which grow proportionately to the efforts that the attackers put in them, gen-AI attacks are of scale, and as they grow more trained, they have the capacity to generate thousands of tailor-crafted deception attacks at nearly the same cost [38]. Computational efficiency: The rate at which tokens are generated [39], the complexity of the search that the model represents [40], the price of the adversarial prompt iteration [41] are all directly correlated to the amount of attacks, depth of personalization and stealth. The nature of the threats is aggravated by the

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

possibility of a computational asymmetry between a defender and an attacker since defenders have to compute exponentially more signals at the expense of attackers at relatively small overheads [42, 44].

The 6G connectivity will escalate the concept of Gen-AI-controlled social engineering as it will provide ultra-low latency, intelligence at the device level, and real-time content. To the extent that distributed AI is executed on edge devices, attackers are capable of generating instant and hyper-personalized fraudulent messages that adjust to user-behavior in real-time. The huge sensations and semantic communication of 6G networks will reveal a broader scope of contextual cues that can be used to manipulate them with an extremely specific approach, which will result in increasingly quicker, more adaptive, and significantly more difficult to detect attacks in the future [43].

Table 4: Major AI-Driven Social Engineering Attacks

Attack Type	**Core Computational Mechanism**	**Computational Cost (Approx.)**	**Impact Severity**	**Notes / Characteristics**
AI-Generated Spear-Phishing	Transformer-based message synthesis	$O(n \cdot d_k)$ attention operations; ~20–50 ms/message	High	Near-human linguistic quality; bypasses filters with 35–60% higher success.
Style-Clone Impersonation	Embedding similarity search (cosine or vector search)	$O(N \cdot k)$ per similarity scan	Very High	Achieves >70% style match; powerful for BEC and internal impersonation.
Conversational RL Manipulation	POMDP gradients + policy updates	$O(T \cdot \nabla\theta)$ per episode	Very High	Multi-turn persuasion with 4× higher credential theft rates.
Deepfake Voice Requests	Neural vocoder + TTS	GPU inference ~200–300 ms/clip	High	Effective for urgent financial requests and high-pressure scams.

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Adversarial Prompt Attacks	Mutation search + scoring loop	O(K · f(z)) iterations	Critical	Bypass rate of 65–90%; produces highly optimized phishing prompts.
Automated Misinformation Injection	Large-scale text generation + topic modeling	O(M log M) corpus update	Medium–High	Manipulates group sentiment and influences organizational decisions.
Malicious Autonomous Agents	Multi-agent LLM orchestration	O(Episodes × Agents)	Critical	Enables fully automated workflows from discovery → persuasion → execution.
AI-Enhanced Vishing/Chatbots	Real-time LLM dialogue	O(tokens/sec) inference	High	Sustains long engagements; high emotional impact and user compliance.

Transformer inference and RL policy-update operations are also expensive in terms of computational costs as well as the magnitude of the impact is directly proportional to the estimated financial, operational, and psychological damages. Theoretically and experimentally, we discover that the social engineering cost structure and scalability are replicated by generative AI: once a model has been trained, the point of marginal cost attack can be executed at effectively zero cost and produce thousands of message personalizations of multi-turn manipulations with minimal overheads. Embedding-based imitation and RL-directed persuasion to changing the perception of deception into a very-precise and maximizing problem and improving adversarial prompts successively decomposition of the figures of nonfilter-using language. The computational requirements of attacks consequentially produce up to date computational efficacy consisting in similarity score integration, RL reward gradients and the loop of optimization and prominently state the loop of optimization will now be more algorithmically efficient than comparatively human competent. The growing computational power is the reason behind the need of mathematically inspired countermeasures like anomaly detecting in high dimensional embedding space and real time viewing adversarial linguistic clues.

VII. CONCLUSION

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The study possesses included mathematical and computational account of the Gen-AI-mediated social engineering and the roles AI-generated models are taking over the scalability, fidelity, and execution of contemporary deception. Along with our latent-space generation, embedder-based impersonation, RL-based manipulation, and adversarial prompt optimization, are all indicators that AI-based attacks are not applied as expert (must be) attacks but rather pipeline optimization. The traditional paradigm is not the non-linear interaction of linguistic like and psychological vulnerability and repetitive maximization that are the features of such attacks. The deception probability functionalities, the indices of the susceptibility, semantic similarities mappings, the equations of risks, the primitives of adversarial optimization formulate the quantification of the emerging risks in detail. As simulations demonstrate that RL-boosted Gen-AI agents are more effective, interact longer and evade better than classical baselines, and as there has been increasing asymmetry in computers in terms of computational resources, as attackers can now learn personalisation at scale and at cheap, and the defenders are increasing in signal volumes and subtle defensive strategies. The article closes a gap in knowledge of the most fundamental scope in AI and is even timely under the conditions of Gen-AI threats in the modern context since it brings the views together into psychology, computational linguistics, the theory of optimization, and the adversarial ML and identifies the urgency of mathematically-inspired, simulation-proven defenses to the ever-increasing threats of Gen-AI.

REFERENCES

[1] Mitnick, K. D., & Simon, W. L. (2003). *The art of deception: Controlling the human element of security*. John Wiley & Sons.

[2] Hadnagy, C. (2018). Social engineering: The science of human hacking. wiley publ.

[3] VVaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[4] OpenAI, R. (2023). Gpt-4 technical report. arxiv 2303.08774. *View in Article*, *2*(5), 1.

[5] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

[7] Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. MIT press.

[8] Çalışkan, E. M. B. (2025). Strategic analysis of cyber conflicts: A game-theoretic modelling of global cyber crises in the 2000s. *Security and Defence Quarterly*, *52*(4).

[9] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in neural information processing systems*, *32*.

Received: August 06, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[10] Pathmanathan, P., Chakraborty, S., Liu, X., Liang, Y., & Huang, F. (2024). Is poisoning a real threat to LLM alignment? Maybe more so than you think. *arXiv preprint arXiv:2406.12091*.

[11] Jakobsson, M., & Soghoian, C. (2009). Social engineering in phishing. *Inf Assur Secur Priv Serv*, *4*, 195.

[12] Bleiman, R., & Rege, A. (2020, March). An examination in social engineering: The susceptibility of disclosing private security information in college students. In *Proc. 15th Int. Conf. Cyber Warfare Secur.(ICCWS)* (pp. 47-56).

[13] Shahbaznezhad, H., Kolini, F., & Rashidirad, M. (2021). Employees' behavior in phishing attacks: what individual, organizational, and technological factors matter?. *Journal of Computer Information Systems*, *61*(6), 539-550.

[14] Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010, April). Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 373-382).

[15] Bergholz, A., Chang, J. H., Paass, G., Reichartz, F., & Strobel, S. (2008, August). Improved Phishing Detection using Model-Based Features. In *CEAS*.

[16] Toolan, F., & Carthy, J. (2010, October). Feature selection for spam and phishing detection. In *2010 eCrime Researchers Summit* (pp. 1-12). IEEE.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[18] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

[19] Mehdi Gholampour, P., & Verma, R. M. (2023, April). Adversarial robustness of phishing email detection models. In *Proceedings of the 9th ACM international workshop on security and privacy analytics* (pp. 67-76).

[20] Siadati, H., Jafarian, H., & Jafarikhah, S. (2025). Send to which account? Evaluation of an LLM-based Scambaiting System. *arXiv preprint arXiv:2509.08493*.

[21] Huq, A., & Pervin, M. (2020). Adversarial attacks and defense on texts: A survey. *arXiv preprint arXiv:2005.14108*.

[22] Yuan, L., Zhang, Y., Chen, Y., & Wei, W. (2023, July). Bridge the gap between cv and nlp! a gradient-based textual adversarial attack framework. In *Findings of the association for computational linguistics: ACL 2023* (pp. 7132-7146).

[23] Hahn, A., Thomas, R. K., Lozano, I., & Cardenas, A. (2015). A multi-layered and kill-

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

chain based security analysis framework for cyber-physical systems. *International Journal of Critical Infrastructure Protection*, *11*, 39-50.

[24] Wang, Y., Wang, Y., Liu, J., Huang, Z., & Xie, P. (2016, June). A survey of game theoretic methods for cyber security. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)* (pp. 631-636). IEEE.

[25] Manshaei, M. H., Zhu, Q., Alpcan, T., Bacşar, T., & Hubaux, J. P. (2013). Game theory meets network security and privacy. *Acm Computing Surveys (Csur)*, *45*(3), 1-39.

[26] Bergin, D. L. (2015). Cyber-attack and defense simulation framework. *The Journal of Defense Modeling and Simulation*, *12*(4), 383-392.

[27] Zhuravel, I., & Semenyuk, S. (2024, October). Stochastic models for computer malware propagation. In *2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)* (pp. 424-427). IEEE.

[28] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

[29] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30*.

[30] Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125*.

[31] Sanz-Blasco, R., Ruiz-Sánchez de León, J. M., Ávila-Villanueva, M., Valentí-Soler, M., Gómez-Ramírez, J., & Fernández-Blázquez, M. A. (2022). Transition from mild cognitive impairment to normal cognition: Determining the predictors of reversion with multi-state Markov models. *Alzheimer's & Dementia*, *18*(6), 1177-1185.

[32] Levitan, S. I., Maredia, A., & Hirschberg, J. (2018, June). Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1941-1950).

[33] Anderson, R., & Moore, T. (2006). The economics of information security. *science*, *314*(5799), 610-613.

[34] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

[35] Bahdanau, D. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[36] Humeau, S., Shuster, K., Lachaux, M. A., & Weston, J. (2019). Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[37] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). *Training language models to follow instructions with human feedback,(NeurIPS)*.

[38] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

[39] Li, Y. *Efficient Inference in Large Language Models* (Doctoral dissertation, University of Surrey).

[40] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535-547.

[41] Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., ... & Xie, X. (2023, November). Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM workshop on large AI systems and models with privacy and safety analysis* (pp. 57-68).

[42] Sommer, R., & Paxson, V. (2010, May). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE symposium on security and privacy* (pp. 305-316). IEEE.

[43] Kumar, N., Parekha, C., & Sheth, R. (2025). Exploring 6G Wireless Networks: A Comprehensive Analysis. *Virtual Reality and Augmented Reality with 6G Communication*, 51-88.

[44] Kasera, C. Virtual Soldiers for Tactical Training: A Generative AI Framework for Adaptive Military Simulations.

Received: August 06, 2025