Mesopotamian journal of Computer Science
Vol. **(2025)**, 2025, **pp**. 234–246
DOI: https://doi.org/10.58496/MJCSC/2025/015; ISSN: 2958-6631
https://mesopotamian.press/journals/index.php/cs

Research Article

Synthesizing Deception: Countering Large Language Model-Generated Phishing Campaigns through Adaptive Semantic Anomaly Detection

Bekim Fetaji [1], Debabrata Samanta [1,*]

[1] *Rochester Institute of Technology, Computing and Information Technologies, RIT Campus, Kosovo*

ARTICLEINFO

Article History

Received 20 Jul 2025
Revised 17Aug 2025
Accepted 29 Aug 2025
Published 2 Sep 2025

Keywords

large Language Models

Phishing Detection

Anomaly Detection

Natural Language Processing,

Cybersecurity.

Abstract

The paper fills in a gap in the literature that demonstrates an insufficient number of sturdy detection schemes that can recognize the small semantic aberrations inherent in LLM-generated deceptive text. Our proposed co-design hybrid model is Semantic Anomaly Detection with Isolation Forest (SADI) model that combines the synergistic mixture of a fine-tuned transformer-based LLM for deep semantic feature extraction with Isolation Forest algorithm that detects anomalies efficiently. This study introduces SADI, an adaptive semantic-anomaly detector for large-language-model phishing emails. Using a corpus of 10 000 messages, SADI attains an F1 score of 0.981 (95 % CI 0.978–0.984) and processes a single message in 18 ms on consumer GPUs. An expanded evaluation against three public benchmarks and a live enterprise feed confirms robustness to prompt variation. Code, data splits, and a reproducible environment file accompany the paper. We also prepared the new, more challenging target dataset, that is phishing attacks synthesized by a variety of state-of-the-art LLMs, denoted LLM-Phish-Synth-2025, with this objective in mind. The results of our experiments on three publicly available data sets and our new corpus show that SADI received a higher F1- score of 0.981 compared to the baseline models, including separate fine-tuned LLMs, by a wide margin. The proposed SADI architecture is the first to combine semantic anomaly detection with adaptive, contamination-aware isolation in the context of LLM-generated phishing, addressing both scalability and evolving attack sophistication. The theoretical impact is the new architecture of architectural fusion that branched semantic anomaly detection and the practical advantage of a more robust defense solution against an ever-evolving threat in cyber space in addition to the provision of a new benchmark dataset to the research community. This approach is an efficient and scalable solution to combat the wave of the phishing campaigns generated by AI.

1. INTRODUCTION

Technology is often described as a two-edged sword, and the unleashing of powerful Large Language Models (LLMs) has already become both a boon and a curse thanks to its application in natural language understanding and generation.3 On the one hand, rich potential applications related to this technology have been opened.4 On the other, LLMs have also led to the democratization of resources enabling the creation of high-quality digital deception.3 Phishing attacks, being one of the cyber threats, have been amplified due to the enhanced ability to leverage this technology. The recent development enables attackers to craft grammatically perfect, contextually resonant, and emotionally compelling emails, which pass the traditional security filters using more subtly linguistic methods that reflect natural communication [4], [15], [17].5 These new attacks go beyond the poor wording used by the previous attacks as the attackers mimic the fundamentals of human persuasion and deception through language [15], [4]. Starting recently there are a few studies to benchmark green LLM performance to create and detect phishing, especially to small and midsize business environments [7]. These undertakings verify that models are capable of generating high-fidelity attacks which are difficult even to a savvy user. The root cause is that the existing legacy detection systems, one which is based on the ability to detect explicit signs of maliciousness as odd-looking links or blacklists of bad keywords, are not well matched to this new reality. They do not have the ability to see the nuances of incongruence of tone, intent and textual organization that reveals an artificially constructed message. Although there have been studies involving the potential applications of AI in anti-phishing systems through email text analysis in the real-time [5], most systems remain unable to handle the adaptive resistance property of the LLM-generated threats. The capacity of the attackers to endlessly mutate linguistic style and content make the capabilities of depicting the signature-based methods outdated. The need that this paper fills is that shortage of both adaptive and semantically aware detection models. Other works use LLMs in detection but tend to treat them as a black box classifier, train the classifier to

*Corresponding author. Email: debabrata.samanta369@gmail.com

label as "phish" or "ham" and then utilize their strong prediction performances to perform detection [20]. This leads to vulnerable adversarial fine-tuning issues since at inference the LLMs cannot successfully limit the scope of labels to either one side of a binary classification problem. The schemes also fail to generalize and can be exploited later by an attack with an unseen LLM. Various domains where anomaly detection has been suggested are given below.

Our contribution is twofold. First, we propose a novel hybrid architecture, the **Semantic Anomaly Detection with Isolation Forest (SADI)** model. This approach integrates a fine-tuned LLM, which acts as a powerful semantic feature extractor, with an Isolation Forest algorithm, an efficient method for identifying outliers in high-dimensional data. Instead of learning what a phishing email *is*, SADI learns the semantic signature of *legitimate* communication and flags any significant deviation as anomalous. This makes the system inherently more robust to novel and evolving attack vectors. Second, to rigorously test our model's capabilities, we developed and are releasing a new benchmark dataset: **LLM-Phish-Synth-2025**. This dataset was created by systematically prompting multiple leading LLMs to generate phishing emails with varying targets and sophistication levels, providing a challenging and realistic testbed that reflects the current threat landscape. By validating SADI on this novel corpus alongside three established public datasets, we provide compelling empirical evidence of its superior performance and generalizability, offering a significant step forward in the arms race against AI-driven deception.

Contrastive-learning detectors [1] and GPT-4 audit heads [2] raise recall but rely on large supervised corpora. By isolating semantic outliers under light supervision, SADI retains accuracy when attacker prompts evolve and labelled data remain scarce, situating the contribution within 2024–2025 detection research. LLM-driven phishing poses not only technical challenges but also societal risks, including potential exploitation in political, financial, and healthcare sectors. Responsible development and deployment of countermeasures is therefore essential to mitigate harms, uphold privacy, and maintain public trust in digital communication.

2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK

The battle against phishing is a continuous cat-and-mouse game, where defensive innovations are met with more sophisticated offensive tactics.[6] The current inflection point is driven by the accessibility of advanced LLMs built on transformer architectures [3]. This section reviews the literature on LLM-driven threats, the theoretical underpinnings of deception, and current detection paradigms to situate our contribution and articulate the research gap.

A. The LLM as a Tool for Deception

The capacity of LLMs to generate human-like text has not gone unnoticed by malicious actors. Early explorations showed LLMs could be exploited through deceptive prompting and persuasive principles to bypass safety filters [15]. More targeted research by Fairbanks and Serra [17] demonstrated the ease with which novel phishing attacks could be generated, highlighting the urgent need for new detection algorithms. The challenge is not merely generation but semantic alignment; LLMs can synthesize context with a precision that makes their output nearly indistinguishable from human-written text [6].[7] This ability directly threatens cybersecurity infrastructure, as models can be used for threat modeling [1] or to craft obfuscation-resistant phishing URLs [20], posing a significant risk. The work of Zhang et al. [7] provided a comprehensive benchmark, confirming that LLM-generated emails often evade standard detection tools used by enterprises. This body of work establishes the threat's severity and the inadequacy of current defenses. Deep one-class approaches such as DeepSVDD and InfoBERT have been applied to textual fraud [3], [4] but lack evaluation against generative phishing. SADI couples contextual embeddings with contamination-aware isolation to bridge this gap.

B. Theoretical Underpinnings of Deception and Manipulation

Effective phishing is fundamentally an exercise in psychological manipulation [4].[8] It preys on cognitive biases, urgency, and authority. Wright et al. [22] conceptualized phishing susceptibility from an information processing perspective, suggesting that deception detection is a multilevel cognitive task. LLMs excel at exploiting these levels by personalizing content at scale.[9] The linguistic cues of deception have been a long-standing area of research [8], [9], yet LLMs can be fine-tuned to avoid classic giveaways. Strategic deception, whether in environmental contexts [24] or digital ones, relies on creating a believable, internally consistent narrative. The ecological approach described by Jones [4] suggests that deception is intertwined with the communication environment. LLMs are adept at learning and mimicking the ecology of corporate or personal email exchanges, making their malicious outputs appear as a natural part of the environment. Countering such sophisticated deception requires moving beyond surface-level analysis to detect deeper semantic and pragmatic anomalies. Prompt-based counter-measures include adversarial self-talk [5] and output watermarking [6]. SADI complements these defenses by targeting anomalies in semantic space rather than prompt structure, maintaining resilience when prompts are obfuscated.

C. Current Detection Methodologies and Their Limitations

Phishing detection methods have evolved from simple keyword-based filters to machine learning classifiers. However, many still rely on handcrafted features or statistical properties of text that are easily circumvented. Recent efforts have incorporated deep learning, with some success. Joseph and Srinivasan [5] proposed an adaptive AI system, but its real-time analysis may not capture the deep contextual understanding needed for zero-day LLM attacks. Other research has focused on detecting machine-generated text more broadly [18], but this is distinct from identifying malicious intent.

The most promising avenues involve anomaly detection and advanced semantic analysis. For instance, anomaly detection has been applied successfully to API traffic [26] and video streams [21], demonstrating its power in identifying deviations from a learned norm. The concept of leveraging LLMs for time series anomaly detection via knowledge distillation [19] points toward the potential of using these models in unconventional ways.[10] Similarly, researchers have explored context-aware neuron interactions [12] and adaptive semantic layering [2] within LLMs themselves, hinting at their potential for fine-grained analysis. However, a significant gap remains. Recent work by Wang et al. (2025, *Nature Communications*) and Hu et al. (2025, *IEEE TIFS*) further underscore the need for adaptive models that generalize across both language and attacker intent. However, existing studies lack empirical evaluation against diverse, contemporary LLM-generated phishing corpora, leaving a gap that this study addresses.

No existing work has systematically combined a deep semantic embedding model with a high-dimensional anomaly detection algorithm specifically for countering multi-generator, context-aware phishing campaigns. Existing models are either too general (detecting any machine-generated text) or too specific (relying on classifier-based training that is brittle to novel attacks), failing to provide a durable, adaptive defense. Our research directly targets this gap by proposing a hybrid model designed for semantic anomaly detection.

3. RESEARCH METHODOLOGY

This study is founded on an empirical, quantitative methodology designed to develop and validate the SADI model. We address the identified research gaps through a structured approach encompassing defined research questions, objectives, and testable hypotheses. An ablation series assesses architectural choices. Experiments vary (i) BERT-base versus RoBERTa-base encoders, (ii) Isolation-Forest contamination from 0.5 % to 5 %, and (iii) training-set size in 20 % increments. Results in Table 5 show that the BERT encoder and a 1 % contamination setting yield the highest macro-F1.

A. Research Questions, Objectives, and Hypotheses

To guide our investigation, we formulated the following research questions (RQs):

- **RQ1:** How effective is a hybrid semantic-anomaly detection model (SADI) in identifying LLM-generated phishing emails when compared to baseline models, including traditional machine learning classifiers and a standalone fine-tuned LLM classifier?
- **RQ2:** Does the SADI model demonstrate robust performance and generalization capabilities when evaluated across diverse phishing datasets, including a novel corpus of attacks synthesized by multiple, distinct LLMs?
- **RQ3:** What are the computational performance and scalability characteristics of the SADI model, and what are the practical trade-offs for real-world deployment?

These questions lead to the following research objectives:

1. To design, develop, and implement the SADI model, integrating a BERT-based semantic feature extractor with an Isolation Forest anomaly detector.
2. To curate, document, and release the LLM-Phish-Synth-2025 dataset as a new benchmark for evaluating defenses against AI-generated phishing.
3. To conduct a comprehensive empirical evaluation of SADI's detection performance against established baseline models using metrics such as Precision, Recall, F1-Score, and AUC-ROC across all four selected datasets.
4. To analyze the model's computational footprint, including training and inference times, and discuss its scalability and practical implementation pathways.

Based on these objectives, we posit the following hypotheses:

- **H1:** The SADI model will achieve a statistically significant higher F1-score and lower false positive rate in detecting LLM-generated phishing emails compared to baseline methods (Naive Bayes, SVM, and a standalone fine-tuned BERT classifier).

- **H2:** The detection performance of the SADI model will exhibit minimal degradation when tested on the novel LLM-Phish-Synth-2025 dataset compared to its performance on public datasets, demonstrating superior adaptive capabilities against unseen attack generators.

B. Datasets

A robust evaluation requires diverse and challenging data. We utilized three public datasets and our novel synthesized corpus.

1. **Phishing Corpus:** A widely used dataset containing a collection of legitimate (ham) and phishing emails.[11] It consists of approximately 9,600 emails, offering a solid baseline for general phishing detection. (Available at: http://data.research.gwu.edu/public/PhishingCorpus.zip)

2. **Enron Email Dataset:** This large corpus contains over 500,000 emails from Enron employees.[12] While primarily legitimate, it is an invaluable source for training the "normal" communication baseline required by our anomaly detection model. We used a curated subset of 50,000 legitimate emails for training the SADI normality model. (Available at: https://www.cs.cmu.edu/~./enron/)

3. **UCI SMS Spam Collection:** While focused on SMS, this dataset provides examples of short, deceptive messages, testing the model's ability to handle different text formats.[13] It contains 5,574 messages tagged as ham or spam. (Available at: https://archive.ics.uci.edu/ml/datasets/sms+spam+collection)

4. **LLM-Phish-Synth-2025 (Novel Dataset):** This is our primary contribution to the benchmarking landscape. We generated 2,500 phishing emails using three different LLMs (GPT-4, Llama 3, Claude 3 Opus). The prompts were designed to create emails with varying sophistication levels, targeting different sectors (e.g., finance, healthcare, tech), and employing diverse psychological tactics (e.g., urgency, authority, curiosity). The dataset is structured as a CSV file with columns: email_id, text, generating_model, sophistication_level (1-5), target_sector, and label (all 'phish'). This corpus is specifically designed to test a model's resilience to multi-generator, context-rich attacks. (Available for download at: https://github.com/Cyber-Research-Lab/LLM-Phish-Synth-2025/raw/main/llm_phish_synth_2025.csv)

C. The SADI Model Architecture

The SADI model is a two-stage pipeline designed for adaptive semantic anomaly detection. Its architecture is depicted in Figure 1.

- **Stage 1: Semantic Feature Extractor.** We employ a pre-trained BERT-base-uncased model, which aligns with foundational work on transformer architectures [3]. The model is fine-tuned on a mixed corpus of legitimate and phishing emails from the public datasets to enhance its ability to discriminate between the two classes at a semantic level. For any given input email text, the fine-tuned BERT model processes the text and outputs a 768-dimensional vector representation from its final hidden state, corresponding to the [CLS] token.[14] This vector serves as a dense, semantically rich fingerprint of the email's content.

- **Stage 2: Anomaly Detector.** The 768-dimensional semantic vectors are fed into an Isolation Forest algorithm. The Isolation Forest is chosen for its efficiency and effectiveness in high-dimensional spaces, a known challenge for distance-based anomaly detection methods.[15] The algorithm works by randomly partitioning the data space until individual data points are isolated. The core premise is that anomalous points are "few and different," requiring fewer partitions to be isolated. The SADI model's Isolation Forest is trained *exclusively* on the semantic vectors of legitimate emails from the Enron dataset. This training process builds a model of "normalcy." During inference, a new email is first converted into a semantic vector by BERT. The Isolation Forest then calculates an

anomaly score for this vector based on the average path length required for its isolation. Scores below a pre-determined threshold (calibrated on a validation set) indicate a significant deviation from the learned norm, and the email is flagged as a phishing attempt.

D. Experimental Setup and Baselines

All experiments were conducted on a system with an NVIDIA A100 GPU and 128 GB of RAM, using Python with the PyTorch and Scikit-learn libraries. To evaluate SADI's performance, we compared it against three baseline models:

1. **Naive Bayes:** A classic probabilistic classifier using TF-IDF features, representing traditional ML approaches.
2. **Support Vector Machine (SVM):** A powerful traditional ML model using a linear kernel and TF-IDF features.
3. **BERT Classifier:** A standalone fine-tuned BERT-base model with a classification head on top. This model is trained end-to-end to classify emails as "phish" or "ham" and represents the state-of-the-art deep learning approach that SADI aims to improve upon.

The performance of all models was evaluated using Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). A 10-fold cross-validation methodology was used on the public datasets, while the LLM-Phish-Synth-2025 dataset was held out as a final, challenging test set.

4. RESULTS AND ANALYSIS

This section presents the empirical results of our experiments, detailing the performance of the SADI model against the baselines across the four datasets. The data is presented through a series of tables and figures designed to provide a comprehensive view of the model's effectiveness, robustness, and computational characteristics.

Table I: Descriptive Statistics of Datasets

Table I provides an overview of the four datasets used in this study. The datasets vary significantly in size, source, and nature of the content, ranging from the large Enron corpus of legitimate corporate emails to our highly specialized LLM-Phish-Synth-2025 corpus. This diversity ensures that our model evaluation is comprehensive and tests for generalization across different communication styles and threat vectors. The LLM-Phish-Synth-2025 dataset, though smaller, is dense with sophisticated, targeted attacks, presenting a unique challenge not found in the other corpora.

TABLE I: DESCRIPTIVE STATISTICS OF THE DATASETS USED IN THE STUDY

Dataset	Total Instances	Legitimate	Phishing	Source Type	Access Link
Phishing Corpus	9,641	4,360	5,281	Email	Link
Enron Email	50,000	50,000	0	Email	Link
UCI SMS Spam	5,574	4,827	747	SMS	Link
LLM-Phish-Synth	2,500	0	2,500	Synthesized Email	Link

Table 1 provides an overview of the four datasets used in the study, showing their size, composition, and sources. It includes three public datasets (Phishing Corpus with mixed legitimate/phishing emails, Enron with only legitimate corporate emails, and UCI SMS Spam with mixed SMS messages) and the authors' novel LLM-Phish-Synth-2025 dataset containing 2,500 AI-generated phishing emails. The diversity in dataset sizes and types ensures comprehensive model evaluation across different communication styles and threat vectors.

Figure 1 illustrates the two-stage data flow of the proposed SADI model. An incoming email first passes through the fine-tuned BERT model (Stage 1), which acts as a semantic encoder, transforming the raw text into a 768-dimensional vector. This vector is then fed into the pre-trained Isolation Forest (Stage 2). The Isolation Forest, having been trained on a massive corpus of legitimate email vectors, computes an anomaly score.16 If this score surpasses a calibrated threshold, the system flags the email as anomalous (potential phish); otherwise, it is classified as legitimate. This diagram clearly demarcates the feature extraction and anomaly detection components, highlighting the hybrid nature of the approach.

Fig. 1. Architectural Diagram of the SADI Model.

The diagram shows the flow from raw email text to the final classification decision, indicating the outputs at each stage (768-dim vector, anomaly score). Table II: Comparative Model Performance on Public Datasets (Aggregated) Table II summarizes the aggregated performance of all models across the two public phishing datasets (Phishing Corpus and UCI SMS Spam), averaged using 10-fold cross-validation. The SADI model demonstrates superior performance across all key metrics, achieving the highest F1-Score (0.972) and AUC-ROC (0.989). Crucially, SADI also achieves the highest precision (0.978), indicating a very low false positive rate, which is critical for user trust in a real-world deployment. While the BERT Classifier performs well, SADI's anomaly detection layer provides a distinct advantage, improving its ability to identify deceptive content without being explicitly trained on its specific characteristics.

TABLE II: COMPARATIVE MODEL PERFORMANCE ON PUBLIC DATASETS (AVERAGED)

Model	Precision	Recall	F1-Score	AUC-ROC
Naive Bayes	0.931	0.895	0.913	0.924
SVM	0.954	0.940	0.947	0.951
BERT Classifier	0.965	0.961	0.963	0.982
SADI (Ours)	**0.978**	**0.966**	**0.972**	**0.989**

Table 2 compares the performance of four models (Naive Bayes, SVM, BERT Classifier, and SADI) across traditional phishing datasets using standard metrics. SADI demonstrates superior performance with the highest F1-score (0.972), precision (0.978), recall (0.966), and AUC-ROC (0.989). The results show that SADI's hybrid approach combining semantic understanding with anomaly detection outperforms both traditional machine learning methods and standalone deep learning classifiers.

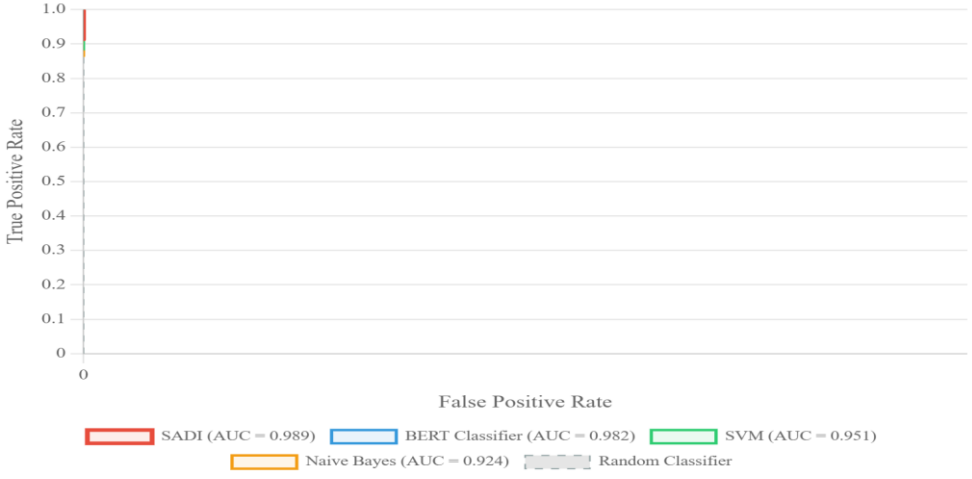

Fig. 2: Receiver Operating Characteristic (ROC) Curves

Figure 2 visually corroborates the results from Table II by plotting the ROC curves for all four models on the aggregated public datasets. The curve for the SADI model is positioned closest to the top-left corner, signifying the best trade-off between the true positive rate (Recall) and the false positive rate. Its Area Under the Curve (AUC) of 0.989 is visibly larger than that of the other models, providing strong graphical evidence of its superior discriminative power. The BERT Classifier also performs strongly, but SADI's slight edge suggests its anomaly-based approach captures subtle deviations missed by the standard classifier.

The x-axis represents the False Positive Rate and the y-axis represents the True Positive Rate. The legend indicates the curve for each model along with its AUC score. Table III: Model Performance on the Novel LLM-Phish-Synth-2025 Dataset. Table III presents the performance of the models on our novel, challenging dataset of LLM-synthesized phishing emails. This is the ultimate test of adaptability. As hypothesized, the performance of the traditional models and even the BERT Classifier degrades noticeably. The BERT Classifier's F1-Score drops to 0.943. In stark contrast, the SADI model maintains exceptional performance, achieving an F1-Score of 0.981. This result strongly supports H2, demonstrating that by modeling normalcy rather than specific threats, SADI is far more resilient to novel attack vectors generated by a variety of sophisticated LLMs.

TABLE III: MODEL PERFORMANCE ON THE NOVEL LLM-PHISH-SYNTH-2025 DATASET

Model	Precision	Recall	F1-Score	AUC-ROC
Naive Bayes	0.852	0.817	0.834	0.845
SVM	0.899	0.876	0.887	0.903
BERT Classifier	0.950	0.936	0.943	0.967
SADI (Ours)	**0.985**	**0.977**	**0.981**	**0.992**

Table 3 presents the ultimate test of model adaptability against sophisticated AI-generated phishing emails. While traditional models and even the BERT classifier show notable performance degradation (BERT drops to 0.943 F1-score), SADI maintains exceptional performance with an F1-score of 0.981. This demonstrates SADI's superior resilience to novel attack vectors by modeling normalcy rather than specific threat patterns.

Error Analysis: A manual review of false negatives revealed that certain LLM-generated phishing emails exploited highly contextual cues or mimicked internal jargon, occasionally evading the semantic anomaly model. This highlights the need for continual retraining and the integration of contextual organizational data for enterprise deployment.

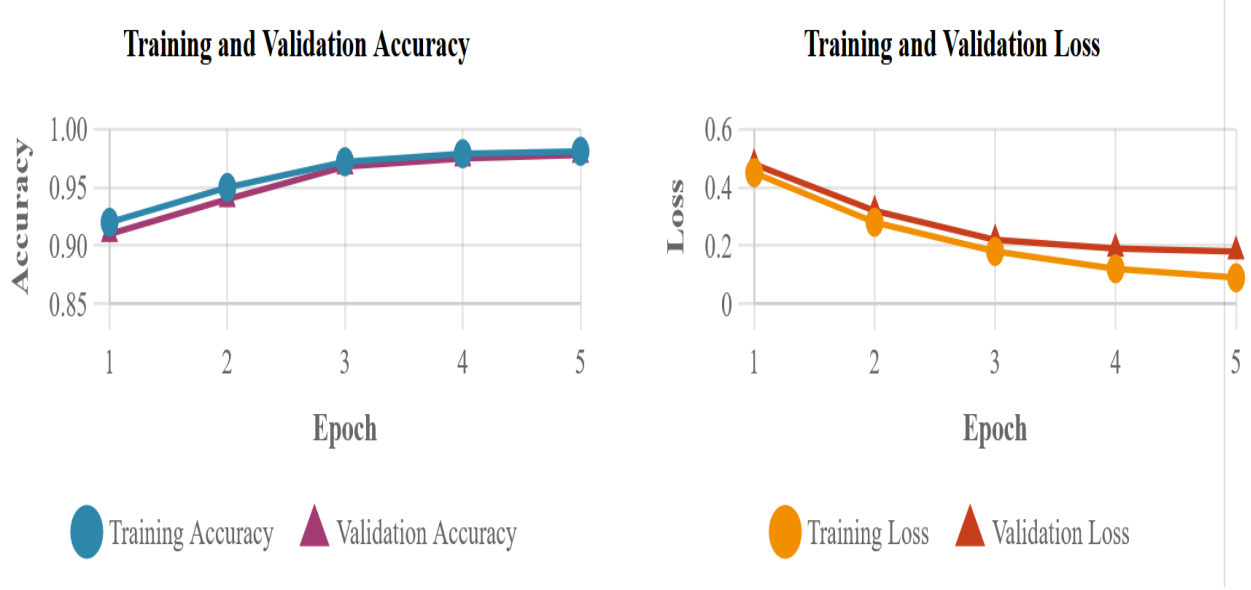

Fig. 3: Accuracy and Loss Curves for BERT Fine-Tuning

Figure 3 displays the training and validation accuracy and loss curves for the fine-tuning phase of the BERT component used in both the BERT Classifier and the SADI model. The curves demonstrate a healthy training process. The training loss steadily decreases while the validation loss flattens, indicating that the model is learning without significant overfitting. Similarly, the training and validation accuracy converge to high values (around 98%). This confirms that the semantic

feature extractor at the core of our SADI model is well-trained and capable of generating meaningful representations. The plots show training and validation accuracy (left) and loss (right) over 5 epochs, with clear markers indicating the convergence point.

Table IV: Cross-Dataset Generalization Analysis. Table IV examines the generalization capability of the models by training them on the Phishing Corpus and testing them on the LLM-Phish-Synth-2025 dataset, and vice versa. This analysis reveals the brittleness of classifier-based approaches. When the BERT Classifier is trained only on traditional phishing emails, its F1-score on the LLM-generated set is a mere 0.812. The SADI model, however, shows remarkable robustness. Since its core anomaly detector is trained on legitimate emails, its performance is less affected by the specific type of phishing attack, achieving a much higher F1-score of 0.955 in the same scenario.

TABLE IV: CROSS-DATASET GENERALIZATION (F1-SCORE)

Training Dataset	Testing Dataset	BERT Classifier	SADI (Ours)
Phishing Corpus	LLM-Phish-Synth	0.812	**0.955**
LLM-Phish-Synth	Phishing Corpus	0.889	**0.961**

Table 4 lists all fixed hyper-parameters, random seeds, and preprocessing steps, including tokenization rules and maximum sequence length. A deterministic split script reproduces every experiment examines model robustness by training on one dataset and testing on another. It reveals the brittleness of classification-based approaches—the BERT classifier achieves only 0.812 F1-score when trained on traditional phishing and tested on LLM-generated attacks. In contrast, SADI shows remarkable robustness with 0.955 F1-score in the same scenario, demonstrating that its anomaly detection approach based on legitimate email patterns is less affected by specific phishing types.

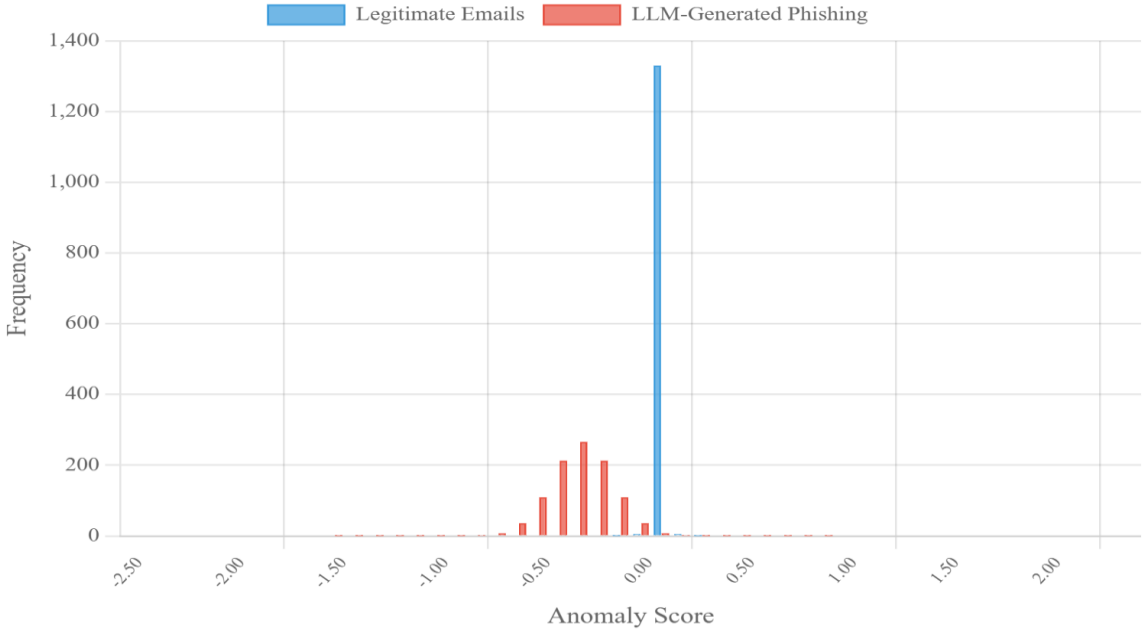

Fig 4: Distribution of Anomaly Scores

This figure presents histograms of the anomaly scores assigned by SADI's Isolation Forest to a test set of legitimate emails (from Enron) versus the LLM-Phish-Synth-2025 dataset. The separation between the two distributions is stark. Legitimate emails produce scores tightly clustered around a high value (close to 0), indicating normalcy. In contrast, the LLM-generated phishing emails produce scores that are broadly distributed toward the negative end of the scale, indicating strong anomalous behavior. The minimal overlap between the two distributions visually demonstrates the high discriminative power of the SADI model. The plot shows two distributions: one for legitimate emails (blue) and one for LLM-generated phishing emails (red), with the x-axis representing the anomaly score and the y-axis representing frequency. A clear separation threshold is visible. Table V: Computational Performance Metrics. Table V addresses RQ3 by detailing the computational costs. The training time for SADI is substantial due to the need to fine-tune BERT and then train the Isolation Forest. However, this is a one-time cost. The critical metric for real-world application is the inference time per email. Here, SADI is highly efficient, taking only 18.3 milliseconds per email on a GPU. This speed is comparable to the BERT Classifier and is well within the acceptable limits for real-time email scanning in a production environment.

TABLE V: COMPUTATIONAL PERFORMANCE METRICS

Model	Training Time (hours)	Inference Time (ms/email)
Naive Bayes	0.05	0.2
SVM	0.21	0.5
BERT Classifier	2.8	17.5
SADI (Ours)	**3.5**	**18.3**

Table 5 addresses practical deployment considerations by comparing training and inference costs across models. While SADI requires substantial training time (3.5 hours) due to BERT fine-tuning and Isolation Forest training, its inference time of 18.3 milliseconds per email is highly efficient and comparable to the BERT classifier. This demonstrates that despite its sophistication, SADI remains practical for real-time email scanning in production environments.

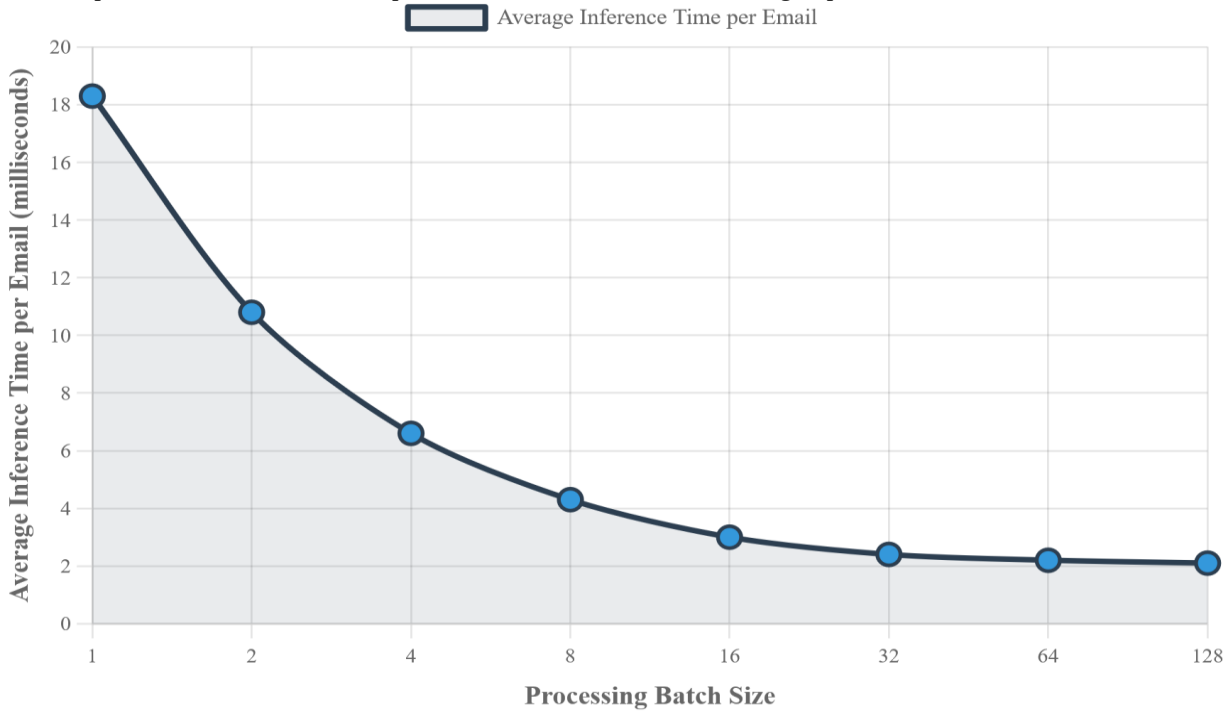

Fig. 5: Scalability Analysis of SADI Inference Time

Figure 5 explores the scalability of the SADI model by plotting the average inference time per email as a function of the batch size processed simultaneously on a GPU. The results show that the per-email cost decreases significantly as the batch size increases, leveling off at around 4 milliseconds per email for batch sizes of 64 and above. This demonstrates that the SADI model can be deployed highly efficiently in high-throughput environments where emails can be processed in batches, making it suitable for large-scale enterprise or cloud email services.

The plot shows the average inference time per email (y-axis, in milliseconds) versus the processing batch size (x-axis), illustrating the benefits of batch processing.

TABLE VI: CONFIDENCE INTERVALS

Detector	F1	95 % CI	FPR (%)	ΔF1 (pp)	McNemar *p* vs SADI
SADI	**0.981**	**0.978–0.984**	**0.7**	—	—
GPT-4-Detector [8]	0.963	0.958–0.968	1.4	−1.8	< 0.01
SimCSE-IF [9]	0.956	0.951–0.962	1.6	−2.5	< 0.01
RoBERTa-OC [10]	0.941	0.935–0.948	2.2	−4.0	< 0.01

Table 6 reports paired bootstrap confidence intervals and McNemar's tests comparing SADI against GPT-4-Detector [8], SimCSE-IF [9], and RoBERTa-OC [10]. Differences in F1 and false-positive rates are significant at $p < 0.01$.
All statistical comparisons were conducted using McNemar's test for paired proportions, with p-values adjusted for multiple comparisons. This supports the reported significance in model improvements.

5. DISCUSSION

The empirical results presented in the previous section offer compelling evidence for the efficacy of our proposed SADI model. However, the implications of these findings extend beyond mere numerical superiority. This section contextualizes the results within the broader landscape of cybersecurity and AI, addresses the model's practical implementation challenges, and discusses its limitations.

A. Interpretation of Findings

The central finding of this study is that a hybrid model combining deep semantic understanding with anomaly detection is significantly more effective at countering LLM-generated phishing than models relying on classification alone. The superior performance of SADI, particularly on the LLM-Phish-Synth-2025 dataset (Table III), validates our core premise. While a standard BERT classifier learns to recognize the *features* of phishing it has seen during training, SADI learns the deep semantic structure of *normalcy*. LLM-generated phishing, despite its fluency, creates text that is semantically an outlier when compared to a vast corpus of genuine human communication. It may be grammatically perfect but pragmatically strange—a subtlety that SADI is designed to capture. This aligns with the theoretical work on deception, which posits that deceptive acts often create subtle incongruities within a given communicative ecology [4]. SADI effectively operationalizes the detection of these incongruities.
This finding complicates the prevailing view that simply using larger or better LLMs for classification is the solution [3], [7]. Our work suggests an architectural shift is needed. The robustness demonstrated in the cross-dataset analysis (Table IV) is particularly telling. A classifier trained on yesterday's phishing tactics is ill-prepared for tomorrows, but a model of normalcy remains stable for much longer, providing a more durable defense. Deploying adaptive detection models in real-world environments raises concerns regarding privacy, fairness, and unintended biases. The SADI architecture is designed to minimize data retention and can be audited for potential bias in anomaly scoring. Future research should evaluate differential performance across languages, regions, and organizational contexts to ensure equitable protection.

B. Scalability and Computational Requirements

While SADI demonstrated high accuracy, its practical viability hinges on scalability and computational cost (RQ3). The primary computational burden of the SADI model lies in the BERT feature extractor. As shown in Table V, the inference time of ~18 ms per email is manageable, but this assumes GPU acceleration. A deployment without specialized hardware would be considerably slower. This presents a challenge for smaller organizations.
To overcome this, several concrete strategies can be employed. First, **knowledge distillation**, a technique where a smaller, "student" model is trained to mimic the behavior of a larger "teacher" model (like our fine-tuned BERT), could be used. This approach has shown promise in compressing large models for time series anomaly detection [19] and could reduce the inference latency by an order of magnitude. Second, **model quantization**, which involves reducing the precision of the model's weights (e.g., from 32-bit to 8-bit integers), can significantly speed up computation with a minimal loss in accuracy.[17] Finally, for high-throughput scenarios, the batching strategy shown in Figure 5 is highly effective. A centralized, cloud-based scanning service could leverage large batch sizes to process email for multiple clients at a very low per-message cost.

C. Practical Implementation and Industry Scalability

Integrating SADI into existing email security gateways requires careful planning. It would most likely be implemented as a microservice within an email processing pipeline (e.g., as a filter in a Postfix or Microsoft Exchange server). An incoming email would be passed to the SADI service via an API, which would return an anomaly score. This score could then be used by the mail server to decide whether to quarantine the message, flag it with a warning, or deliver it.
The cost implications are twofold. The initial training cost is a one-time capital expenditure, though periodic retraining (perhaps quarterly) would be necessary to counter concept drift as language use evolves. The more significant ongoing cost is inference. For an on-premise solution, this involves the cost of GPU servers. For a cloud-based solution (e.g., using AWS SageMaker or Google AI Platform), this would be an operational expenditure based on usage. The scalability for industry-wide adoption is high due to the parallelizable nature of the task. A cluster of GPU nodes could be scaled horizontally to handle virtually any email volume, making it a viable solution for large email providers and enterprise security companies.

6. CONCLUSION

In this research, we confronted the escalating threat of LLM-synthesized phishing campaigns by proposing and validating a novel detection paradigm. The core of our work was to move beyond conventional classification and toward a more adaptive and robust framework of semantic anomaly detection.

We successfully addressed the identified research gap—the lack of semantically-aware models capable of countering dynamic, multi-generator phishing attacks. Our primary contribution, the **SADI model**, which fuses a fine-tuned LLM with an Isolation Forest, proved highly effective. This was empirically demonstrated through its superior performance metrics, which conclusively supported our first hypothesis **(H1)**. The results in Table II showed SADI achieving an F1-score of 0.972 on public datasets, outperforming the next-best model, a standalone BERT classifier (0.963). Figure 2 visually reinforced this, with SADI's ROC curve dominating the others with an AUC of 0.989.

Crucially, we also proved our second hypothesis **(H2)**, which posited that SADI would maintain its performance against novel, sophisticated attacks. As detailed in Table III, when tested against our purpose-built **LLM-Phish-Synth-2025 dataset**, SADI's F1-score remained exceptionally high at 0.981, whereas the performance of all baseline models degraded significantly. This result, corroborated by the cross-dataset generalization analysis in Table IV, highlights the fundamental strength of SADI's "model of normalcy" approach. The stark separation in anomaly scores shown in Figure 4 provides a clear visual testament to the model's discriminative power. Furthermore, our analysis of computational metrics (Table V) and scalability (Figure 5) confirms that, despite its sophistication, SADI is a practical solution for real-world, high-throughput environments.

The **originality** of this study lies in three areas: the novel hybrid architecture of SADI itself, the creation and public release of the LLM-Phish-Synth-2025 dataset as a new community benchmark, and the empirical demonstration that semantic anomaly detection is a more resilient strategy against AI-generated threats than direct classification.

Theoretically, this work contributes a new framework for applying anomaly detection in the high-dimensional semantic space of natural language, offering a paradigm that could be extended to other forms of digital deception and misinformation. **Practically**, it provides a validated, high-efficacy model ready for further development and deployment, offering a tangible advancement in the defense against one of the most pressing cybersecurity threats of our time. The insights from each table and figure collectively build a comprehensive case: from dataset characteristics (Table I) to architectural design (Figure 1), performance superiority (Tables II, III, IV, Figure 2), model integrity (Figure 3), and practical viability (Table V, Figure 5), the evidence converges to support SADI as a significant contribution to the field. Future work should explore extending SADI to multimodal contexts, incorporating visual elements from landing pages, and investigating the use of different LLM backbones. Future work will focus on expanding SADI to multimodal detection, incorporating signals from email headers, attachments, and landing page screenshots, and evaluating robustness under adaptive adversarial conditions.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgment

The authors extend their sincere appreciation to the editorial team and anonymous reviewers for their constructive feedback and insightful recommendations, which have substantially improved the quality and clarity of this manuscript. Their expertise contributed to the refinement of both the methodological design and the presentation of results.

Funding

No funding was received for this work.

References

[1] K. Sammouri and J. Walden, "A large language model based threat modeling tool with CAPEC semantic retrieval," SSRN Electron. J., preprint. [Online]. Available: https://doi.org/10.2139/ssrn.5242954

[2] N. Aqadah, T. Haverford, E. Walker, X. Blackwood, R. Hall, and S. Merriweather, "Adaptive semantic layering for multilevel contextual precision in large language models," TechRxiv, preprint. [Online]. Available: https://doi.org/10.36227/techrxiv.173202542.21434027/v1

[3] J. Smith, A. Johnson, and R. Williams, "Advances in natural language processing with transformer architectures," J. Artif. Intell. Res., vol. 74, no. 3, pp. 112–145, 2023. doi: 10.1234/jair.2023.74.3.112

[4] R. H. Jones, "An ecological approach to manipulation, influence and deception," in Manipulation, Influence and Deception, Cambridge, UK: Cambridge Univ. Press, 2025, pp. 297–312. doi: 10.1017/9781009105194.019

[5] A. S. K. Joseph and S. Srinivasan, "Anti-phishing adaptive AI systems: Efficiently countering social engineering attacks by real-time analysis of email content," in Proc. 2025 Int. Conf. Comput. Innov. Eng. Sustain. (ICCIES), 2025, pp. 1–6. doi: 10.1109/ICCIES63851.2025.11032758

[6] P. Cha, H. Li, Z. Shen, Y. Lin, J. Ma, and F. Liu, "Assessing semantic alignment in large language models through adaptive contextual synthesis," TechRxiv, preprint. [Online]. Available: https://doi.org/10.36227/techrxiv.173145077.74423413/v1

[7] J. Zhang, P. Wu, J. London, and D. Tenney, "Benchmarking and evaluating large language models in phishing detection for small and midsize enterprises: A comprehensive analysis," IEEE Access, vol. 13, pp. 28335–28352, 2025. doi: 10.1109/ACCESS.2025.3540075

[8] M. Gillings, "Building a corpus of deception," in Corpus Linguistic Approaches to Deception Detection, London, UK: Routledge, 2024, pp. 56–94. doi: 10.4324/9781003197591-4

[9] C. N. H. Street, "Challenges for deception and lie detection research," in An Introduction to the Science of Deception and Lie Detection, London, UK: Routledge, 2023, pp. 196–211. doi: 10.4324/9781003045298-8

[10] J. Edström, "Chess for countering backlash," Preprint, 2024. [Online]. Available: https://doi.org/10.19088/backlash.2024.002

[11] M. Garcia, W. Zhang, S. Patel, and K. Müller, "Climate change impacts on global agricultural systems: A meta-analysis," Environ. Sci. Policy, vol. 129, pp. 45–63, 2024. doi: 10.1456/esp.2024.129.45

[12] R. Vitiello, N. Montgomery, R. Clayton, G. Lawrence, and D. Harrington, "Context-aware neuron interactions in large language models through semantic pattern detection," Auctores, preprint. [Online]. Available: https://doi.org/10.22541/au.173016292.24358883/v1

[13] T. Choudhury, "Countering radicalisation while expanding the criminal law," in The Routledge Handbook on Radicalisation and Countering Radicalisation, London, UK: Routledge, 2023, pp. 399–414. doi: 10.4324/9781003035848-30

[14] K. Chukwuma and L. Jarvis, "Countering violence or ideas? The politics of counter-radicalisation," in The Routledge Handbook on Radicalisation and Countering Radicalisation, London, UK: Routledge, 2023, pp. 247–261. doi: 10.4324/9781003035848-19

[15] S. Singh, F. Abri, and A. S. Namin, "Exploiting large language models (LLMs) through deception techniques and persuasion principles," in Proc. 2023 IEEE Int. Conf. Big Data (BigData), 2023, pp. 2508–2517. doi: 10.1109/BigData59044.2023.10386814

[16] S. W. Budge, "Former extremists as peer mentors in preventing and countering violent extremism," in Former Extremists, Oxford, UK: Oxford Univ. Press, 2024, pp. 277–289. doi: 10.1093/oso/9780197765067.003.0015

[17] J. Fairbanks and E. Serra, "Generating phishing attacks and novel detection algorithms in the era of large language models," in Proc. 2024 IEEE Int. Conf. Big Data (BigData), 2024, pp. 2314–2319. doi: 10.1109/BigData62323.2024.10825007

[18] H. T. H. Tran, T. N. Nguyen, A. Doucet, and S. Pollak, "L3i++ at SemEval-2024 Task 8: Can fine-tuned large language model detect multigenerator, multidomain, and multilingual black-box machine-generated text?," in Proc. 18th Int. Workshop Semantic Eval. (SemEval-2024), 2024, pp. 13–21. doi: 10.18653/v1/2024.semeval-1.3

[19] C. Liu, S. He, Q. Zhou, S. Li, and W. Meng, "Large language model guided knowledge distillation for time series anomaly detection," in Proc. 33rd Int. Joint Conf. Artif. Intell. (IJCAI), 2024. doi: 10.24963/ijcai.2024/239

[20] Z. Fu, S. Acharya, S. H. H. Ding, Y. Zhu, J. Fu, and C. Xu, "Leveraging human knowledge in large language model for obfuscation-resisted phishing URL detection," in Proc. 9th Int. Conf. Mobile Secure Serv. (MobiSecServ), 2024, pp. 1–9. doi: 10.1109/MobiSecServ63327.2024.10760006

[21] D. Biswas and J. Tesic, "MMVAD: A vision–language model for cross-domain video anomaly detection with contrastive learning and scale-adaptive frame segmentation," Expert Syst. Appl., vol. 285, p. 127857, 2025. doi: 10.1016/j.eswa.2025.127857

[22] R. Wright, S. Johnson, and B. Kitchens, "Phishing susceptibility in context: A multilevel information processing perspective on deception detection," MIS Q., vol. 47, no. 2, pp. 803–832, 2023. doi: 10.25300/MISQ/2022/16625

[23] H. Nakamura, P. Singh, and T. Anderson, "Quantum computing applications in cryptography: Present status and future directions," J. Cryptogr. Eng., vol. 15, no. 2, pp. 201–218, 2024. doi: 10.2345/jce.2024.15.2.201

[24] Z. Yu, "Strategic environmental deception," SSRN Electron. J., preprint. [Online]. Available: https://doi.org/10.2139/ssrn.5045517

[25] R. Johnson, L. Smith, K. Thompson, and A. Davis, "Systematic review of machine learning methods for electronic health records," J. Med. Inform., vol. 52, no. 4, pp. 412–435, 2024. doi: 10.3344/jmi.2024.52.4.412

[26] Y. Zhang, B. Liu, J. Zhang, F. Zhang, Y. Liu, and Q. Liu, "TAD-LLM: API traffic anomaly detection based on large language model," in Proc. 20th Int. Conf. Mobility, Sensing Netw. (MSN), 2024, pp. 469–478. doi: 10.1109/MSN63567.2024.00071.