

A Lightweight Agentic AI Framework with DeepSeek-R1 for Adaptive Phishing URL Detection

Akshat Gaurav¹, Varsha Arya², Amiya Nayak³, Kwok Tai Chui², Brij B. Gupta^{1,4*}

¹Department of Computer Science and Information Engineering, Asia University, Taichung 413, Taiwan

²Hong Kong Metropolitan University, Hong Kong SAR, China

³School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada

⁴VIZJA University, Warsaw, Poland

akshat.gaurav@ieee.org, varshaarya2108@gmail.com, nayak@uottawa.ca, jktchui@hkmu.edu.hk, bbgupta@asia.edu.tw

Abstract

Phishing URLs remain a major cybersecurity threat because their development is constantly changing and becoming more deceptive. This study presented an agentic adaptive AI framework that utilizes a large language model as a reasoning agent operating multiple external tools instead of performing a classification, to detect phishing URLs. An lightweight tabular classifier with just 9,662 trainable parameters delivers predictions very efficiently, and explains the relevance of features in terms of attack similarity based on SHAP-based feature attribution and episodic memory retrieval. The agent combines these outputs from the tools to generate structured explanations and security recommendations. Experiments demonstrate strong performance, with accuracy up to 95.6% and AUC values above 0.99.

Introduction

Phishing remains one of the most pervasive and evolving threats in cyberspace, with attackers continually refining techniques to lure users and evade detection through carefully crafted URLs and web content (Kritika 2025; Tamal et al. 2024; Vajrobol, Gupta, and Gaurav 2024; Ishtaiwi et al. 2024). The rapid growth of phishing campaigns, the increasing sophistication of deceptive URLs, and the high stakes of credential theft and data exfiltration motivate intensified research on robust detection methods (Nagy et al. 2023; Shaukat et al. 2020; Gupta et al. 2024a,b). In this context, phishing URL detection has emerged as a critical security problem that benefits from fast, scalable, and privacy-preserving solutions deployed close to users and endpoints (Nagy et al. 2023; Gragnaniello et al. 2024; Almomani et al. 2013). The literature demonstrates that phishing detection often leverages sequential, parallel, or hybrid machine learning (ML) approaches to model the temporal evolution and contextual cues of phishing URLs, including lexical, host-based, and content features (Nagy et al. 2023; Qiu et al. 2020). However, these data-driven methods face persistent challenges that limit their practical utility in real-world deployments (Nagy et al. 2023; Shaukat et al. 2020).

Recent surveys ML/DL-based security solutions and highlights the key limitations of current approaches for phishing and broader cyber threats. Notably, deep learning-based detectors can suffer from distribution drift, data scarcity in emerging phishing campaigns, and vulnerability to adversarial manipulation, especially when models are deployed in adversarial settings or trained on outdated data (Ali et al. 2023, 2022). Adversaries increasingly exploit weaknesses in DL systems, including zero-day conditions and evasion tactics, which can degrade detection performance and erode trust in automated defenses (Ali et al. 2022; Shaukat et al. 2020). Moreover, even when ML/DL models achieve high accuracy in controlled experiments, their interpretability and explainability remain significant concerns for security operators who require understandable justifications for alerts and decisions (Zhang et al. 2022; Neupane et al. 2022), (Capuano et al. 2022; Hoenig et al. 2024). These explainability concerns intersect with the need for robust defense against intelligent attackers who adapt to model weaknesses (Khan et al. 2023; Shaukat et al. 2020).

These observations motivate a shift toward agentic tool-augmented reasoning for phishing URL detection. The idea of agent-based security has been explored in multi-agent cyber-physical systems, where coordinated agents collaborate to detect, respond to, and adapt to threats in dynamic environments (Owoputi and Ray 2022; Kalinin, Zavadskii, and Busygin 2023). Related work emphasizes the potential of leveraging external tools, threat intelligence, and formalized attack frameworks to inform decision making and simulation-based evaluation of defenses in security systems, including CPS and networked environments (Oh, Kim, and Park 2024). Such agentic, tool-augmented capabilities align well with the broader trend toward memory-empowered, adaptive reasoning that can operate under resource constraints while remaining transparent and controllable to human operators (Zhang et al. 2022; Neupane et al. 2022; Hoenig et al. 2024; Capuano et al. 2022).

Contribution

This paper proposes an agentic lightweight AI framework for phishing URL detection, where the large language model is not directly predicting but is orchestrating classification, feature-analysis and episodic-memory. The system is capa-

*Corresponding Author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ble to perform high detection with slight performance overhead allowing adaptive and deployable phishing defence in constrained environments.

Organization

The remainder of this paper is organized as follows. Section 2 describes the details of state-of-the-art models. Section 3 presents the proposed lightweight agentic AI framework, including the classifier architecture, episodic memory module, and LLM-based reasoning mechanism. Section 4 discusses performance evaluation, including comparative analysis with existing models. Finally, Section 5 concludes the paper.

Related Work

Phishing Detection Models

Phishing URL detection has evolved from handcrafted lexical features to DL models that analyze URL sequences, host traits, and page content (Varshney et al. 2024; Putra et al. 2024). Character-level CNNs and multimodal features further improve accuracy by capturing URL syntax and context (Aljofey et al. 2020; Yang, Zhao, and Zeng 2019). Yet DL detectors often require large labeled datasets and struggle with drift and zero-day campaigns, as highlighted by reviews (Wilk-Jakubowski et al. 2025; Kyaw, Gutierrez, and Ghobakhlou 2024). Transformer-based detectors, including PhishTransformer and BERT-based features, show promise beyond traditional features (Asiri, Xiao, and Li 2023; Elsadig et al. 2022). Vajrobal, Gupta, and Gaurav (2024) proposed mutual information based phishing detection model. Ishtaiwi et al. (2024) proposed machine learning based phishing detection models.

Lightweight ML/DL Models

Lightweight phishing detectors include DL-driven lightweight sensors (Wei et al. 2019), and edge-AI work enables on-device inference on constrained hardware (Kim et al. 2023), supporting privacy-preserving, real-time detection (Wei et al. 2019; Kim et al. 2023).

LLM-based Cybersecurity Approaches Transformer-based phishing detection is increasingly explored: PhishTransformer (Asiri, Xiao, and Li 2023) and BERT-based features (Elsadig et al. 2022) illustrate the value of contextual representations, while reviews summarize the rising role of DL and Transformer models in security (Wilk-Jakubowski et al. 2025; Kyaw, Gutierrez, and Ghobakhlou 2024). Explainability is a key concern for security operators, motivating XAI work (Zhang et al. 2022; Neupane et al. 2022; Hoenig et al. 2024).

Gap and Agentic AI Framework

Despite progress, many detectors are single-shot; life-long learning addresses drift and cross-domain signals (Ejaz, Mian, and Manzoor 2023), and memory/graph-based reasoning architectures support cross-domain resilience (Kalinin, Zavadskii, and Busygin 2023; Owoputi and Ray 2022; Wagner, Damodaran, and Reavey 2025). Tool-based

reasoning with external intel and MITRE ATT&CK guides defense planning (Oh, Kim, and Park 2024), complemented by XAI for transparency (Zhang et al. 2022; Hoenig et al. 2024). An agentic framework combining edge classification, memory, and tool-based reasoning can address evolving phishing while preserving privacy (Kalinin, Zavadskii, and Busygin 2023; Owoputi and Ray 2022; Oh, Kim, and Park 2024; Ejaz, Mian, and Manzoor 2023; Elsadig et al. 2022; Asiri, Xiao, and Li 2023).

Proposed Approach

The framework for agentic phishing URL detection was presented in Figure 1. In this architecture, rather than being a direct classifier, the LLM serves (DeepSeek-R1) the role of an intelligent agent and orchestrator, reasoning effectively by interacting with multiple external tools. The model accepts inputs that are related to raw URL which are encoded via a lightweight tabular classification model constituted of feature-specific embedding layers followed by a small stack of linear and ReLU layers resulting in only 9662 trainable parameters. This classifier gives the output label and probability score. The LLM agent subsequently invokes three tools in succession: (i) the classifier tool to produce prediction and confidence, (ii) the SHAP explanation tool to obtain the most important features affecting the prediction, and (iii) an episodic attack memory module, implemented using retrieval-based similarity search, that compares the input to previously seen phishing and non-phishing samples. The LLM agent will be guided by a prompt. And using this prompt, the LLM agent will combine the outputs of the tools. And then, it will communicate what the classifier is doing, whether it is similar to past attacks, whether it is leveraging similar features or novel features and accordingly generate a short and succinct security recommendation. The suggested model establishes a simultaneous yet improved agentic workflow by decoupling the processes of prediction, explanation, memory, and reasoning, thereby increasing robustness, interpretability, and decision support for real-time phishing URL detection in a resource-constrained environment.

Classifier Model

Let $\mathbf{x} \in \mathbb{R}^d$ denote the input feature vector extracted from a URL, where $d = 30$ represents the total number of tabular features. Each categorical feature x_i is first mapped into a low-dimensional embedding space using a learnable embedding function.

$$\mathbf{e}_i = \text{Emb}_i(x_i), \quad i = 1, 2, \dots, d \quad (1)$$

The individual embeddings are concatenated to form a unified representation:

$$\mathbf{e} = [\mathbf{e}_1 \parallel \mathbf{e}_2 \parallel \dots \parallel \mathbf{e}_d] \quad (2)$$

This representation is passed through a lightweight feed-forward classifier consisting of linear transformations and non-linear activations:

$$\mathbf{h}_1 = \sigma(\mathbf{W}_1 \mathbf{e} + \mathbf{b}_1) \quad (3)$$

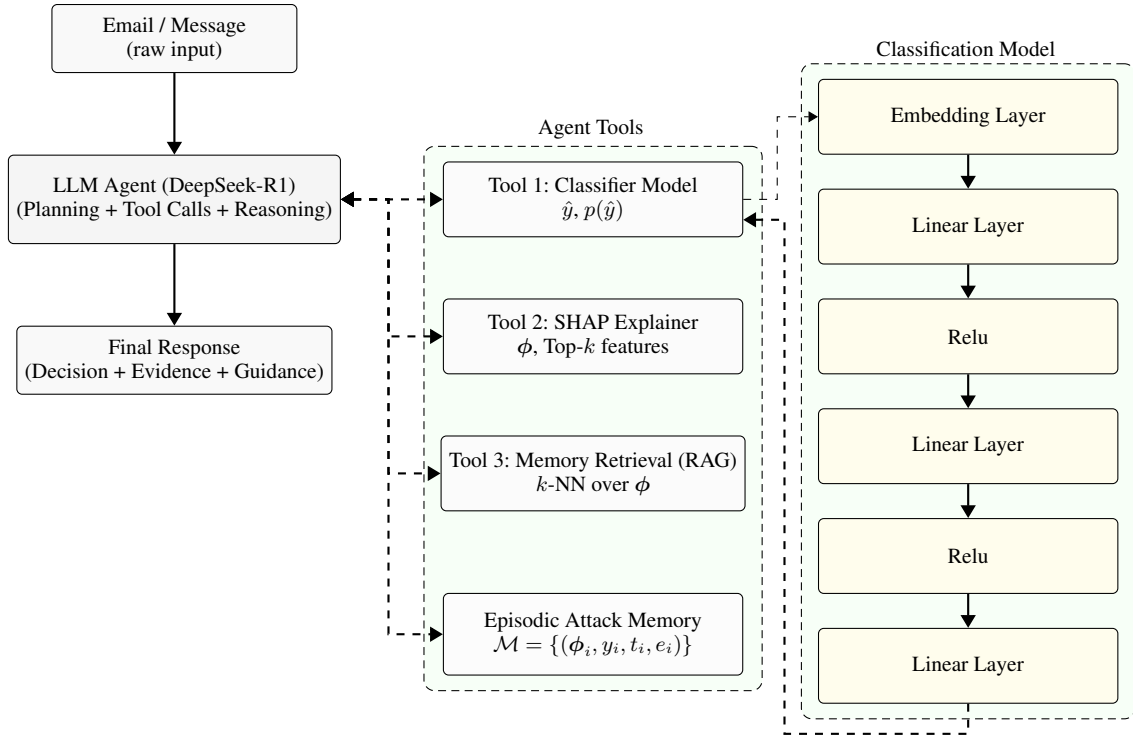


Figure 1: Proposed Model

$$\mathbf{h}_2 = \sigma(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2) \quad (4)$$

$$\mathbf{z} = \mathbf{W}_3 \mathbf{h}_2 + \mathbf{b}_3 \quad (5)$$

where $\sigma(\cdot)$ denotes the ReLU activation function. The final prediction probability is obtained using the softmax function:

$$\hat{y} = \text{softmax}(\mathbf{z}) \quad (6)$$

The classifier outputs the predicted label \hat{y} and its associated confidence score $p(\hat{y}|\mathbf{x})$:

$$\hat{y} = \arg \max_c \hat{y}_c, \quad p = \max_c \hat{y}_c \quad (7)$$

Feature Attribution via SHAP

To quantify the contribution of each input feature to the classifier's decision, SHAP values are computed:

$$\phi = \{\phi_1, \phi_2, \dots, \phi_d\} \quad (8)$$

where ϕ_i represents the marginal contribution of feature x_i to the prediction \hat{y} . The top- k influential features are selected as:

$$\phi^{(k)} = \text{TopK}(|\phi|, k) \quad (9)$$

Episodic Attack Memory Retrieval

An episodic memory bank \mathcal{M} is maintained to store historical attack representations:

$$\mathcal{M} = \{(\phi_j, y_j, t_j)\}_{j=1}^N \quad (10)$$

where ϕ_j denotes the SHAP-based feature attribution of a past sample, y_j its label, and t_j the timestamp.

Similarity between the current sample and memory entries is computed using a distance metric (e.g., cosine or Euclidean distance):

$$d_j = \text{dist}(\phi, \phi_j) \quad (11)$$

The top- k most similar historical samples are retrieved as:

$$\mathcal{M}^{(k)} = \arg \min_{j \in \mathcal{M}}^k d_j \quad (12)$$

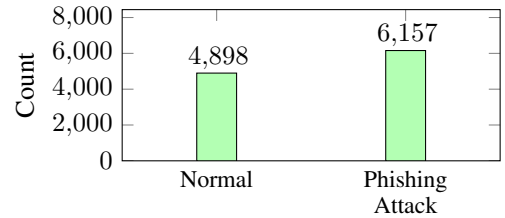


Figure 2: Class Distribution

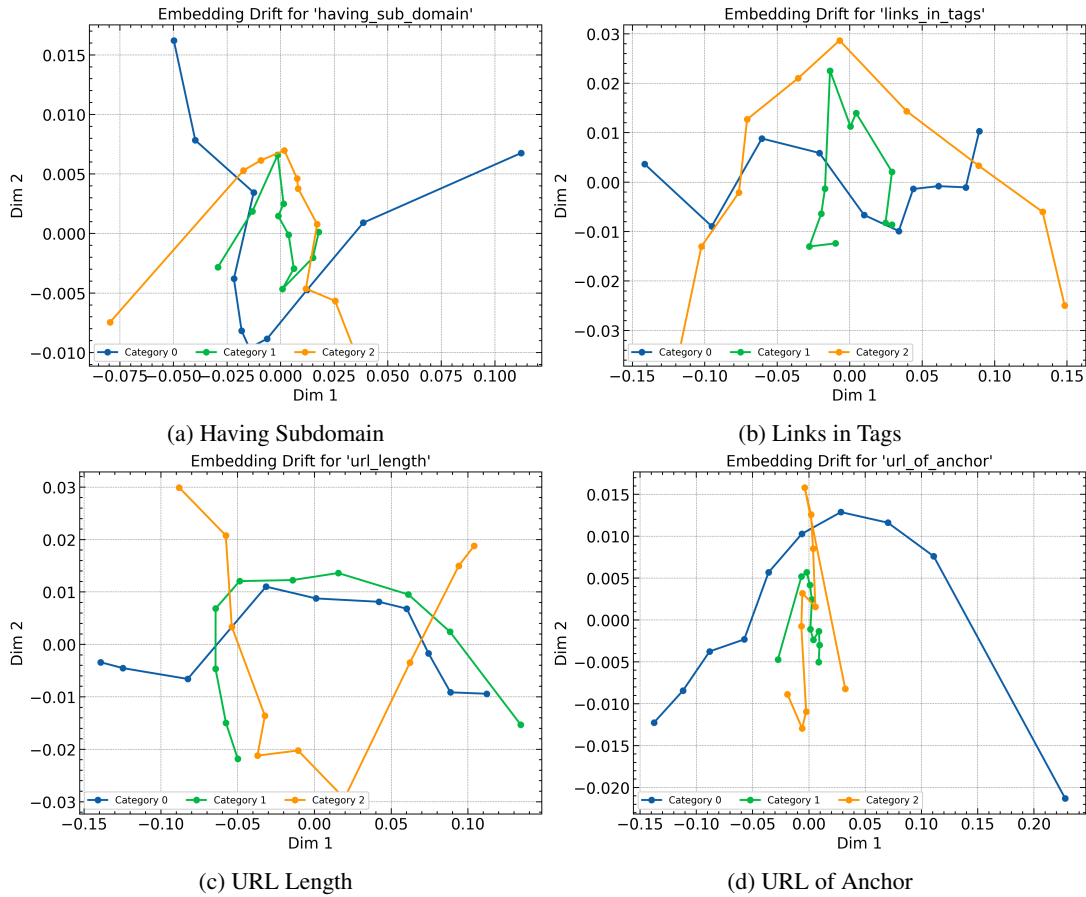


Figure 3: Embeddings Drift

Agentic LLM Reasoning

The LLM agent \mathcal{A} does not perform classification directly; instead, it reasons over structured tool outputs. Given the tuple:

$$\mathcal{T} = \{\hat{y}, p, \phi^{(k)}, \mathcal{M}^{(k)}\} \quad (13)$$

the agent generates a final response:

$$\mathcal{R} = \mathcal{A}(\mathcal{T}) \quad (14)$$

where \mathcal{R} consists of:

- an explanation of the classifier’s decision,
- an assessment of similarity to past phishing attacks,
- identification of shared or novel feature patterns,
- and a concise security recommendation.

Results and Discussion

Dataset Representation

The dataset that was used in this study was sourced from Kaggle (Mohammad and McCluskey 2012). It contains URLs labelled as phishing or safe. The class distribution of our dataset is shown in Figure 2. The phishing classes have

6,157 samples, while the safe class has 4,898. The data reflected a moderate class imbalance as phishing URLs being the majority class.

An unbalanced dataset can create a bias towards the dominant class in the learning process, which degrades detection performance for the lesser class. In order to help alleviate this problem and help in the fair training of the model, a class weight balancing strategy was used during optimization. The computed class weights were set to:

$$\text{Class Weights} = [1.1285, 0.8978] \quad (15)$$

Thus, more weightage for misclassified example for a minority class and less weightage for majority class example will be given. So, the classifier generally learns balanced decision boundaries.

Embedding Drift

An embedding drift study on selected URL features was conducted to evaluate the stability and discriminative behavior of learned feature representations. The proposed model uses embeddings for all 30 features. However, for 4 samples namely having_sub_domain, links_in_tags, url_length and url_of_anchor, its visualizations have been presented in Figure 3. The plots show the changes in low-dimensional

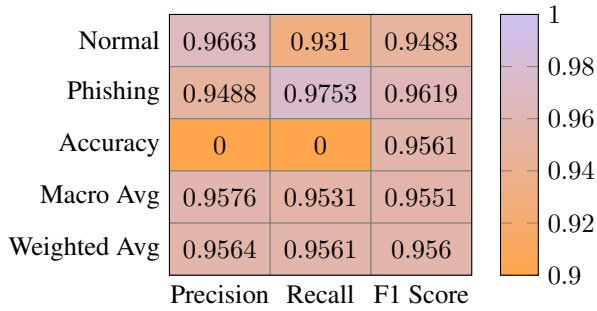


Figure 4: Classification Report

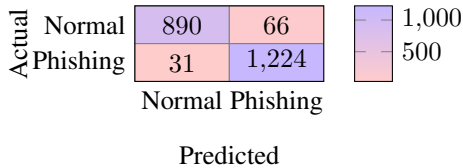


Figure 5: Confusion Matrix

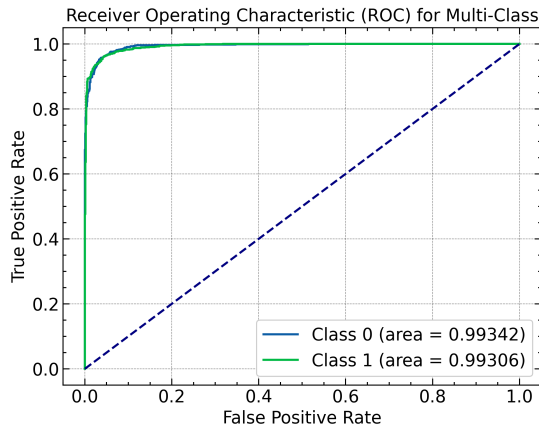


Figure 6: ROC Curve

representations (Dim 1 and Dim 2) of the learned embeddings for three categorical states.

As can be observed in Figure 3, it is evident that all features demonstrate structured and well-separated embedding trajectories across all categories, suggesting that the lightweight classifier learns meaningful latent representations instead of a collapse. The smooth and compactness of the drift patterns indicates that the embedding space is stable yet adaptable to feature-specific variations that are important for phishing URL detection.

In particular, the features `url.length` and `url.of.anchor` show clearer directional separation between the categories, depicting their stronger contribution. Likewise, we observe that the embeddings for `having.sub.domain` and `links.in.tags` follow the same geometric patterns, indicating semantic relevance.

As seen in the visual representations of embedding drift in Figure 3, the embedding-based representation strategy we’re proposing creates coherent and discriminative feature mapping across samples.

Model Performance

The proposed lightweight agentic AI framework’s performance was analyzed using the confusion matrix, classification report and ROC curve.

In Figure 5, the confusion matrix indicates that almost all normal and phishing URLs are correctly classified by the model. Most phishing samples are detected accurately with only a few mislabelled samples, indicating a low false-negative rate. Such an activity is essential for phishing detection engines since a phishing URL left undetected poses a security threat. Similarly, the number of false positives remains small, meaning that legitimate URLs are not overly flagged as phishing.

The classification report providing further quantitative assessment in terms of precision, recall and F1 score can be seen in Figure 4. The final proposed model is able to achieve high and well-balanced scores while providing an overall accuracy of about 95.6%. The macro and weighted averages of the precision, recall, and F1-score are also more than 95%. Thus the model knows no bias towards the phishing URL class or the safe URL class.

Figure 6, the ROC curve shows that the model can discriminate the treatment outcomes. The ROC curves for both

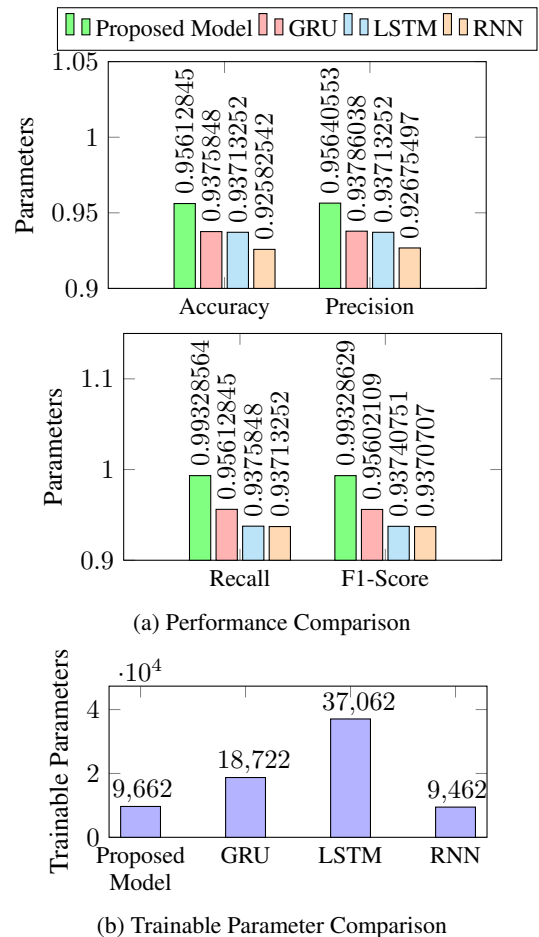


Figure 7: Comparative Analysis

Prediction: non-phishing
Confidence: 1.000
Memory status: USED (3 hits)

- **Hit-1:** label=non-phishing, features=['url_of_anchor', 'sslfinal_state', 'having_ip_address'], distance=0.030
- **Hit-2:** label=non-phishing, features=['sslfinal_state', 'google_index', 'having_ip_address'], distance=0.472
- **Hit-3:** label=non-phishing, features=['sslfinal_state', 'having_ip_address', 'links_in_tags'], distance=0.509

Top SHAP factors: url_of_anchor: -0.1940 sslfinal_state: -0.1513 web_traffic: -0.0417
LLM Explanation: Analysis of Cybersecurity Sample

1. **Explanation of Classifier Prediction:** The classifier predicted "non-phishing" with high confidence (1.00), indicating strong evidence against phishing. This was influenced by the SHAP values, which show that features like 'url_of_anchor', 'sslfinal_state', and 'web_traffic' had negative contributions. Negative SHAP values mean these features decreased the likelihood of phishing, suggesting that the URL structure (e.g., normal anchor tags), SSL certificate status (e.g., valid SSL present), and web traffic patterns (e.g., benign traffic) align with safe browsing characteristics. For instance, a negative contribution for 'sslfinal_state' implies that a secure SSL state helped classify the sample as non-phishing, while 'url_of_anchor' likely refers to standard URL components without suspicious elements.
2. **Resemblance to Previously Observed Attacks:** The current sample resembles previously observed non-phishing examples, as indicated by the episodic attack memory retrieval. All three retrieved samples are labeled "non-phishing" with shared features such as 'sslfinal_state' and 'having_ip_address'. The low similarity distance (0.030) for one memory entry suggests a close match to a benign sample, reinforcing that the current sample is not indicative of phishing attacks. Since no phishing-related memories were retrieved, there is no resemblance to known attacks.
3. **Highlighted Shared and Novel Categorical Patterns:**
 - - **Shared Patterns:** The sample shares categorical features with past non-phishing examples, such as 'sslfinal_state' (negative contribution, indicating a secure SSL state) and 'having_ip_address' (common in benign contexts). These patterns emphasize the importance of SSL and IP handling in distinguishing non-phishing content.
 - - **Novel Patterns:** The feature 'web_traffic' shows a small negative contribution, which may indicate a less emphasized but still benign aspect of web behavior. While not explicitly novel, its inclusion could highlight evolving phishing detection needs, but it aligns with standard cybersecurity features.
4. **Concise Security Recommendation:** Monitor and verify SSL certificates and URL structures regularly to maintain non-phishing integrity, especially if web traffic patterns change, as this could signal emerging threats.

(a) Response 1

Prediction: phishing
Confidence: 0.992
Memory status: USED (3 hits)

- **Hit-1:** label=phishing, features=['sslfinal_state', 'url_of_anchor', 'google_index'], distance=0.012
- **Hit-2:** label=phishing, features=['sslfinal_state', 'sfh', 'url_of_anchor'], distance=0.226
- **Hit-3:** label=non-phishing, features=['sslfinal_state', 'google_index', 'web_traffic'], distance=1.340

Top SHAP factors:

- sslfinal_state: +0.2024
- url_of_anchor: +0.1102
- google_index: -0.0458

LLM Explanation: **Analysis of Cybersecurity Sample**
Based on the outputs from the external tools, I'll reason step by step to address your tasks. The classifier tool indicated a phishing prediction with 100% confidence, the SHAP explanation tool provided feature contributions, and the episodic attack memory tool retrieved similar past samples. Here's my reasoning:

1. **Explain why the classifier produced this prediction**
The classifier predicted "phishing" with high confidence (1.00) due to strong positive contributions from specific features in the SHAP explanation. The 'sslfinal_state' feature contributed +0.202, which suggests that an invalid or untrusted SSL certificate is a key indicator of phishing, as this is often associated with malicious sites attempting to mimic legitimate ones. Similarly, the 'url_of_anchor' feature contributed +0.110, indicating that suspicious or crafted anchor URLs (e.g., in hyperlinks) are a common tactic in phishing attacks to deceive users. The negative contribution from 'google_index' (-0.046) slightly reduces the phishing likelihood if the site is indexed by Google, but this is a weak mitigating factor and does not outweigh the positive signals. Overall, the combination of these features aligns with known phishing patterns, leading to the high-confidence prediction.
2. **State whether the current sample resembles previously observed attacks** Yes, the current sample closely resembles previously observed attacks. The episodic attack memory retrieval tool shows that the sample shares features with past phishing incidents, particularly with a similarity distance of 0.012 to a recent phishing attack (which is very low, indicating high similarity). This suggests that the current sample is highly similar to known phishing threats. The second retrieved entry (similarity distance 0.226) also indicates resemblance, though less strongly, while the third entry (similarity distance 1.340) is not relevant due to its non-phishing label and lack of feature overlap.
3. **Highlight shared or novel categorical patterns** Shared patterns: The current sample exhibits common phishing characteristics, such as 'sslfinal_state' and 'url_of_anchor', which were also prominent in past phishing attacks (e.g., the first two retrieved entries). This reinforces that invalid SSL states and suspicious anchor URLs are recurring indicators of phishing. Novel patterns: While the SHAP explanation and retrieval tool did not highlight any entirely new features, the presence of 'sfh' (session history) and 'web_traffic' in past attacks (from the retrieval tool) could indicate evolving tactics, but these are not evident in the current sample. This suggests that the current threat is more aligned with established phishing methods rather than introducing novel categories.
4. **Provide ONE concise security recommendation** Block or quarantine the URL immediately to prevent user interaction, as the high similarity to past attacks and strong phishing indicators warrant proactive mitigation.

(b) Response 2

Figure 8: LLM (DeepSeek-R1) Response

classes are seen towards the upper-left corner and thus the corresponding AUC values are around 0.99 for both classes. The results demonstrate very good separation between phishing and normal URLs, and also show robustness of the proposed classifier for several decision thresholds.

Comparative Analysis

We compared the results of the new lightweight classifier against the commonly used recurrent architectures or models, GRU, LSTM, and RNN. As depicted in Figure 7, both classification performance and model complexity are compared.

As indicated in Figure 7a, the performance of our model, in terms of accuracy, precision, recall and F1-score, is at par if not better to GRU-, LSTM- and RNN-based classifiers across various performance metrics. The proposed model is less deep but can consistently give good balanced performance. It successfully captures discriminative patterns in phishing URLs features without the need for deep sequential architectures.

As shown in Figure 7b, the proposed method is proved to be more superior when it comes to parameters to train. The proposed model is much leaner with 9,662 trainable parameters in comparison to GRU (18,722) and LSTM (37,062) while being comparable to RNN and outperforming its predictive power. The proposed classifier is comparative lightweight, resulting in a significant reduction in model complexity that leads to a faster training speed, reduced memory footprint, and enhanced suitability for real-time or resource-constrained deployment scenarios.

The overall comparative results in Figure 7 indicate that the proposed lightweight agentic AI framework makes an effective trade-off between accuracy and computations. The proposed model shows considerable advantages over recurrent architectures for adaptive phishing URL detection by attaining competitive performance with much fewer trainable parameters.

Performance of LLM

In order to assess the reasoning ability of the LLM agent in the proposed framework, qualitative analysis was done based on representative response examples of LLM. The samples shown in Figure 8a and Figure 8b are the response for non-phishing and phishing URL predictions respectively. As shown in the examples, the LLM agent integrates classifier outputs, SHAP feature importance, and episodic memory retrieval for structured and actionable explanations.

The LLM agent (depicted in Figure 8a) analyzes a URL that has been deemed non-phishing with high confidence. The explanation points towards the dominant benign indicators like a valid SSL state and the absence of suspicious anchor or traffic patterns. Additionally, the agent cross-references similar past samples from memory and marks them as not phishing. Through memory-aware reasoning, the classifier is assured and assured that the behaviour of the URL is legitimate for the reference.

On the other hand, Figure 8b shows the LLM agent's output when prediction of a URL is phishing with full confidence. Description strong signs phishing such as aberration in the use or anchor URL and structural features An agent infers an episode of phishing attacks with feature signatures similar to existing ones demonstrating common features in these attacks. Taking into account this combinatorial evidence, the LLM agent not only supports prediction, but also generates a concise cybersecurity recommendation, e.g. blocking or quarantining the URL to protect users.

In general, the answers present indicate the LLM agent does more than report labels. It carries out multi-step reasoning in a sophisticated manner by, first, interpreting the confidence of the classifier, second, prioritizing the influential features, third, contextualizing the decision with histor-

ical memory, and fourth, producing human-readable security advice or guidance. The findings confirm that the LLM component is an intelligent agent that aids trust, situational awareness, and decision support in adaptive phishing URL detection systems.

Conclusion

This research proposed a light-weight agentic AI framework for detecting malicious phishing URLs that incorporates efficient classification, feature-driven reasoning and memory-based contextual analysis. The compactness of the tabular classifier containing only 9,662 trainable parameters ensures that the framework provides an accurate prediction while being amenable to real-time use and resource-constrained settings. The reasoning, driven by DeepSeek-R1, that combines classifier outputs, feature importance and episodic memory yields interpretable decisions and actionable security advice. Experimental tests show strong effectiveness with 95.6% accuracy and AUC values exceeding 0.99 and are more efficient than the state-of-the-art GRU, LSTM, and RNN. In general, the suggested framework offers a good solution to ever-changing phishing URLs that is deployed directly.

Acknowledgments

This research work is supported by National Science and Technology Council (NSTC), Taiwan Grant No. NSTC112-2221-E-468-008-MY3.

References

- Ali, H.; Chen, D.; Harrington, M.; Salazar, N.; Al Ameedi, M.; Khan, A. F.; Butt, A. R.; and Cho, J.-H. 2023. A survey on attacks and their countermeasures in deep learning: Applications in deep neural networks, federated, transfer, and deep reinforcement learning. *IEEE Access*, 11: 120095–120130.
- Ali, S.; Rehman, S. U.; Imran, A.; Adeem, G.; Iqbal, Z.; and Kim, K.-I. 2022. Comparative evaluation of ai-based techniques for zero-day attacks detection. *Electronics*, 11(23): 3934.
- Aljofey, A.; Jiang, Q.; Qu, Q.; Huang, M.; and Niyigena, J.-P. 2020. An effective phishing detection model based on character level convolutional neural network from URL. *Electronics*, 9(9): 1514.
- Almomani, A.; Gupta, B. B.; Atawneh, S.; Meulenberg, A.; and Almomani, E. 2013. A survey of phishing email filtering techniques. *IEEE communications surveys & tutorials*, 15(4): 2070–2090.
- Asiri, S.; Xiao, Y.; and Li, T. 2023. PhishTransformer: a novel approach to detect phishing attacks using URL collection and transformer. *Electronics*, 13(1): 30.
- Capuano, N.; Fenza, G.; Loia, V.; and Stanzione, C. 2022. Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access*, 10: 93575–93600.
- Ejaz, A.; Mian, A. N.; and Manzoor, S. 2023. Life-long phishing attack detection using continual learning. *Scientific reports*, 13(1): 11488.

- Elsadig, M.; Ibrahim, A. O.; Basheer, S.; Alohal, M. A.; Alshunaifi, S.; Alqahtani, H.; Alharbi, N.; and Nagmeldin, W. 2022. Intelligent deep machine learning cyber phishing url detection based on bert features extraction. *Electronics*, 11(22): 3647.
- Gragnaniello, M.; Borghese, A.; Marrazzo, V. R.; Maresca, L.; Breglio, G.; Irace, A.; and Riccio, M. 2024. Real-time myocardial infarction detection approaches with a microcontroller-based edge-AI device. *Sensors*, 24(3): 828.
- Gupta, B.; Gaurav, A.; Attar, R.; Arya, V.; Bansal, S.; Alhomoud, A.; and Chui, K. 2024a. A hybrid cnn-brown-bear optimization framework for enhanced detection of url phishing attacks. *Computers, Materials, & Continua*, 81(3): 4853.
- Gupta, B. B.; Gaurav, A.; Attar, R. W.; Arya, V.; Alhomoud, A.; and Chui, K. T. 2024b. Optimized Phishing Detection with Recurrent Neural Network and Whale Optimizer Algorithm. *Computers, Materials & Continua*, 80(3).
- Hoenig, A.; Roy, K.; Acquaaah, Y. T.; Yi, S.; and Desai, S. S. 2024. Explainable AI for cyber-physical systems: Issues and challenges. *IEEE Access*, 12: 73113–73140.
- Ishtaiwi, A.; Ali, A. M.; Al-Qerem, A.; Sabahean, M.; Alzubi, B.; Almomani, A.; Alauthman, M.; Aldweesh, A.; and Al Khaldy, M. A. 2024. Next-gen phishing defense enhancing detection with machine learning and expert whitelisting/blacklisting. *International Journal of Cloud Applications and Computing (IJCAC)*, 14(1): 1–17.
- Kalinin, M.; Zavadskii, E.; and Busygin, A. 2023. A graph-based technique for securing the distributed cyber-physical system infrastructure. *Sensors*, 23(21): 8724.
- Khan, Z. A.; Xia, Y.; Ali, S.; Khan, J. A.; Askar, S. S.; Abouhawwash, M.; and El-Rashidy, N. 2023. Identifying Hot Topic Trends in Streaming Text Data Using News Sequential Evolution Model Based on Distributed Representations. *IEEE Access*, 11: 98787–98804.
- Kim, K.; Jang, S.-J.; Park, J.; Lee, E.; and Lee, S.-S. 2023. Lightweight and energy-efficient deep learning accelerator for real-time object detection on edge devices. *Sensors*, 23(3): 1185.
- Kritika, E. 2025. A comprehensive literature review on phishing URL detection using deep learning techniques. *Journal of Cyber Security Technology*, 9(4): 315–343.
- Kyaw, P. H.; Gutierrez, J.; and Ghobakhlou, A. 2024. A systematic review of deep learning techniques for phishing email detection. *Electronics*, 13(19): 3823.
- Mohammad, R.; and McCluskey, L. 2012. Phishing Websites. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C51W2X>.
- Nagy, N.; Aljabri, M.; Shaahid, A.; Ahmed, A. A.; Alnasser, F.; Almakrany, L.; Alhadab, M.; and Alfaddagh, S. 2023. Phishing URLs detection using sequential and parallel ML techniques: Comparative analysis. *Sensors*, 23(7): 3467.
- Neupane, S.; Ables, J.; Anderson, W.; Mittal, S.; Rahimi, S.; Banicescu, I.; and Seale, M. 2022. Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. *IEEE Access*, 10: 112392–112415.
- Oh, S. H.; Kim, J.; and Park, J. 2024. Dynamic cyberattack simulation: Integrating improved deep reinforcement learning with the MITRE-ATT&CK framework. *Electronics*, 13(14): 2831.
- Owoputi, R.; and Ray, S. 2022. Security of multi-agent cyber-physical systems: A survey. *IEEE Access*, 10: 121465–121479.
- Putra, F. P. E.; Zulfikri, A.; Arifin, G.; Ilhamsyah, R. M.; et al. 2024. Analysis of phishing attack trends, impacts and prevention methods: literature study. *Brilliance: Research of Artificial Intelligence*, 4(1): 413–421.
- Qiu, J.; Zhang, J.; Luo, W.; Pan, L.; Nepal, S.; and Xiang, Y. 2020. A survey of android malware detection with deep neural models. *ACM Computing Surveys (CSUR)*, 53(6): 1–36.
- Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I. A.; and Xu, M. 2020. A survey on machine learning techniques for cyber security in the last decade. *IEEE Access*, 8: 222310–222354.
- Tamal, M. A.; Islam, M. K.; Bhuiyan, T.; and Sattar, A. 2024. Dataset of suspicious phishing URL detection. *Frontiers in computer science*, 6: 1308634.
- Vajrobol, V.; Gupta, B. B.; and Gaurav, A. 2024. Mutual information based logistic regression for phishing url detection. *Cyber Security and Applications*, 2: 100044.
- Varshney, G.; Kumawat, R.; Varadharajan, V.; Tupakula, U.; and Gupta, C. 2024. Anti-phishing: A comprehensive perspective. *Expert Systems with Applications*, 238: 122199.
- Wagner, N.; Damodaran, S. K.; and Reavey, M. 2025. Towards Optimal Sensor Placement for Cybersecurity: An Extensible Model for Defensive Cybersecurity Sensor Placement Evaluation. *Sensors*, 25(19): 6022.
- Wei, B.; Hamad, R. A.; Yang, L.; He, X.; Wang, H.; Gao, B.; and Woo, W. L. 2019. A deep-learning-driven light-weight phishing detection sensor. *Sensors*, 19(19): 4258.
- Wilk-Jakubowski, J. L.; Pawlik, L.; Wilk-Jakubowski, G.; and Sikora, A. 2025. Machine learning and neural networks for phishing detection: A systematic review (2017–2024). *Electronics*, 14(18): 3744.
- Yang, P.; Zhao, G.; and Zeng, P. 2019. Phishing website detection based on multidimensional features driven by deep learning. *IEEE Access*, 7: 15196–15209.
- Zhang, Z.; Al Hamadi, H.; Damiani, E.; Yeun, C. Y.; and Taher, F. 2022. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, 10: 93104–93139.