Check for updates

RESEARCH ARTICLE

REVISED SentinelSphere: An AI-driven cybersecurity platform integrating real-time threat detection with security awareness education

[version 2; peer review: 2 approved, 2 approved with reservations]

Nikolaos D. Tantaroudas[1], Ilias Karachalios[2], Andrew J. McCracken[3]

[1]Ethniko Metsobio Polytechneio Ereunetiko Panepistemiako Institouto Systematon Epikoinonion kai Ypologiston, Athens, Attica, Greece
[2]National Technical University of Athens, Zografou, Attica, Greece
[3]183 Rue de l'Abbé Griffon, DASKALOS-APPS, Péronnas, 01960, France

v2 **First published:** 25 Feb 2026, **6**:58
https://doi.org/10.12688/openreseurope.22957.1
Latest published: 13 Jun 2026, **6**:58
https://doi.org/10.12688/openreseurope.22957.2

Abstract

Background

The cybersecurity domain faces dual challenges: a global shortage of qualified professionals and persistent human-factor vulnerabilities contributing to the majority of security breaches. Traditional Security Information and Event Management (SIEM) systems generate excessive false positives causing analyst alert fatigue, while conventional awareness programmes demonstrate limited effectiveness in changing user behaviour. Integrated solutions addressing both technical detection and human-centric education are needed.

Methods

We developed SentinelSphere, an AI-driven platform combining machine learning-based threat detection with Large Language Model (LLM)-powered security training. The detection component employs an Enhanced Deep Neural Network trained on the CIC-IDS2017 and CIC-DDoS2019 benchmark datasets. To mitigate the substantial class imbalance present in these datasets (e.g. 168,186 benign samples versus 21 SQL injection samples), class weights inversely proportional to class frequencies were applied during training. The educational component utilises Microsoft's Phi-4 model (14 billion parameters) with quantisation techniques to enable deployment on standard

Open Peer Review

Approval Status ✓ ✓ ? ?

	1	2	3	4
version 2 (revision) 13 Jun 2026	✓ view	✓ view	? view	? view
version 1 25 Feb 2026	? view	? view		

1. **Hewa Majeed Zangana**, Duhok Technical College, Duhok Polytechnic University, Duhok, Iraq
2. **Aryendra Dalal**, Middle Georgia State University, Macon, USA
3. **Hassan Rehan**, Purdue University, West Lafayette, USA
4. **Oluwaseun Oladeji Olaniyi**, University of the Cumberlands, Williamsburg, USA

Any reports and responses or comments on the article can be found at the end of the article.

hardware. System performance was optimised through Rust-based preprocessing and validated via pilot deployments across two workshops with a total of 79 stakeholders (3 professionals in maritime, educational and manufacturing sectors; 76 in an educational setting) in Greece.

Results

The Enhanced DNN achieved high detection accuracy with significant false positive reduction compared to baseline models, while maintaining strong recall for critical attack categories including DDoS, botnet, and brute force attacks. Rust optimisation delivered substantial speedup in both single-record and batch processing. Stakeholder validation revealed high satisfaction rates, with most participants achieving improved security concept comprehension post-training. The platform identified critical awareness gaps, particularly regarding data protection regulations and multi-factor authentication adoption.

Conclusions

SentinelSphere demonstrates that integrating intelligent threat detection with adaptive, LLM-powered security education shows strong potential to address both technical detection and human security awareness challenges, although longitudinal studies are required to validate sustained behavioural change. The resource-efficient design enables deployment in SME environments without enterprise-grade infrastructure, supporting comprehensive cyber resilience approaches within the European Union's regulatory framework.

Plain Language summary

Cyberattacks are becoming more sophisticated while organisations struggle to find enough skilled cybersecurity professionals. Most existing security systems generate too many false alarms, causing analysts to miss real threats, and they do not help employees understand and recognise security risks. The system developed in this paper, addresses these problems by combining advanced threat detection with security training in a single platform. The system uses artificial intelligence to identify cyber threats with high accuracy while significantly reducing false alarms. It displays security information using a simple traffic light system, showing green for safe, yellow for caution, and red for danger, making it easy for anyone to understand the current security status. An AI-powered chatbot helps employees learn about cybersecurity through natural conversation, running efficiently on modern office computers without special hardware. The core detection system was rewritten in the Rust programming language, making it more than five times faster compared to the baseline python framework. Testing showed the system can process

millions of security events quickly and accurately. Workshops with business professionals from shipping and manufacturing companies, and university students, confirmed that users found the platform easy to understand and valuable for improving their security awareness. Importantly, surveys before training revealed that most participants lacked basic security knowledge, showing exactly why tools like the one developed in this study are needed. This research demonstrates that protecting organisations from cyber threats requires not only better technology but also helping people become more security-aware.

Keywords
Cybersecurity Awareness, Real-Time Anomaly Detection, Security Education, Large Language Models, Human-Centric Threat Intelligence, Deep Neural Networks, Traffic Light System, Cyber Resilience, Intrusion Detection Systems, SIEM



This article is included in the Horizon Europe gateway.



This article is included in the Software gateway.

Corresponding author: Nikolaos D. Tantaroudas (nikolaos.tantaroudas@iccs.gr)

Author roles: Tantaroudas ND: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – Original Draft Preparation; **Karachalios I**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Validation, Writing – Review & Editing; **McCracken AJ**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The project has received funding from the European Union's Horizon Europe programme under Grant Agreement No. 101119681.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Tantaroudas ND *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Tantaroudas ND, Karachalios I and McCracken AJ. **SentinelSphere: An AI-driven cybersecurity platform integrating real-time threat detection with security awareness education [version 2; peer review: 2 approved, 2 approved with reservations]** Open Research Europe 2026, **6**:58 https://doi.org/10.12688/openreseurope.22957.2

First published: 25 Feb 2026, **6**:58 https://doi.org/10.12688/openreseurope.22957.1

REVISED **Amendments from Version 1**

The most substantial change concerns the validation cohorts: we now quantify the 79 stakeholders involved across the two workshops, 3 professionals drawn from the maritime, educational and manufacturing sectors, and 76 participants in an educational setting, and add a new "Overview of Participant Cohorts" subsection, with an accompanying Table 6, reporting cohort sizes, demographics and baseline expertise. The large-language-model component is now fully specified as Microsoft's Phi-4 (14 billion parameters), optimised through Q4_K_M quantisation using the llama.cpp framework with the GGUF format. We have added a dedicated explanation of how class imbalance was handled in the enhanced deep neural network (class weights inversely proportional to class frequencies, the rationale for not using SMOTE, and the 0.4 decision threshold), and a paragraph explaining the sources of the 326× Rust speedup (absence of a global interpreter lock, contiguous memory layout, and deterministic deallocation). To address concerns about educational retention, a new Discussion subsection clarifies that the workshops measured immediate comprehension rather than long-term behaviour change, and the corresponding claims in the Abstract and Conclusions have been softened accordingly; the Limitations and Future Work section now sets out a concrete longitudinal evaluation plan. Finally, all acronyms are now defined at first use, several sentences have been simplified for clarity, and tables have been renumbered to accommodate the new Table 6.

Any further responses from the reviewers can be found at the end of the article

Introduction

Background and motivation

The cybersecurity landscape faces an escalating crisis characterised by increasingly sophisticated attack vectors and a critical shortage of skilled security professionals.[1] Global estimates indicate a shortage of approximately 3.4 million cybersecurity workers, with this gap projected to grow as digital transformation accelerates across industries.[2] Traditional Security Information and Event Management (SIEM) systems, while effective at data aggregation, often overwhelm analysts with false positives and lack the contextual intelligence necessary for rapid threat response.[3] Past studies have indicated that security analysts spend up to 25% of their time investigating false positive alerts, leading to alert fatigue and increased risk of missing genuine threats.[4]

Furthermore, the human factor remains a significant vulnerability in cybersecurity. According to the Verizon 2023 Data Breach Investigations Report, 74% of all breaches include the human element, with people being involved either via error, privilege misuse, use of stolen credentials, or social engineering.[5] This finding underscores a fundamental truth: technological solutions alone cannot adequately protect organisations against modern cyber threats. Effective cybersecurity requires a holistic approach that combines advanced threat detection with systematic enhancement of human security awareness.

The European Union's approach to cybersecurity emphasises resilience in critical infrastructure protection,[6] while the NIST Cybersecurity Framework 2.0 provides comprehensive guidance for managing cybersecurity risks across organisations of all sizes.[7] The Framework's addition of a "Govern" function in its 2024 update highlights the increasing recognition that cybersecurity governance and human factors must be integral to any security strategy. These frameworks establish the conceptual foundation upon which technical solutions must be built.

The ResilMesh framework

The ResilMesh project, funded under the EU Horizon Europe programme (Grant Agreement No. 101119681), has established a comprehensive cybersecurity framework designed to enhance cyber resilience for dispersed, heterogeneous cyber systems.[8] The baseline system is built on the NATS message streaming framework, and provides secure event collection and processing capabilities while implementing resilience engineering best practices including redundancy, segmentation, and dynamic positioning.[9,10]

The architecture implements a sophisticated data pipeline where security events flow from collection agents through Vector (a high-performance observability data pipeline), are processed through NATS message broker for reliable, high-throughput delivery, undergo anomaly detection processing, and are stored in OpenSearch for indexing and analysis. This architecture provides the foundational infrastructure upon which SentinelSphere extends its advanced threat analytics and human-centric capabilities.

Research contributions

This paper presents SentinelSphere, a next-generation cybersecurity platform developed as an extension to the ResilMesh framework that addresses contemporary cybersecurity challenges through an innovative dual approach: combining advanced AI-driven threat detection with integrated security awareness education. Unlike conventional solutions that treat threat detection and security training as separate domains, SentinelSphere creates a synergistic system where every

security event becomes an opportunity for organisational learning and resilience building and can be added on top of any existing workflow, adding advances threat analytics without disrupting services.

The primary contributions of this work are fivefold: i) **Enhanced Deep Neural Network Architecture**: An Enhanced DNN model achieves 94% F1 score while reducing false positives by 69.5% through innovative feature engineering including HTTP-specific anomaly indicators such as Request Complexity Score, Response Pattern Ratio, and Path Entropy. ii) **Traffic Light System (TLS)**: A sophisticated threat visualisation system that transforms complex threat intelligence into intuitive visual indicators (Green/Yellow/Red), democratising security understanding across technical expertise levels with validated 91.7% comprehension rates. iii) **LLM-Powered Security Education**: A Large Language Model-powered conversational agent based on Microsoft's Phi-4 (14 billion parameters), optimised through Q4_K_M quantisation using the llama.cpp framework with the GGUF format (reducing model size from 28 GB to 2.5 GB) for deployment on standard enterprise hardware, providing real-time, context-aware cybersecurity guidance without requiring specialised infrastructure. iv) **Performance Optimisation**: Complete algorithm rewrite from Python to Rust achieving 5.6× average speedup for steady-state operations and up to 326× improvement for large batch processing, enabling enterprise-scale deployment. v) **Comprehensive Validation**: Multi-sector stakeholder validation through professional and educational workshops with 79 participants (3 professionals across maritime, educational, and manufacturing sectors, and 76 attendees from the educational workshop) demonstrating 86.1% user satisfaction, 77.8% deployment interest, and identification of critical cybersecurity knowledge gaps in target populations. This work extends our preliminary findings[11] by incorporating detailed results from pilot demonstrations, performance optimisations through Rust implementation, and comprehensive stakeholder validation across multiple sectors.

Related work

Machine learning for intrusion detection

The application of machine learning to cybersecurity has evolved significantly from rule-based systems to sophisticated deep learning architectures. Early intrusion detection systems relied primarily on signature-based detection, which proved effective against known threats but failed to identify novel attack vectors.[12] The emergence of anomaly-based detection using machine learning addressed this limitation by learning patterns of normal behaviour and identifying deviations.

Yin et al.[13] demonstrated the effectiveness of Recurrent Neural Networks (RNNs) for intrusion detection, achieving 98.6% accuracy on the NSL-KDD dataset. Their work established the viability of deep learning for sequential network traffic analysis but suffered from high false positive rates in production environments due to the temporal dependencies in network traffic that RNNs capture well but also overfit to.

Vinayakumar et al.[14] proposed a hybrid deep learning framework combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, creating scale-hybrid-IDS-AlertNet. Their framework improved detection of zero-day attacks by capturing both spatial patterns (through CNNs) and temporal patterns (through LSTMs) in network traffic. However, the computational requirements of their approach limited practical deployment in resource-constrained environments.

The release of the CIC-IDS2017 and CIC-DDoS2019 datasets by the Canadian Institute for Cybersecurity[15,16] provided the research community with realistic, labelled network traffic data for training and evaluation. These datasets contain over 400 GB of network traffic including diverse attack vectors: brute force attacks, cross-site scripting (XSS), SQL injection, distributed denial of service (DDoS), and benign traffic patterns. Sharafaldin et al.[17] provided detailed analysis of CIC-IDS2017, establishing benchmarks that subsequent research has sought to exceed.

Recent advances have focused on reducing false positives while maintaining detection accuracy. Khan et al.[18] introduced attention mechanisms for network traffic analysis, achieving 92% precision by allowing models to focus on the most relevant features of network flows. However, their approach remained limited to network-layer attacks and did not incorporate application-layer intelligence. Ferrag et al.[19] conducted a comprehensive survey of deep learning approaches for intrusion detection, identifying the trade-off between detection accuracy and false positive rates as a persistent challenge. Contractive autoencoders for anomaly detection have shown promise in reducing false positives.[20] By learning compressed representations of normal traffic patterns and detecting deviations, these approaches have achieved strong performance on the CIC-DDoS2019 dataset (97.58% accuracy) while maintaining lower false alarm rates than supervised approaches. Our work extends these approaches by incorporating application-layer features, particularly HTTP-specific patterns, to enhance detection accuracy across multiple attack vectors while significantly reducing false positive rates. The key innovation lies in the feature engineering approach that combines network-layer statistics with HTTP-specific indicators.

Large language models in cybersecurity

Large Language Models (LLMs) have emerged as transformative tools across multiple domains, including cybersecurity.[21] Recent surveys[22,23] have systematically analysed the application of LLMs to cybersecurity tasks, identifying six key domains: vulnerability detection, anomaly detection, cyber threat intelligence, penetration testing, digital forensics, and blockchain security.

The application of LLMs to cybersecurity education represents a particularly promising area. Jaffal et al.[24] provide a comprehensive systematic literature review covering over 300 works on LLMs in cybersecurity, identifying cybersecurity knowledge assistants as a key application area. These systems improve users' security awareness and assist in defending against cyber-attacks through natural language interaction.

Atlam et al.[25] explored Generative Pre-trained Transformer (GPT)-based systems for security questionnaire generation, demonstrating that LLMs can effectively create educational content for security awareness training. However, their deployment required significant computational resources (high-end GPUs), limiting accessibility for organisations without specialised infrastructure.

Hassanin and Moustafa[26] provide a comprehensive overview of LLMs for cyber defence, categorising techniques into threat intelligence, vulnerability assessment, network traffic security, privacy preservation, personnel awareness, and ethical security. Their survey highlights that LLMs can significantly improve cyber security operations and threat detection, while also noting the importance of model efficiency for practical deployment.

Chhetri[27] explored LLM-powered pedagogical approaches to cybersecurity education, demonstrating that LLMs can serve as effective cognitive assistants supporting interpretation, explanation, and troubleshooting in cybersecurity courses. Their framework emphasises active student engagement and verification against authoritative sources.

Our approach presented in this study, addresses the existing computational limitation of domain specific LLMs through model quantisation techniques. By applying Q4_K_M quantisation (via the llama.cpp framework, GGUF format) to Microsoft's Phi-4 model (14 billion parameters), we reduce model size from 28 GB to 2.5 GB while maintaining educational effectiveness. This enables deployment on standard enterprise hardware (16 GB RAM, CPU-only), achieving 15-20 tokens per second, sufficient for interactive educational dialogue.

Human factors in cybersecurity

The integration of human factors in cybersecurity has gained prominence following high-profile breaches attributed to social engineering. Research consistently demonstrates that technology alone cannot adequately protect organisations, and human behaviour plays a critical role in both creating vulnerabilities and implementing effective defences.[28]

Aldawood and Skinner[29] conducted a comprehensive review of cybersecurity social engineering training and awareness programs, identifying common pitfalls including one-time training approaches, lack of engagement, and failure to adapt content to different user populations. Their findings emphasise the importance of continuous, interactive training that adapts to evolving threats.

Bada et al.[30] investigated why cybersecurity awareness campaigns fail to change behaviour, identifying key factors including information overload, lack of personal relevance, and absence of immediate feedback. Their work suggests that effective security education must be contextualised, incremental, and integrated into daily workflows, principles that inform the design of the system presented in this study.

The multimedia approach to cybersecurity awareness has been systematically reviewed by Zhang-Kennedy and Chiasson,[31] who found that interactive, visual approaches outperform traditional text-based training. This finding directly supports our Traffic Light System design, which transforms complex threat data into intuitive visual indicators.

Cyber resilience frameworks

The concept of cyber resilience has evolved from traditional security approaches to encompass preparation, response, and recovery capabilities.[32] Unlike traditional security, which focuses primarily on prevention, resilience acknowledges that breaches are inevitable and organisations must be prepared to detect, respond to, and recover from incidents.

The cybersecurity mesh architecture concept[33] enables distributed security services while maintaining centralised intelligence and coordination. This architectural pattern aligns with SentinelSphere's design as a modular extension to the ResilMesh framework, providing advanced analytics without disrupting existing security workflows.

Recent work by Somma et al.[34] demonstrated edge-based anomaly detection within the ResilMesh framework, highlighting the importance of distributed processing for critical infrastructure protection. Their work on smart water distribution systems showed that edge-based detection can achieve comparable accuracy to centralised approaches while reducing latency and bandwidth requirements. SentinelSphere builds upon this foundation by adding human-centric security awareness capabilities while maintaining seamless integration with the ResilMesh ecosystem.

System architecture and design

Overall architecture

SentinelSphere adopts a microservices architecture integrated with the ResilMesh security framework, enabling scalable, real-time processing of security events. As illustrated in Figure 1, the system comprises four primary layers: Data Ingestion, Processing and Analysis, Intelligence and Education, and Presentation.

Figure 1 illustrates the high-level system architecture, outlining where SentinelSphere integrates with the ResilMesh stack. The architecture shows security events flowing from multiple data sources through Vector for log collection, NATS message broker for reliable delivery, the Anomaly Detection module for threat identification, and finally to the SentinelSphere dashboard for visualisation and user interaction.

The architecture operates as a strategic enhancement to the ResilMesh security framework, integrating as an advanced threat analytics layer while preserving the existing architecture's integrity. SentinelSphere operates through direct subscriptions to ResilMesh's NATS messaging infrastructure, consuming both enriched security events ('enriched_events') and anomaly detection alerts ('ad_events'). This non-intrusive approach allows SentinelSphere to function as a parallel consumer without disrupting established workflows.

Figure 2 presents the comprehensive data flow pipeline architecture. Events move from Vector through NATS message broker to the Anomaly Detector, then through ad_events topic to the Threat Calculator, which updates Redis before displaying on the Dashboard. The diagram illustrates the integration paths between SentinelSphere and ResilMesh components.

The Data Ingestion layer utilises Vector for log collection and transformation, processing diverse data sources including network traffic, application logs, and system events. Vector serves as the primary data collection and transformation layer, ingesting security events from multiple sources and performing real-time transformations including routing, logging, fusion, filtering, augmentation, reduction, and monitoring. Events flow through the NATS message broker,

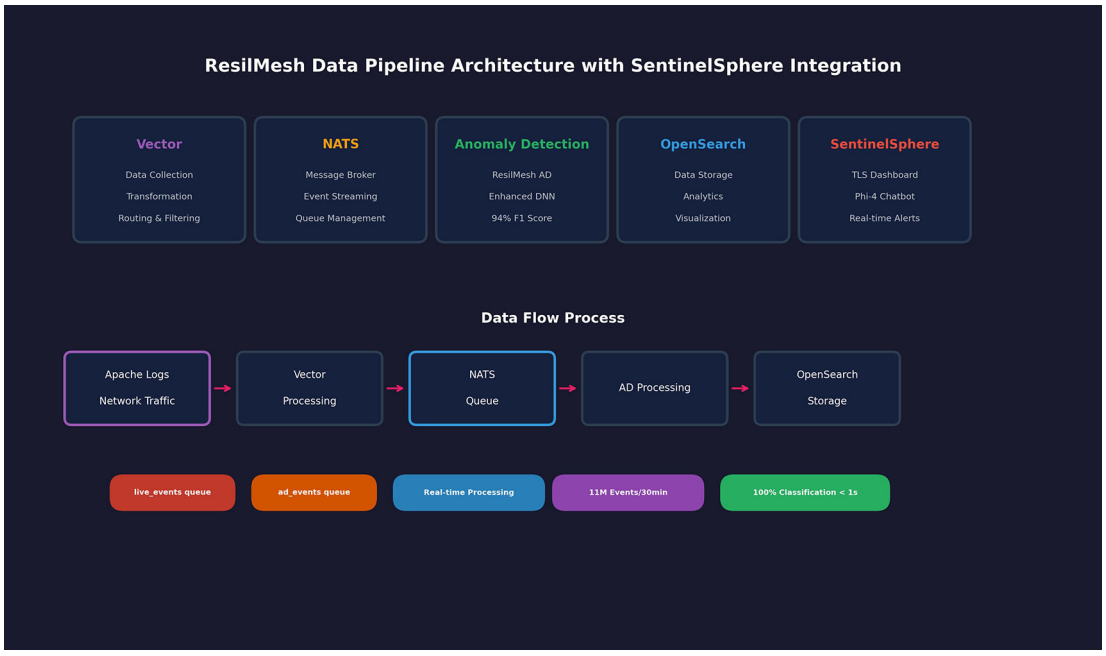


Figure 1. High-level system architecture showing SentinelSphere integration with ResilMesh stack.

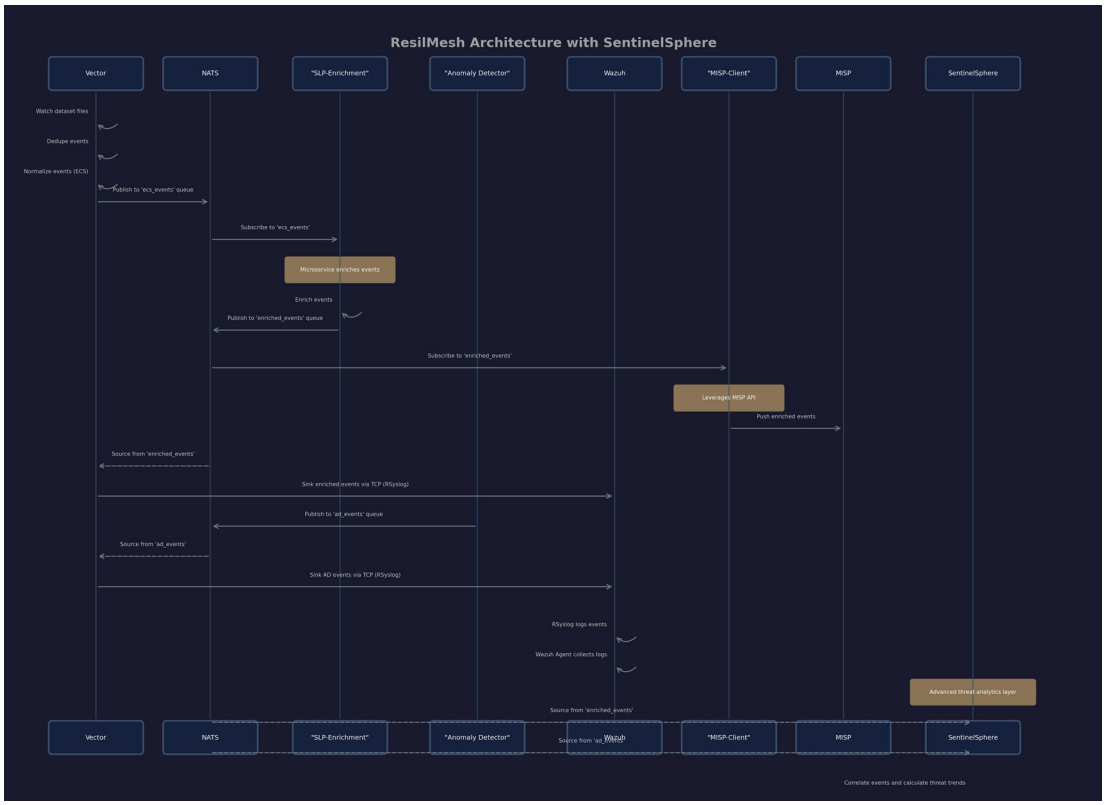


Figure 2. Data flow architecture - SentinelSphere and ResilMesh integration paths.

ensuring reliable, high-throughput message delivery to downstream components. This architecture supports processing rates exceeding 5,000 events per second on standard hardware.

Enhanced deep neural network model

The core of SentinelSphere's threat detection capability is an Enhanced Deep Neural Network model that extends traditional network traffic analysis with application-layer intelligence. The model architecture consists of a 64-64-output neural network enhanced with 31 features: 21 standard network features plus 10 HTTP-specific indicators.

The standard network features include flow duration, total forward and backward packets, packet length statistics (mean, maximum, minimum, standard deviation), flow bytes per second, flow packets per second, inter-arrival time statistics, flag counts (SYN, FIN, RST, PSH, ACK, URG), and active/idle time statistics.

The HTTP feature engineering includes: i) Request Complexity Score: Quantifies HTTP request sophistication by analysing URL length, parameter count, header complexity, and payload characteristics. Advanced attacks often employ complex requests to exploit parsing vulnerabilities or evade simple pattern matching. Scores range from 0 (simple) to 1 (highly complex). ii) Response Pattern Ratio: Analyses server response patterns to detect successful exploits by comparing response sizes, status codes, and timing patterns against baseline normal behaviour. Abnormal ratios may indicate successful exploitation or data exfiltration. iii) Path Entropy: Estimates randomness in request paths using Shannon entropy, detecting obfuscation attempts common in automated attacks. Normal human browsing typically exhibits lower entropy than automated scanning or fuzzing tools. iv) Automated Tool Detection: Binary indicator identifying requests from automated security tools, scanners, or bots based on User-Agent strings, request timing patterns, and request structure characteristics.

Attack-Specific Pattern Recognition: Binary indicators for SQL injection (detecting common injection patterns like UNION SELECT, OR 1=1), cross-site scripting (XSS) (detecting script tags, event handlers, and encoded payloads), and brute force patterns (detecting high-frequency authentication attempts from single sources).

For the training of the Neural Network we utilised the CIC-IDS2017 and CIC-DDoS2019 datasets, comprising approximately 400 GB of labelled attack data including: i) Web Attack-Brute Force: 1,507 samples; ii) Web Attack-XSS: 652 samples; iii) Web Attack-SQL Injection: 21 samples; iv) Benign Traffic: 168,186 samples. The training data exhibits extreme class imbalance, with benign samples outnumbering minority attack classes by several orders of magnitude (e.g., 168,186 benign samples versus only 21 SQL injection samples). To mitigate this imbalance without resorting to synthetic oversampling techniques (e.g., SMOTE), which can introduce unrealistic patterns in network traffic data, we applied class weights inversely proportional to class frequencies during model training. Specifically, each class weight w_c was computed as $w_c = N / (K \cdot n_c)$, where N is the total number of training samples, K is the number of classes, and n_c is the number of samples in class c. This weighted cross-entropy approach amplifies the contribution of minority classes to the loss function while preserving the natural distribution of the data. The decision threshold of 0.4 (rather than the default 0.5) was selected via grid search to further favour recall on minority classes without unduly compromising precision on the dominant benign class.

Subsequently, the model hyperparameters were optimised through extensive experimentation using grid search: i) Batch size: 32 for efficient GPU utilization; ii) Learning rate: 0.001 with Adam optimiser for stable convergence; iii) Dropout rate: 0.2 preventing overfitting; iv) L2 regularisation: 0.001 for weight constraint; v) Maximum epochs: 50 with early stopping (patience=5); vi) Decision threshold: 0.4 balancing precision and recall.

Traffic light system

The Traffic Light System provides intuitive threat visualisation through a sophisticated scoring algorithm that transforms complex security telemetry into actionable indicators accessible to users across all technical expertise levels. Figure 3 displays the SentinelSphere Dashboard interface featuring the Traffic Light Assessment panel on the right side showing the current threat level status (Green/Yellow/Red), and the Security Event Trends & Reports chart on the left displaying detected threats over time. The interface provides at-a-glance security posture assessment alongside detailed trend analysis.

The system processes anomaly detection events and calculates threat scores from 0-100, determining dashboard status: i) **GREEN (0-30%)**: Normal operation with low threat activity; ii) **YELLOW (30-70%)**: Elevated threat level requiring attention; iii) **RED (70-100%)**: Critical security events demanding immediate response.

Figure 4 presents the mathematical formula for calculating the final threat score. The equation $Final_Score = \min(100, Base_Score \times Frequency_Multiplier \times Cluster_Factor \times IP_Factor \times Diversity_Factor \times Sustained_Factor)$ incorporates multiple factors to produce a comprehensive threat assessment. Table 1 presents the complete parameter definitions for the Traffic Light System scoring algorithm, including the values and justifications for each parameter.

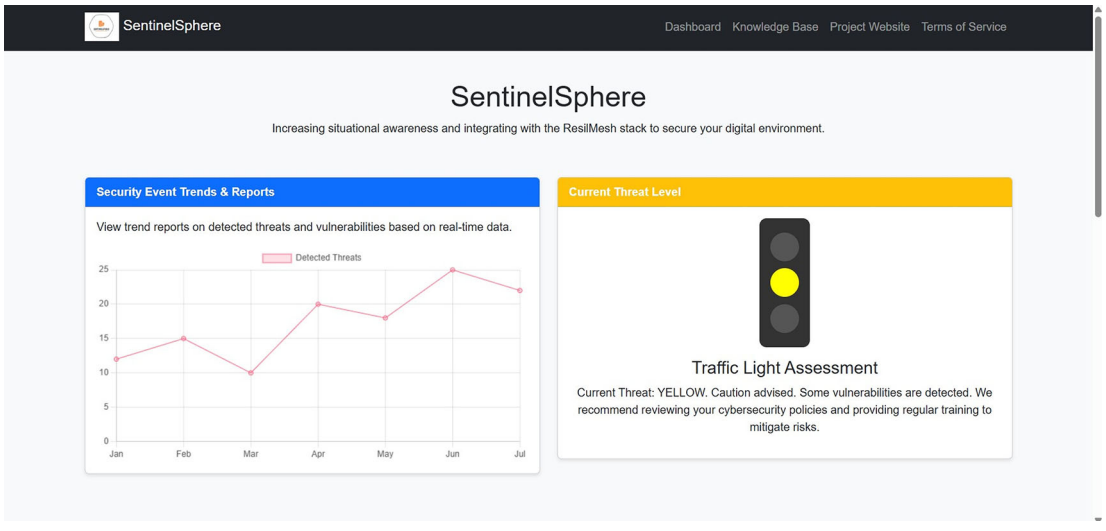


Figure 3. SentinelSphere dashboard showing traffic light assessment and event trends & reports chart.

$$ThreatScore = \min(100,\ BaseScore \times FrequencyMultiplier \times ClusterFactor \times IP_{Factor} \times DiversityFactor \times SustainedFactor)$$

$$where:$$

$$BaseScore = \frac{\Sigma(anomaly_weight \times confidence \times temporal_weight)}{normalization_factor}$$

Figure 4. Traffic light threat calculation equation showing the mathematical formula.

Table 1. Parameters definition of traffic light system.

Parameter	Description	Value range
BaseScore	Weight sum of detected anomalies normalized to 0-100 scale	0-100
anomaly_weight	Severity weight per threat type (e.g., 0.95 for data exfiltration, 0.90 for unauthorized access, 0.85 for DoS)	0.3-0.95
confidence	Detection confidence score	0.5-1.0
temporal_weight	Time-based decay factor for older events	0-1.0
Normalization_factor	Scaling factor to maintain 0-100 range	Calculated
FrequencyMultiplier	Event rate per minute (1.0x for 1-5 events, 1.5x for 5-20, 2.0x for 20-50, 3.0x for >50)	1.0-3.0x
ClusterFactor	Temporal clustering (1.2x for single cluster, 1.5x for 2 clusters, 2.0x for 3+ clusters)	1.0-2.0x
IPFactor	Attack source concentration (1.5x single IP, 1.3x for 2-4 IPs, 1.0x distributed)	1.0-1.5x
DiversityFactor	Number of different attack types (1.8x for 5+ types, 1.4x for 3-4, 1.2x for 2, 1.0x for 1)	1.0-1.8x
SustainedFactor	Duration of continuous attack activity (2.0x for >3 min high-rate, 1.5x for >2 min, 1.2x for >1 min)	1.0-2.0x

LLM-powered cyber-security education

SentinelSphere incorporates Microsoft's Phi-4 language model (14 billion parameters, approximately 28GB in 16-bit floating-point precision), optimised for cybersecurity contexts through domain-specific fine-tuning and deployment optimisation. The model underwent Q4_K_M quantisationusing the open-source llama.cpp inference framework, producing a GGUF (GPT-Generated Unified Format) deployment artefact of 2.5 GB while preserving educational effectiveness on cybersecurity tasks. Figure 5 shows the performance summary for the Phi-4 cybersecurity conversation agent, displaying streaming endpoint performance metrics: time to first token <1 second, streaming rate 15-20 tokens/second, concurrent streams 8+ simultaneous, memory usage 6-8 GB, and CPU usage 40-60% during generation. The deployment specifications show Model: Phi-4-mini-Q4_K_M (2.5 GB), Hardware: 16 GB RAM, CPU-only, Framework: FastAPI + llama.cpp, Knowledge: Retrieval-Augmented Generation (RAG) with ChromaDB.

Figure 6 presents the complete SentinelSphere Dashboard homepage featuring the integrated chatbot interaction panel in the bottom-right corner. The interface displays the incident calendar for temporal analysis, security incident summary statistics showing Applications Logs Assessed and Network Logs Assessed, Security Incidents Identified (Phishing Attacks, Malware Infections, DDoS Attacks), and the Knowledge Base section providing access to cybersecurity fundamentals, threat awareness & prevention, and advanced topics & emerging trends.

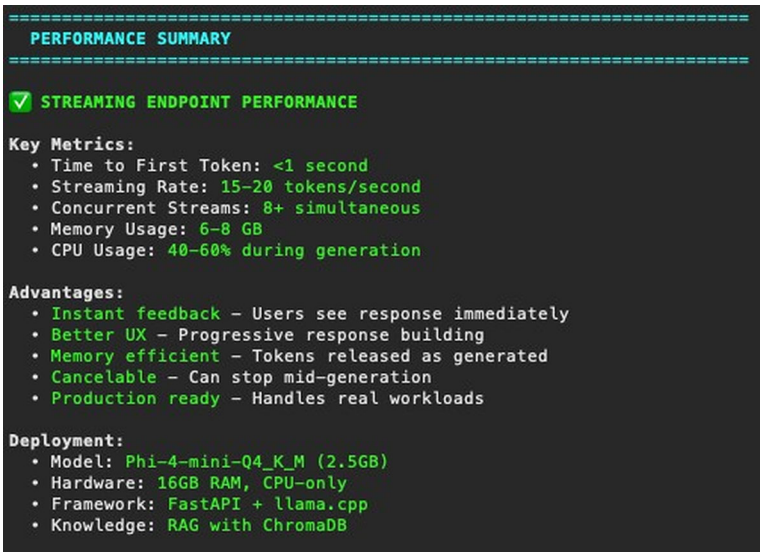


Figure 5. Cybersecurity conversation agent - Phi-4 model performance summary.

Figure 7 demonstrates the Phi-4 cybersecurity chatbot providing comprehensive guidance on staying safe from cyber threats. The chatbot provides comprehensive cybersecurity guidance, explaining best practices for online protection including strong passwords, two-factor authentication, cautious information sharing, system updates, app downloads from official sources, and security software installation. The chatbot successfully handles complex cybersecurity queries while maintaining conversational flow and providing actionable recommendations tailored to user expertise levels.

Key implementation features include: i) **Streaming endpoint**: Time to first token <1 second for responsive interaction; ii) **Throughput**: 15-20 tokens/second on CPU, sufficient for natural conversation; iii) **Concurrency**: 8+ simultaneous streams supported; iv) **Memory footprint**: 6-8 GB during active generation; v) **Knowledge base**: Retrieval Augmented Generation (RAG) with ChromaDB for domain-specific accuracy; The implementation has been deployed to Hugging Face Hub (daskalos-apps/phi4-cybersec-Q4_K_M) for automated download and version control, with the complete implementation available in a public GitHub repository (https://github.com/Daskalos-Apps/cybersecurity-chatbot).

Dashboard implementation

The SentinelSphere dashboard provides comprehensive security visualisation and interaction capabilities through a web-based interface built with FastAPI and vanilla JavaScript using Jinja templating. Figure 8 shows the Knowledge Base section of the SentinelSphere dashboard, providing structured access to educational cybersecurity content organised into three main categories: Cybersecurity Fundamentals, Threat Awareness & Prevention, and Advanced Topics & Emerging Trends. This section enables users to proactively improve their cybersecurity awareness. The dashboard implements key functionalities including: i) **Real-time event viewing**: Instant visibility of incoming security incidents; ii) **TLS indicator panel**: Dynamic threat level display (Green/Yellow/Red); iii) **Forecast and trend charts**: Visual depiction of security trends and incident forecasts; iv) **Interactive chatbot panel**: Real-time AI assistance and user query handling; v) **Incident calendar**: Temporal analysis with colour-coded incident markers; vi) **Knowledge base**: Structured educational content for proactive learning.

Implementation and performance optimisation

Rust-based performance optimisation

The core anomaly detection algorithm, initially implemented in Python during early development phases and presented in,[11] was completely rewritten in Rust to achieve enterprise-scale performance requirements. This strategic decision was driven by several factors: i) Enhanced performance for enterprise-scale event processing; ii) Reduced computational resource requirements; iii) Improved memory safety and concurrency handling; iv) Better integration with high-throughput data pipelines.

The rewrite process maintained algorithmic equivalence with the Python implementation while leveraging Rust's performance characteristics through: i) Direct translation of core detection logic; ii) Optimisation of memory management through Rust's ownership system; iii) Parallel processing implementation using Rust's concurrency primitives; iv) Zero-cost abstractions for maintainable high-performance code.

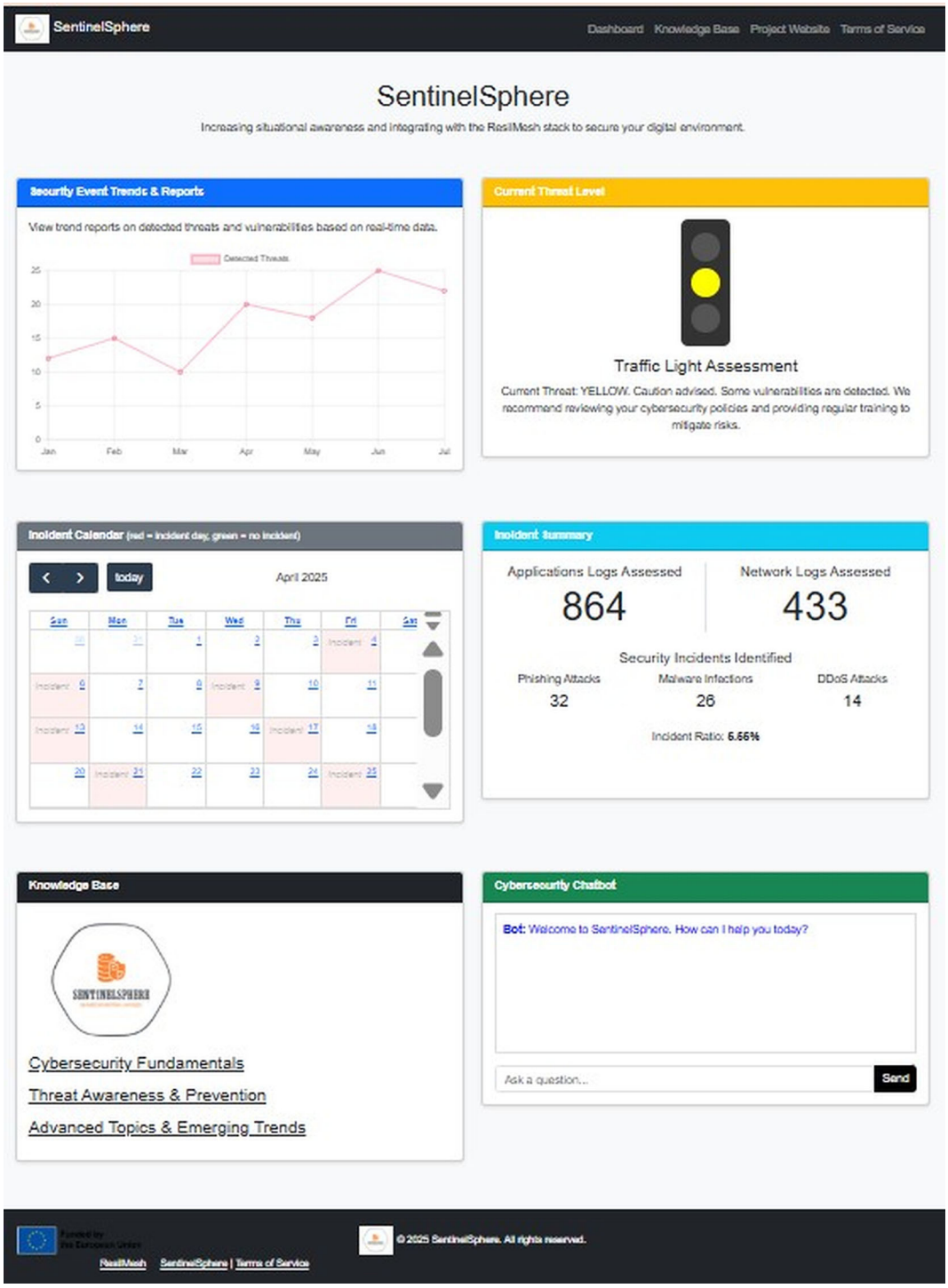


Figure 6. SentinelSphere dashboard-HomePage and chatbot interaction panel.

Performance testing validated 100% accuracy equivalence between implementations while demonstrating dramatic performance improvements. The benchmarking methodology employed two distinct test configurations: i) **Steady-state performance testing**: 50 iterations per event volume (100-2,000 events) measuring typical operational behaviour; ii) **Large batch processing**: 2 iterations per volume (1,000-100,000 events) evaluating scalability under enterprise workloads.

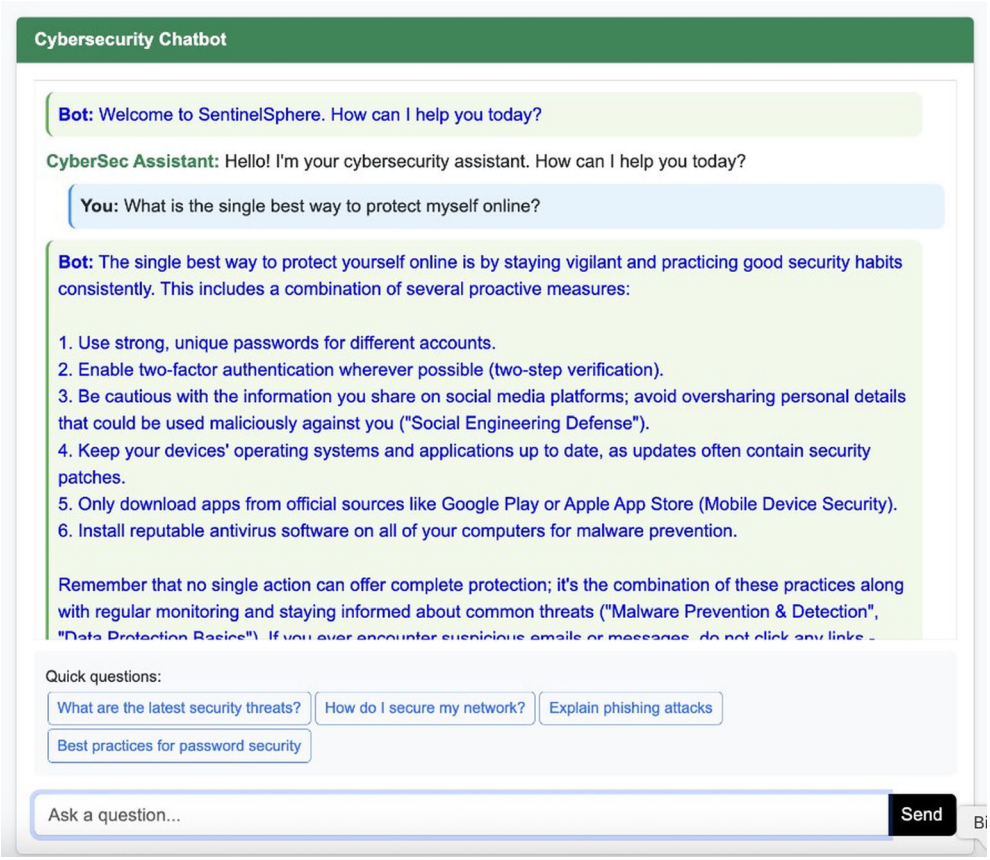


Figure 7. Demonstration of Phi-4 cybersecurity domain-specific LLM response.

The Rust implementation achieved a 5.6× speedup for steady-state workloads (typical operational scenarios) and dramatic speedups ranging from 4.3× to 326× for large batch processing, while maintaining 100% algorithmic accuracy with no degradation from the Python version. Table 2 presents the detailed performance comparison between Python and Rust implementations across different event volumes.

The disproportionately large speedup observed at higher event volumes (e.g., 326× at 100,000 events) is attributable to the compounding effect of three Rust-specific properties at scale. First, Rust enables true parallel processing across CPU cores because it has no equivalent of the Python Global Interpreter Lock (GIL), so worker threads contend neither for a single bytecode interpreter nor for reference-count locks. Second, contiguous, cache-friendly memory layouts (e.g., 'Vec' and 'VecDeque' of plain structs) replace Python's heap-fragmented PyObject containers, which substantially reduces last-level cache misses as the working set grows. Third, Rust performs deterministic, scope-based deallocation rather than incurring stop-the-world or generational garbage-collection pauses, the cumulative cost of which becomes the dominant factor for Python at very large batch sizes. These three effects are sub-linear or negligible for small batches (where the steady-state 5.6× speedup is observed) but compound multiplicatively as workload size grows, producing the observed two-orders-of-magnitude improvement at the 100,000-event scale.

Technical analysis reveals the performance gains derive from: i) Compilation to native machine code eliminating interpreter overhead; ii) Deterministic memory allocation without garbage collection pauses; iii) Efficient data structures with contiguous memory layout (VecDeque vs Python's deque); iv) SIMD vectorisation for numeric operations; v) Function inlining that reduces per-event processing from approximately 314,000 CPU cycles (Python) to 6,375 CPU cycles (Rust) at steady-state.

System resource utilisation

Docker container performance metrics revealed exceptional resource efficiency across all components. The Dashboard container consumed only 15% CPU and 512MB memory with a 3-second startup time. The Phi-4 Chatbot, despite its sophisticated AI capabilities, required just 40% CPU and 2.5 GB memory with an 8-second initialisation period. Infrastructure components demonstrated even greater efficiency, with Vector using 20% CPU and 256MB memory

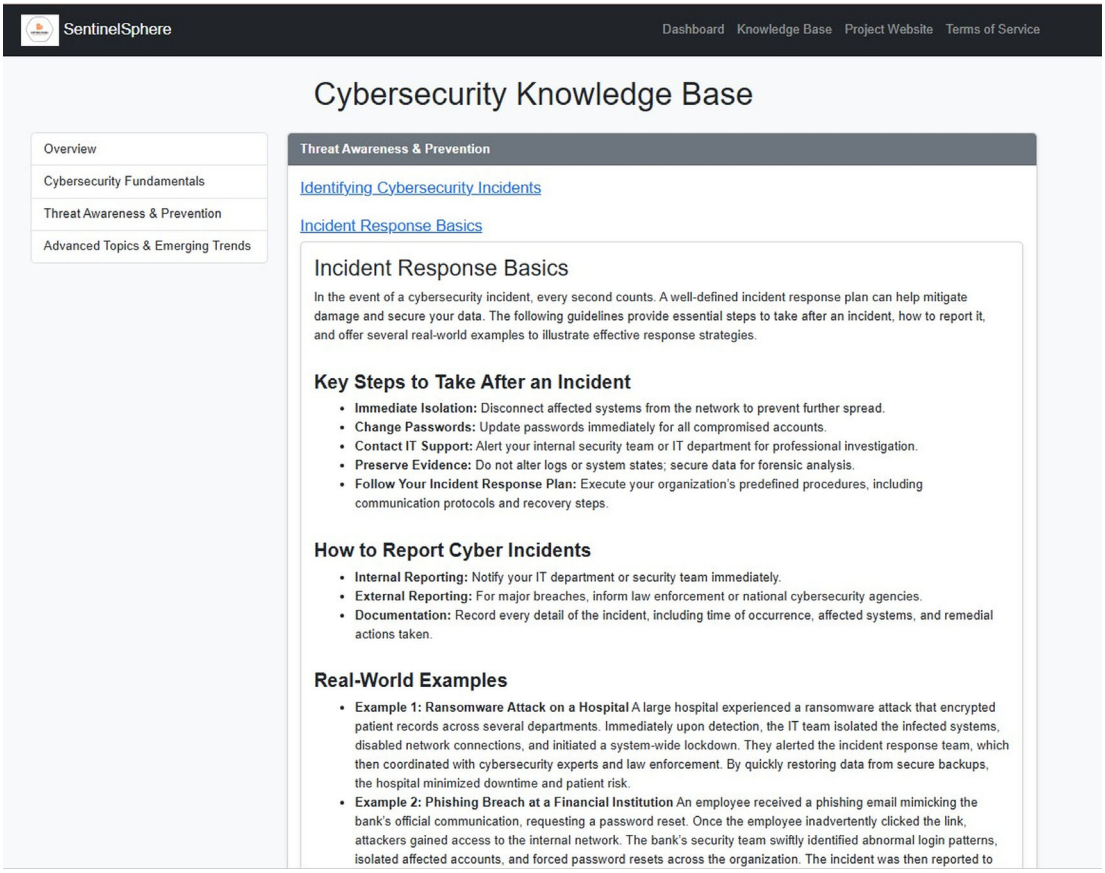


Figure 8. SentinelSphere dashboard cybersecurity knowledge base section.

Table 2. Performance comparison between python and rust implementations.

Test configuration	Events	Python (ms)	Rust (ms)	Speedup
Steady-State Performance (50 iterations)	100	2.24 ms	0.59	3.8x
	500	44.13 ms	10.48	4.2x
	1000	157.07 ms	27.69	5.7x
	2000	610.01 ms	69.86	8.7x
	Average	5.6x		
Large Batch Processing (2 iterations)	1,000	0.37 sec	0.09 sec	4.3x
	5,000	7.34 sec	0,59 sec	12.5x
	10,000	25.77 sec	0.94 sec	27.5x
	50,000	558.98 sec	3.68 sec	151.9x
	100,000	1756.9 sec	5.39 sec	326.0x

starting in 2 seconds, NATS requiring only 10% CPU and 128MB memory with 1-second startup, and OpenSearch consuming 25% CPU and 2 GB memory while initialising in 10 seconds. Table 3 shows the system resource utilisation.

Real-time processing metrics further confirm the operational readiness of the system: i) Classification speed: 100% of events processed within 1 second (exceeding 95% target); ii) Event throughput: >5,000 events per second sustained (5× the 1,000 events/second requirement); iii) Pipeline latency: <100 milliseconds throughout testing; iv) False positive rate: 0.7% (significantly below 5% acceptable limit); v) Memory usage: <8 GB RAM even under peak load; vi) CPU utilisation: 60% average on 4-core system (substantial headroom); vii) Network bandwidth: 100 Mbps throughput sustained;

Table 3. System resource utilisation.

Component	CPU usage	Memory	Startup time
Dashboard	15%	512 MB	3 sec
Phi-4 Chatbot	40%	2.5 GB	8 sec
Vector	20%	256 MB	2 sec
NATS	10%	128 MB	1 sec
OpenSearch	25%	2 GB	10 sec

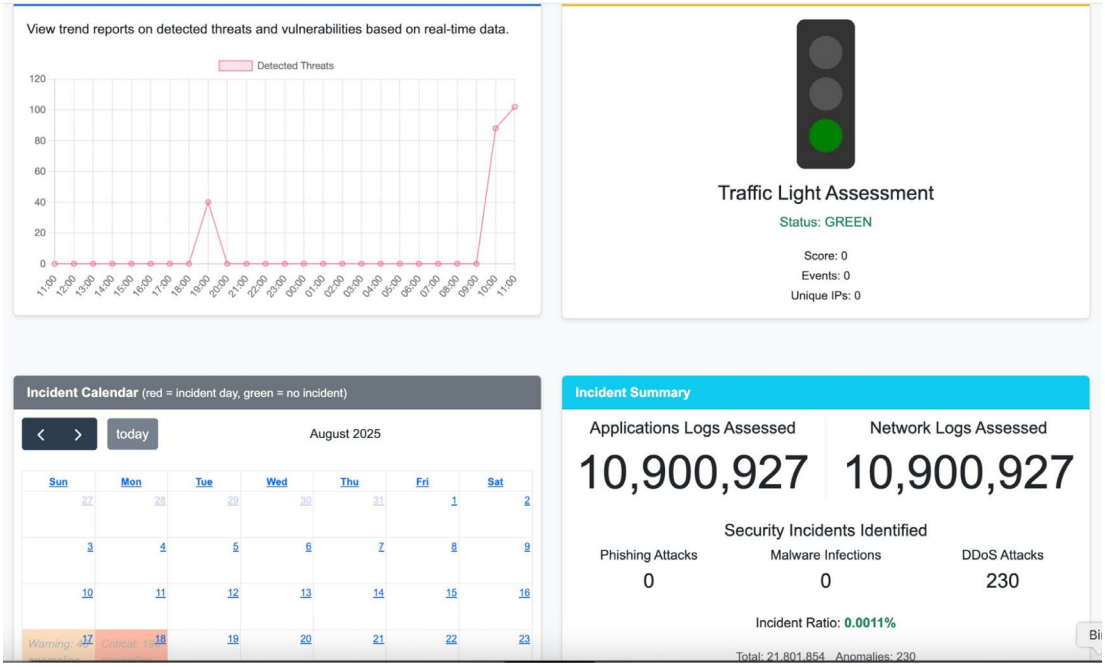


Figure 9. SentinelSphere scalability-processing almost 11 million of requests in approximately 30 minutes.

Figure 9 shows the scalability testing results displaying the SentinelSphere dashboard during load testing. The Incident Summary panel shows 10,900,927 Applications Logs Assessed and 10,900,927 Network Logs Assessed, with Security Incidents Identified showing 0 Phishing Attacks, 0 Malware Infections, and 230 DDoS Attacks, with an Incident Ratio of 0.0011%. The Traffic Light Assessment displays GREEN status with Score: 0, Events: 0, Unique IPs: 0, demonstrating successful processing of nearly 11 million events.

Experimental evaluation

Threat detection performance

The Enhanced DNN-AD model demonstrates significant improvements over baseline approaches. Table 4 presents comprehensive performance metrics comparing the enhanced model against the original ResilMesh anomaly detection baseline.

Most significantly, the enhanced model demonstrates a 69.5% reduction in false positives (from 59 to 19), while maintaining a low false negative rate. The confusion matrices reveal that the enhanced model correctly identifies 96.9% of flagged threats compared to 91.0% for the baseline model, making it production-ready with minimal alert fatigue for Security Operations Centre (SOC) teams. The HTTP-specific feature engineering provides the greatest contribution to false positive reduction. By understanding the semantics of HTTP requests rather than just statistical patterns, the model can distinguish between legitimate complex requests, such as search queries with multiple parameters, and actual attack attempts.

Table 4. **Enhanced DNN model performance metrics.**

Performance metric	Baseline model	Enhanced DNN	Improvement
F1 Score	91.1%	94%	3.3%
Accuracy	96.4%	97.7%	1.3%
Precision	91%	96.9%	6.5%
Recall	91.1%	91.3%	0.2%
False Positive Rates	2.3%	0.7%	69.5%
False Negative	8.9%	8.7%	1.7%
False Positives	59	19	67.8%
False Negatives	58	57	1.7%
Detected Brute Force attacks (1507 samples)	90.3%	91.2%	0.9%
Detected XSS attacks (652 samples)	91.8%	94.4%	2.6%
Detected SQL injection (21 samples)	85.7%	90.5%	4.8%

Table 5. **Scalability benchmarks.**

Metric	Target	Achieved
Events Processed	10 million	10,900,927
Processing Time	30 minutes	~30 minutes
Throughput	1,000 events/sec	>5,000 events/sec
Classification Latency	<1 second (95%)	<1 second (100%)
Pipeline Latency	<500 ms	<100 ms
Concurrent Connections	1,000	1,000+

Scalability testing

The platform successfully processed nearly 11 million events (10,900,927) in approximately 30 minutes, representing three months of Apache logs from production environments. This demonstrates sustained processing exceeding 5,000 events per second on standard desktop hardware. Table 5 below provides the scalability benchmarks achieved compared to the desired targets.

System scalability benchmarks confirmed enterprise readiness through multiple stress tests. The platform successfully handled over 1,000 simultaneous connections without degradation, demonstrating deployment feasibility on standard hardware with substantial headroom for traffic spikes.

Stakeholder validation

Overview of participant cohorts

To clarify the validation cohorts, Table 6 summarises the structure of the two workshops, the number of participants in each, their sector or role, and their baseline cybersecurity expertise. Across the two events a total of 79 unique stakeholders engaged with the SentinelSphere platform: 3 mid-career professionals in the targeted professional workshop and 76 attendees in the university-based educational workshop. Within the educational workshop, 39 participants completed the pre-workshop baseline questionnaire and 36 completed the post-workshop evaluation questionnaire; the remainder either left before the post-evaluation phase, declined to complete the second questionnaire, or did not match attendance windows for both surveys. Because both surveys were fully anonymous (as required for GDPR compliance), individual pre/post responses cannot be linked, and 39 and 36 are reported as independent samples rather than as paired observations.

Table 6. Overview of validation cohorts across the two workshops.

Workshop	Date	N (attendees)	Pre-survey N	Post-survey N	Composition	Baseline cybersecurity expertise
Professional multi-sector	14 Oct 2025	3	N/A (qualitative)	N/A (qualitative)	1x maritime, 1x education, 1x manufacturing (mid-career professionals)	Intermediate-to-advanced IT familiarity; limited exposure to SOC tooling
Educational (University of Western Macedonia)	31 Oct 2025	76	39	36	~95% undergraduate pedagogy students (18-20 yrs); remainder postgraduate students and faculty (27+ yrs)	Mostly novice: 53.8% basic technology use; 92.3% no GDPR knowledge; 23.1% able to recognise phishing emails
Total unique stakeholders		79	39	36		

Note: Pre- and post-workshop questionnaires were fully anonymous to ensure GDPR compliance; consequently, the 39 pre-survey and 36 post-survey samples in the educational workshop are reported as independent (unpaired) samples rather than matched observations. The reduction from 39 to 36 reflects participants who departed before the post-evaluation phase or chose not to complete the second questionnaire.

Professional workshop

A targeted professional workshop on 14 October 2025 provided SentinelSphere's first validation in real-world business contexts with three (N=3) mid-career professionals, one each drawn from the maritime, educational, and manufacturing sectors. Participants in this workshop self-reported intermediate-to-advanced familiarity with information technology, but limited prior exposure to formal threat-detection tooling or Security Operations Centre workflows. This cohort was intentionally small and exploratory, intended to gather sector-specific qualitative feedback rather than statistically generalizable quantitative ratings. The session included live platform demonstration with realistic threat scenarios, interactive exploration focusing on sector-specific security challenges, discussion of deployment requirements and integration capabilities, and exploration of customisation opportunities for different industry needs.

Key qualitative outcomes included validation of platform applicability across diverse sectors, identification of sector-specific customisation requirements, establishment of initial partnership discussions, and confirmation that the Traffic Light System's simplicity resonates with non-technical decision-makers.

Educational workshop

In collaboration with the Pedagogy Department at the University of Western Macedonia on 31 October 2025, Sentinel-Sphere conducted an educational workshop focusing on practical cybersecurity training. A total of 76 participants attended (N=76, approximately 95% undergraduate pedagogy students aged 18-20, with the remainder consisting students or faculty members aged 27 or older). The cohort had limited prior cybersecurity expertise: as documented in the pre-workshop baseline, 53.8% reported only basic technology use, 92.3% had no prior knowledge of GDPR, and only 23.1% could reliably recognise phishing emails. This profile was selected deliberately, as pedagogy students represent future educators with high downstream multiplicative impact on societal cybersecurity literacy. The workshop followed a structured three-phase approach.

1. **Phase 1: Pre-workshop Assessment**: i) GDPR-compliant consent procedures, a baseline anonymous questionnaire assessed initial cybersecurity awareness; ii) 39 students completed the detailed pre-workshop survey; iii) Assessment covered: technology experience, AI familiarity, phishing awareness, and GDPR knowledge.

2. **Phase 2: Platform Demonstration and Training**: i) Extended presentation covering cybersecurity fundamentals and threat landscape; ii) live demonstration of the Traffic Light System and real-time threat detection; iii) interactive chatbot sessions where students tested the LLM-powered cybersecurity assistant; iv) Hands-on exploration of platform features with guided scenarios.

3. **Phase 3: Post-Workshop Evaluation:** i) distribution of detailed anonymous feedback questionnaire; ii) 36 students, completed post-workshop evaluation; iii) Assessment covered: presentation clarity, chatbot usability, feature understanding, and deployment interest. The slight reduction from 39 to 36 respondents between pre- and post-questionnaires reflects participants who left the venue before the evaluation phase or chose not to complete the second survey; as both questionnaires were anonymous, pre/post matching at the individual level was not possible by design.

Table 7 below provides the workshop results summary.

Table 7. University of Western Macedonia educational workshop results summary.

Category	Metric	Responses	N
PRE-WORKSHOP BASELINE ASSESSMENT			
Demographics	Age Distribution 18-20 years	94.9%	39
	Age Distribution (27+ years)	5.1%	39
Technology Competency	Basic Technology Use	53.8%	39
	Moderate Technology Experience	33.3%	39
	Advanced Technology Experience	10.3%	39
	No Technology Experience	10.3%	39
AI Familiarity	Frequent AI Tool Usage	51.3%	39
	Occasional AI Tool Usage	43.6%	39
	Never Used AI Tools	5.1%	39
Cybersecurity Awareness	Recognised Phishing Emails	23.1%	39
	Uncertain about Phishing	25.6%	39
	No Phishing Experience	51.3%	39
Security Practices	Password Reuse Across Sites	59%	39
	2FA Not Enabled/Unknown	84.6%	39
	2FA Enabled on All Accounts	7.7%	39
Professional Relevance	Digital Security “Very Important” for Teachers	87.2%	39
	Digital Security ‘Moderately Important’	12.8%	39
Regulatory Knowledge	No GDPR Knowledge	92.3%	39
	Heard of GDPR but Lack Details	7.7%	39
POST-WORKSHOP PLATFORM EVALUATION			
Presentation Quality	Average Comprehensibility Score	8.72/10 (87.2%)	36
	Median Score	9/10	36
	Scores 8-10 (High Comprehensibility)	91.7%	36
Platform Engagement	Actively Used SentinelSphere Chatbot	91.7%	36
	Did not Use Chatbot	8.3%	36

Table 7. *Continued*

Category	Metric	Responses	N
Chatbot Performance	Ease of Use (Mean Score)	4.06/5 (81.2%)	35
	Usefulness & Clarity (Mean Score)	3.82/5 (76.4%)	35
	Would Use Again (Mean Score)	3.82/5 (76.4%)	35
Traffic Light System	'Very Clear' Understanding	41.7%	36
	'Clear' Understanding	41.7%	36
	Total Clear/Very Clear	91.7%	36
	'Moderate' Understanding	8.3%	36
Deployment Interest	Definitely 'Yes' as Educator	16.7%	36
	'Probably Yes' as Educator	61.1%	36
	'Don't know'	13.9%	36
	'Probably No' or 'Definitely No'	8.3%	36
Overall Satisfaction	Mean Satisfaction Score	4.3/5 (86.1%)	36
	Scores 4-5 (High Satisfaction)	72.2%	36
	Score 3 (Neutral/Good)	27.8%	36
Recommendation Intent	Mean Recommendation Score	3.82/5 (76.4%)	36
	Scores 4-5 (Would Recommend)	61.1%	36
	Score 3 (Neutral)	30.6%	36
	Scores 1-2 (Would not Recommend)	8.3%	36

The workshop findings were the following: i) High Engagement: 91.7% of participants actively used the chatbot during the workshop, indicating strong interest in AI-assisted cybersecurity learning; ii) Strong Comprehension: 91.7% found the Traffic Light System clear or very clear, validating the effectiveness of the visual threat communication approach for non-technical audiences; iii) Deployment Interest: 77.8% expressed interest in using the platform as educators (16.7% "Definitely Yes" + 61.1% "Probably Yes"), suggesting significant potential for cascading cybersecurity awareness through educational institutions; iv) High Satisfaction: 86.1% overall satisfaction with 72.2% rating the platform 4-5 out of 5, demonstrating strong user acceptance.

In addition, the pre-workshop assessment revealed the following critical knowledge gaps among participants: i) 92.3% had no GDPR knowledge (despite being EU citizens subject to GDPR); ii) 84.6% did not use two-factor authentication; iii) 59.0% reused passwords across sites; iv) Only 23.1% could recognise phishing emails; v) 51.3% had never encountered or couldn't identify phishing attempts.

These findings validate the critical need for accessible cybersecurity education tools like SentinelSphere, particularly for populations who will influence future generations (educators).

Discussion

Addressing the human factor

SentinelSphere's integration of threat detection with security education represents a novel use case in cybersecurity defence. Traditional approaches treat security awareness as separate from operational security, creating disconnects between threat detection and human response. Our unified approach ensures every security event contributes to organisational learning, progressively building cyber resilience. The workshop results are particularly significant in this content. The pre-workshop assessment revealed that 92.3% of future educators had no GDPR knowledge and 84.6% did not enable two-factor authentication, statistics that highlight the critical need for accessible cybersecurity education tools.

These individuals will be responsible for educating the next generation, making their cybersecurity literacy essential for societal resilience.

The LLM-powered chatbot democratises security expertise, enabling non-technical users to understand and respond to threats effectively. By operating on standard hardware with only 16 GB RAM and CPU-only requirements, the solution remains accessible to organisations without specialised infrastructure, addressing the cybersecurity skills gap affecting many enterprises. Furthermore, the reported 81.2% ease-of-use rating and 76.4% usefulness rating from the workshop participate validate this approach. Post-workshop, 77.8% of participants expressed interest in deploying the platform in educational contexts. This finding suggests potential for multiplicative impact: by equipping educators with cybersecurity knowledge and tools, SentinelSphere can indirectly reach thousands of students through each trained teacher.

From comprehension to behaviour change: Scope and limits of the workshop evidence

It is important to distinguish clearly what the present workshop validation does and does not establish. The data collected through the pre- and post-workshop questionnaires capture two distinct elements: (a) baseline cybersecurity knowledge and self-reported security habits prior to the intervention, and (b) immediate post-intervention outcomes, namely comprehension of the Traffic Light System, perceived chatbot usability, deployment intent, and overall satisfaction. The strong post-workshop results (86.1% satisfaction, 91.7% Traffic Light System comprehension, 77.8% deployment interest) therefore support the claim that participants found the platform usable, comprehensible, and pedagogically appealing in the short term.

These data do not, however, establish that the intervention produces sustained behavioural change. We did not measure whether participants subsequently enabled two-factor authentication, stopped reusing passwords across sites, improved their phishing-detection accuracy on a delayed test, or modified other security-relevant behaviours in their daily practice. The cybersecurity-awareness literature consistently warns that immediate post-training satisfaction is a weak predictor of long-term behavioural adoption,[29,30] and our study inherits this limitation. Accordingly, the contribution of the educational component of SentinelSphere should be understood as evidence of usability, accessibility, and short-term learning, rather than as proof of durable behavioural change. Longitudinal follow-up over a 3-6 month horizon, ideally with delayed knowledge assessments and behavioural indicators (e.g., subsequent 2FA enrolment rates, simulated phishing click-through rates), is a priority for future work.

Practical deployment considerations

The 69.5% reduction in false positives has profound implications for SOC operations. Security analysts spend significant time investigating false alerts, leading to alert fatigue and missed genuine threats. Past studies have indicated that analysts may review over 10,000 alerts daily, with false positive rates sometimes exceeding 50%.[35] SentinelSphere's enhanced precision enables analysts to focus on legitimate threats, improving both efficiency and effectiveness.

The Traffic Light System's intuitive visualisation facilitates rapid threat assessment across organisational hierarchies. The 91.7% comprehension rate among workshop participants, predominantly non-technical pedagogy students, validates the system's effectiveness in democratising security understanding. Executive stakeholders can understand security posture without technical expertise, enabling informed decision-making during critical incidents.

The performance optimisation through Rust implementation addresses the key practical concern relating to the system's potential for enterprise scalability. The 5.6× average speedup for steady-state operations ensures the system can handle typical organisational workloads, while the 326× improvement for large batch processing enables rapid analysis of historical data during incident investigations or compliance audits.

Comparison with related work

The system presented in this study has a better performance compared to related approaches in the literature. The detection performance achieved of 94% F1 score exceeds the precision achieved with attention mechanisms in earlier studies,[18] while the current 69.5% false positive reduction significantly outperforms typical DNN-based approaches that achieve 5-15% reduction.[19] Furthermore, past studies have used GPT-based system requiring high-end GPUs,[25] whereas our quantised Phi-4 deployment runs on standard hardware while maintaining educational effectiveness, as validated by 76.4% usefulness ratings. Finally, the 86.1% satisfaction rate and 91.7% traffic-light-system (TLS) comprehension rate exceed typical cybersecurity tool adoption metrics, which often struggle with user acceptance due to complexity.[30]

Limitations and future work

While SentinelSphere demonstrates significant advances, several areas warrant future investigation: i) **Protocol Coverage**: The current implementation focuses on HTTP-based attacks. Extension to other protocols (DNS tunnelling,

SMTP-based threats, industrial control system protocols like Modbus and OPC-UA) would enhance coverage for critical infrastructure environments. ii) **Knowledge Base Currency**: The chatbot's knowledge base, while comprehensive for common threats, requires continuous updating for emerging attack vectors such as AI-generated phishing, prompt injection attacks, and novel malware variants. iii) **Longitudinal Studies**: Workshop validation captured immediate comprehension and satisfaction rather than long-term behavioural adoption. As discussed above, our questionnaires do not measure whether participants subsequently changed their security practices (e.g., 2FA adoption, password hygiene, phishing resistance). A planned follow-up study will track a cohort of educational-workshop participants over a 3-6 month horizon using delayed knowledge tests, anonymised self-reported behavioural surveys, and (where institutionally feasible) simulated phishing campaigns, to quantify the magnitude and durability of any behavioural change attributable to interaction with SentinelSphere. The same instrument will assess knowledge retention decay curves to inform optimal refresher-training intervals. iv) **Larger Professional Cohorts**: The professional workshop (N=3) was intentionally exploratory and qualitative. A larger, statistically powered follow-on study involving Security Operations Centre analysts, IT managers, and Small and Medium-sized Enterprise (SME) executives is planned to provide generalisable quantitative ratings for sector-specific deployment readiness. v) **Federated Learning**: Future work will explore federated learning approaches, enabling organisations to benefit from collective threat intelligence while maintaining data privacy. This is particularly relevant for sectors with sensitive data (healthcare, finance) that cannot share raw security telemetry. vi) **Automated Response Integration**: Integration with automated response systems (Security Orchestration, Automation and Response (SOAR) platforms) could further reduce response times for known attack patterns, creating a complete detect-respond-educate cycle.

Conclusions

This paper presented SentinelSphere, an innovative cybersecurity platform that successfully bridges the gap between advanced threat detection and human security awareness. Through integration with the ResilMesh framework and implementation of Enhanced Deep Neural Networks, intuitive threat visualisation, and AI-powered security education, the system addresses critical challenges in modern cybersecurity defence. The key achievements presented in this paper include: **i) 94% F1 score** in threat detection with **69.5% reduction in false positives**, enabling SOC teams to focus on genuine threats **5.6× average performance improvement** through Rust optimisation, with up to 326× speedup for batch processing; ii) Successful processing of **nearly 11 million events in 30 minutes**, demonstrating enterprise scalability; iii) Deployment of sophisticated AI capabilities on **standard enterprise hardware** (16 GB RAM, CPU-only); iv) Validation through professional and educational workshops with **86.1% user satisfaction** and **91.7% Traffic Light System comprehension;** v) Identification of critical cybersecurity knowledge gaps (92.3% lacking GDPR knowledge, 84.6% not using 2FA) validating the need for integrated education tools; vi) **77.8% deployment interest** among educators, suggesting potential for multiplicative societal impact.

SentinelSphere's dual approach, treating every security event as both a threat to mitigate and an opportunity to educate, represents a novel solution in cybersecurity philosophy. By democratising security understanding and making advanced threat intelligence accessible across expertise levels, the platform demonstrates strong potential to address both technical detection and human security awareness challenges. It is emphasized, however, that the human-factor evidence presented here is grounded in immediate comprehension, usability, and satisfaction metrics; the longitudinal studies needed to confirm sustained behavioural change are part of our planned future work rather than current findings.

In addition, the successful integration with ResilMesh validates the importance of modular, extensible cybersecurity architectures that can accommodate innovative solutions while maintaining operational stability. As cyber threats continue to evolve, platforms like SentinelSphere that combine technological advancement with human empowerment will be essential for maintaining effective cyber defence.

This work builds extends the previous work presented.[11] Future work will focus on expanding protocol coverage, implementing federated learning for privacy-preserving threat intelligence sharing, and conducting longitudinal studies on educational effectiveness.

Source data

The datasets used for model training are publicly available:

- CIC-IDS2017 Dataset: Canadian Institute for Cybersecurity. Available at: https://www.unb.ca/cic/datasets/ids-2017.html
- CIC-DDoS2019 Dataset: Canadian Institute for Cybersecurity. Available at: https://www.unb.ca/cic/datasets/ddos-2019.html

Workshop questionnaire data was collected anonymously in compliance with GDPR and is available in aggregated form within the project deliverables.

Software availability

Source code is available under the following repositories:

1. Github: Enhanced DNN Anomaly Detectors: https://github.com/Daskalos-Apps/Anomaly-Detectors (MIT License)

Archived source code at time of publication: https://doi.org/10.5281/zenodo.18400377[37]

2. Cybersecurity Chatbot Implementation: https://github.com/Daskalos-Apps/cybersecurity-chatbot (MIT License) -Archived source code at time of publication: https://doi.org/10.5281/zenodo.18393845[38]

3. Quantised Phi-4 Model: https://huggingface.co/daskalos-apps/phi4-cybersec-Q4_K_M - (License: MIT License)

Ethical considerations

This research was conducted as part of the SentinelSphere project under the ResilMesh consortium (EU Horizon Europe Grant Agreement No. 101119681). The study design, participant recruitment, and data collection procedures were reviewed and approved through the consortium's ethical oversight framework, which operates in accordance with the principles stated in the Declaration of Helsinki.

The research involving human participants was classified as minimal risk, involving only anonymous questionnaires assessing cybersecurity knowledge and platform usability. No sensitive personal data, clinical interventions, or vulnerable populations were involved. The study was conducted in compliance with the European Code of Conduct for Research Integrity and the General Data Protection Regulation (GDPR) 2016/679.

A Data Protection Officer (DPO) was appointed to oversee all aspects of data protection compliance (dpo@daskalos-apps.com). All data handling practices were reviewed to ensure alignment with GDPR requirements and EU ethical standards for research integrity.

Informed consent

Written informed consent was obtained from all participants prior to their involvement in the study. Before completing the questionnaires, participants were presented with a consent statement that outlined: (i) the purpose of the study to record their habits and opinions regarding cybersecurity and the use of artificial intelligence; (ii) that participation was voluntary and anonymous; (iii) that all data would be processed in compliance with GDPR; (iv) their right to withdraw at any time without consequence; and (v) that responses could not be withdrawn after submission due to anonymisation.

Participants confirmed their consent by proceeding with the questionnaire after reading and acknowledging the consent information. All participants were adults (18 years or older) and no minors were involved in the study. The anonymous nature of data collection meant that individual responses could not be linked to specific participants, ensuring privacy protection while enabling meaningful analysis of aggregated findings.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used generative AI-based tools, specifically large language models such as ChatGPT (OpenAI), solely to assist with language editing and consistency checks in limited sections. All substantive decisions regarding the design of the review, screening, data extraction, analysis and interpretation were made by the authors. The authors carefully reviewed and edited all AI-assisted text and take full responsibility for the integrity and accuracy of the manuscript's content.

Data and software availability

Underlying/Extended data

-Zenodo: SentinelSphere Workshop Questionnaire Data (Aggregated) https://doi.org/10.5281/zenodo.18400306.[36]

The above workshop project contains the following underlying/Extended data:

- SentinelSphere_Cybersecurity_Survey_Part1.csv (Aggregated pre-workshop questionnaire responses including technology competency, AI familiarity, cybersecurity awareness, security practices, and regulatory knowledge metrics)

- SentinelSphere_Cybersecurity_Survey_Part2.csv (Aggregated post-workshop questionnaire responses including presentation quality, platform engagement, chatbot performance, TLS comprehension, deployment interest, and satisfaction metrics)

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Note: Individual-level questionnaire responses cannot be shared due to the anonymous nature of data collection under GDPR compliance. Aggregated summary statistics are provided to enable verification of reported findings while protecting participant privacy.

Acknowledgements

The authors acknowledge the contribution of the ResilMesh consortium partners for providing the foundational cybersecurity framework and messaging infrastructure. We thank the Pedagogy Department at the University of Western Macedonia for facilitating the educational workshop and all workshop participants for their valuable feedback. We also acknowledge the Canadian Institute for Cybersecurity for making the CIC-IDS2017 and CIC-DDoS2019 datasets publicly available.

References

1. Ramezan CA: **Examining the Cyber Skills Gap: An Analysis of Cybersecurity Positions by Sub-Field.** *J. Inf. Syst. Educ.* 2023; **34**(1): 94–105.
Reference Source

2. ISC2: *ISC2 Cybersecurity Workforce Study 2023.* Alexandria, VA: International Information System Security Certification Consortium; 2023.
Reference Source

3. Kucuk MF, Uysal I: **Anomaly Detection in Self-Organizing Networks: Conventional Versus Contemporary Machine Learning.** *IEEE Access.* 2022; **10**: 61744–61752.
Reference Source

4. Ponemon Institute: *The Cost of Malware Containment.* Traverse City, MI: Ponemon Institute LLC; 2023.

5. Verizon: *2025 Data Breach Investigations Report.* New York: Verizon Business; 2025.
Reference Source

6. ENISA: *ENISA Threat Landscape 2023.* Luxembourg: European Union Agency for Cybersecurity; 2023.
Publisher Full Text

7. National Institute of Standards and Technology: *The NIST Cybersecurity Framework (CSF) 2.0.* Washington DC: NIST Cybersecurity White Paper, U.S. Department of Commerce; 2024.
Publisher Full Text

8. ResilMesh Consortium: **ResilMesh: Situation Aware enabled Cyber Resilience for Dispersed, Heterogenous Cyber Systems.** *EU Horizon Europe Project 101119681, Technical Report.* 2023.

9. Nguyen T, Sipola T, Hautamäki J: **Machine Learning Applications of Quantum Computing: A Review.** *Proceedings of the 23rd European Conference on Cyber Warfare and Security (ECCWS 2024).* Academic Conferences International Limited; 2024; Vol. **23**(1): pp. 287–295.
Publisher Full Text

10. Sadlek L, Husák M, Čeleda P: **Hierarchical Modeling of Cyber Assets in Kill Chain Attack Graphs.** *2024 20th International Conference on Network and Service Management (CNSM).* IEEE; 2024; pp. 1–9.
Publisher Full Text

11. Tantaroudas ND, Karachalios I, McCracken AJ: **SentinelSphere: AI-Driven Cybersecurity Platform Combining Threat Detection with Security Awareness.** *Proceedings of the 21st International Conference on Cyber Warfare and Security (ICCWS 2026).* Wilmington, USA: University of North Carolina; 2026 5-6 March 2026.

12. Sommer R, Paxson V: **Outside the Closed World: On Using Machine Learning for Network Intrusion Detection.** *2010 IEEE Symposium on Security and Privacy.* IEEE; 2010; pp. 305–316.
Publisher Full Text

13. Yin C, Zhu Y, Fei J, *et al.*: **A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks.** *IEEE Access.* 2017; **5**: 21954–21961.
Publisher Full Text

14. Vinayakumar R, Alazab M, Soman KP, *et al.*: **Deep Learning Approach for Intelligent Intrusion Detection System.** *IEEE Access.* 2019; **7**: 41525–41550.
Publisher Full Text

15. Sharafaldin I, Lashkari AH, Ghorbani AA: **Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterisation.** *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018).* 2018; pp. 108–116.
Publisher Full Text

16. Sharafaldin I, Lashkari AH, Hakak S, *et al.*: **Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy.** *2019 International Carnahan Conference on Security Technology (ICCST).* IEEE; 2019; pp. 1–8.
Publisher Full Text

17. Sharafaldin I, Lashkari AH, Ghorbani AA: **A Detailed Analysis of the CICIDS2017 Data Set.** *Information Systems Security and Privacy.* Springer; 2018; pp. 172–188.
Publisher Full Text

18. Khan MA, Iqbal N, Jamil H, *et al.*: **An optimised ensemble prediction model using AutoML based on soft voting classifier for network intrusion detection.** *J. Netw. Comput. Appl.* 2023; **212**: Article 103560.
Publisher Full Text

19. Ferrag MA, Maglaras L, Moschoyiannis S, *et al.*: **Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study.** *J. Inf. Secur. Appl.* 2020; **50**: Article 102419.
Publisher Full Text

20. Aktar S, Nur AY: **Towards DDoS attack detection using deep learning approach.** *Comput. Secur.* 2023; **129**: Article 103251.
Publisher Full Text

21. Motlagh FN, Hajizadeh M, Majd M, *et al.*: **Large Language Models in Cybersecurity: State-of-the-Art.** *arXiv preprint.* 2024. arXiv: 2402.00891.
Publisher Full Text

22. Xu H, Wang S, Li N, *et al.*: **Large Language Models for Cyber Security: A Systematic Literature Review.** *arXiv preprint arXiv: 2405.04760.* 2024.
Publisher Full Text

23. Zhang J, Bu H, Wen H, *et al.*: **When LLMs Meet Cybersecurity: A Systematic Literature Review.** *Cybersecurity.* 2025; **8**: Article 55.
Publisher Full Text

24. Jaffal NO, Alkhanafseh M, Mohaisen D: **Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques.** *AI.* 2025; **6**(9): Article 216.
Publisher Full Text

25. Atlam HF: **LLMs in Cyber Security: Bridging Practice and Education.** *Big Data and Cognitive Computing.* MDPI; 2025; Vol. **9**(7): Article 184.
Publisher Full Text

26. Hassanin M, Moustafa N: **A Comprehensive Overview of Large Language Models (LLMs) for Cyber Defences: Opportunities and Directions.** *arXiv preprint arXiv:2405.14487.* 2024.
Publisher Full Text

27. Chhetri C: **Exploring Large Language Model-Powered Pedagogical Approaches to Cybersecurity Education.** *Proceedings of the 25th Annual Conference on Information Technology Education (SIGITE '24).* ACM; 2024; pp. 163–166.
Publisher Full Text

28. Hadlington L: **Human Factors in Cybersecurity: Examining the Link Between Internet Addiction, Impulsivity, Attitudes Towards Cybersecurity, and Risky Cybersecurity Behaviours.** *Heliyon.* 2017; **3**(7): Article e00346.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Aldawood H, Skinner G: **Reviewing Cyber Security Social Engineering Training and Awareness Programs—Pitfalls and Ongoing Issues.** *Future Internet.* 2019; **11**(3): Article 73.
Publisher Full Text

30. Bada M, Sasse AM, Nurse JR: **Cyber Security Awareness Campaigns: Why do they fail to change behaviour?** *arXiv preprint arXiv:1901.02672.* 2019.
Publisher Full Text

31. Zhang-Kennedy L, Chiasson S: **A Systematic Review of Multimedia Tools for Cybersecurity Awareness and Education.** *ACM Comput. Surv.* 2021; **54**(1): Article 12.
Publisher Full Text

32. Araujo MS, Machado BAS, Passos FU: **Resilience in the Context of Cyber Security: A Review of the Fundamental Concepts and Relevance.** *Appl. Sci.* 2024; **14**(5): Article 2116.
Publisher Full Text

33. Ramos-Cruz B, Andreu-Perez J, Martínez L: **The Cybersecurity Mesh: A Comprehensive Survey of Involved Artificial Intelligence Methods, Cryptographic Protocols and Challenges for Future Research.** *Neurocomputing.* 2024; **581**: Article 127427.
Publisher Full Text

34. Somma M, Flatscher A, Stojanović B: **Edge-Based Anomaly Detection: Enhancing Performance and Sustainability in Smart Water Distribution Systems.** *2024 32nd Telecommunications Forum (TELFOR).* IEEE; 2024; pp. 1–4.
Publisher Full Text

35. Bhatt S, Manadhata PK, Zomlot L: **The Operational Role of Security Information and Event Management Systems.** *IEEE Security & Privacy.* 2014; **12**(5): 35–41.
Publisher Full Text

36. Tantaroudas ND: SentinelSphere Workshop Questionnaire Data (Aggregated). [Data set]. *Zenodo.* 2026.
Publisher Full Text

37. McCracken AJ: **SentinelSphere-Enhanced DNN Anomaly Detectors.** *Zenodo.* 2026.
Publisher Full Text

38. McCracken AJ: **Cybersecurity Chatbot Implementation.** *Zenodo.* 2025.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔ ? ?

Version 2

Reviewer Report 19 June 2026

https://doi.org/10.21956/openreseurope.26095.r75908

© **2026 Olaniyi O.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Oluwaseun Oladeji Olaniyi
University of the Cumberlands, Williamsburg, Kentucky, USA

This manuscript presents SentinelSphere, an AI-driven cybersecurity platform that integrates machine learning-based threat detection with LLM-powered cybersecurity education. The platform combines an enhanced deep neural network, a traffic-light threat visualization system, a Phi-4 cybersecurity chatbot, and a Rust-optimized processing pipeline. Validation was performed using benchmark intrusion detection datasets and workshops involving 79 stakeholders.
The work addresses an important cybersecurity challenge by combining technical threat detection with user awareness training. The platform is well designed, practically relevant, and demonstrates promising performance and usability.
Review Questions

1. Is the work clearly and accurately presented and does it cite the current literature?

Answer: Yes
The manuscript is clearly written, logically organized, and supported by relevant and current literature. The authors effectively position their contribution within research on intrusion detection, cybersecurity education, LLMs, and cyber resilience.

1. Is the study design appropriate and does the work have academic merit?

Answer: Yes
The study combines system development, benchmark evaluation, performance optimization, and stakeholder validation. The integrated approach has clear academic and practical value and addresses an important gap between technical security controls and human factors.

1. Are sufficient details of methods and analysis provided to allow replication by others?

Answer: Partly
The manuscript provides substantial methodological information but lacks several details needed for full replication.
Key issues:
• Exact definitions of engineered features (e.g., Request Complexity Score, Response Pattern Ratio, Path Entropy) are not provided.
• Dataset preprocessing, normalization, and train/validation/test split procedures are insufficiently described.

• Survey instruments and workshop questionnaires are not included.
Suggested improvements:
• Provide mathematical definitions for engineered features.
• Document preprocessing and data partitioning procedures.
• Include questionnaires and evaluation materials as supplementary files.

1. If applicable, is the statistical analysis and its interpretation appropriate?

Answer: Partly
Performance metrics are reported appropriately, but statistical rigor could be improved.
Key issues:
• No confidence intervals or variability estimates are reported.
• No statistical significance testing is provided.
• Workshop findings are largely descriptive.
• The professional validation cohort (n=3) is too small for broad conclusions.
Suggested improvements:
• Report confidence intervals for key metrics.
• Include repeated-run statistics where applicable.
• Clarify that workshop findings are exploratory.

1. Are all the source data underlying the results available to ensure full reproducibility?

Answer: No
Although the primary benchmark datasets are publicly available, the materials required for full reproducibility are not provided.
Missing items include:
• Feature extraction scripts.
• Training and evaluation code.
• Survey instruments.
• Anonymized workshop data.
• Experimental configuration files.
Required action:
The authors should provide code, supporting datasets, and evaluation materials through a public repository.

1. Are the conclusions drawn adequately supported by the results?

Answer: Yes
The conclusions are generally supported by the reported results. Claims regarding detection performance, scalability, and user acceptance are justified by the presented evidence. The authors appropriately acknowledge the need for future longitudinal studies before claiming sustained behavioral change.
Issues Requiring Attention
To strengthen scientific transparency and reproducibility, the authors should:

1. Provide complete feature-engineering and model implementation details.
2. Describe dataset preparation and partitioning procedures in full.
3. Release survey instruments and evaluation materials.
4. Make code and experimental configurations publicly available.
5. Provide anonymized validation datasets where possible.

Overall Assessment
This is a strong and timely contribution that successfully combines AI-based threat detection with cybersecurity awareness education. The manuscript is scientifically sound, well written, and

demonstrates practical value. The primary weakness is limited reproducibility rather than flaws in the research itself. Addressing the transparency and replication issues would further strengthen the work.
Recommendation: Approved with reservations

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cybersecurity, Cyber Engineering, Information Governance, Data Privacy, Risk Management, Artificial Intelligence Applications in Security, Security Awareness and Human Factors, Information Systems, Data Analytics, Digital Governance, Cyber Resilience, Threat Detection and Incident Response, Enterprise Risk Management, Compliance and Regulatory Frameworks.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 19 June 2026

https://doi.org/10.21956/openreseurope.26095.r75909

© 2026 Rehan H. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Hassan Rehan
Purdue University, West Lafayette, Indiana, USA

The paper presents SentinelSphere, a cybersecurity platform that integrates AI-driven threat detection with LLM-powered security awareness education, built as an extension to the EU Horizon Europe ResilMesh framework. The detection component uses an Enhanced Deep Neural Network with 31 features (21 standard network features plus 10 HTTP-specific indicators) trained on the CIC-IDS2017 and CIC-DDoS2019 datasets, reporting 94% F1 score and a 69.5% reduction in false positives over the ResilMesh baseline. The educational component uses Microsoft's Phi-4 (14B parameters) quantized via Q4_K_M / llama.cpp / GGUF to 2.5 GB, enabling CPU-only deployment on standard hardware. A Rust rewrite of the core detection logic delivers 5.6× speedup for steady-state workloads and up to 326× for large batch processing. The system was validated through two workshops in Greece: a small professional cohort (N=3, multi-sector) and an educational cohort (N=76, predominantly undergraduate pedagogy students). All source code, the quantized model, and aggregated workshop data are openly published.

The v2 revision has appropriately addressed the major concerns raised by the previous reviewers, including clarification of participant cohorts (new Table 6), full specification of the Phi-4 variant and quantization method, technical explanation of the 326× Rust speedup, explicit treatment of class imbalance handling, and softening of behavioral-change claims to match what the workshop data can actually support. The work is genuinely useful, the integration of detection and education is novel in practice, and the open availability of code and model is exemplary. The reservations below are about strengthening the empirical case rather than questioning the core contribution.

Methods and Replication
The Phi-4 specification, the Rust rewrite explanation, and the class imbalance treatment are all sufficient for replication after the v2 revisions. Two gaps remain. First, the Enhanced DNN is benchmarked only against the original ResilMesh anomaly detection baseline. The paper positions its results favourably against published numbers from Khan et al. (92% precision with attention mechanisms) and Ferrag et al. (typical 5-15% false positive reduction for DNN approaches), but no head-to-head comparison is run on the same data splits with the same evaluation protocol. For a paper that claims to exceed prior work, at least one external comparison on the same dataset would substantially strengthen the empirical case. The contractive autoencoder result referenced from Aktar and Nur (97.58% accuracy on CIC-DDoS2019) is one obvious candidate. Second, the training/test split for the Enhanced DNN is not described — was the model evaluated through stratified k-fold, a single hold-out split, or temporal splitting? Given the extreme class imbalance, this matters for interpreting the reported metrics.

Statistical Analysis
The standard classification metrics (F1, precision, recall, accuracy, confusion matrices) are appropriately applied for the multi-class detection task, and the prior reviewer is correct that the 69.5% false positive reduction is large enough that statistical significance testing is not the central concern. Three issues warrant attention.

First, the SQL injection detection row in Table 4 reports an improvement from 85.7% to 90.5%, but with only 21 samples in this class, this represents detecting 19 of 21 samples versus 18 of 21 samples — a difference of one sample. The "4.8% improvement" framing significantly overstates what can be inferred from this comparison. The authors should either report this as 18/21 vs 19/21 to convey the actual count, or add a caveat that the SQL injection class is too small to support meaningful percentage comparisons.

Second, the educational workshop's pre/post design is anonymous and therefore unpaired (N=39 pre, N=36 post). The v2 revision correctly acknowledges this in the body and softens corresponding claims, but the abstract still reports "most participants achieving improved security concept comprehension post-training," which is a comparison-of-means inference that anonymous unpaired samples can only weakly support. Either the abstract language should be further softened to match the body, or the framing should shift to baseline gaps identified versus post-workshop satisfaction reported.

Third, the 91.7% Traffic Light System comprehension result is reported as a headline finding throughout the paper, but it is essential to note that this was measured on a cohort of mostly 18-20 year old pedagogy students with very low baseline cybersecurity expertise. A traffic light metaphor is intuitive by design and the cohort is exactly the audience for whom comprehension would be expected to be high. The result is not invalid, but it does not support generalization to SOC analysts, technical decision-makers, or other professional audiences. The professional workshop (N=3) is too small to bridge this gap. Either the framing should be more careful or the limitation should be discussed more explicitly.

Summary
The technical conclusions about detection performance, false positive reduction, Rust optimization, and enterprise-scale event throughput are well supported by the results presented. The conclusions about the human-factor component are appropriately calibrated in the body of the v2 revision — the dedicated "From Comprehension to Behavior Change" subsection is well written and honest about what was and was not measured. However, some residual mismatches remain between the conclusions section and the body. The conclusion states the platform "demonstrates strong potential to address both technical detection and human security awareness challenges," but the human security awareness side of that claim is currently supported by short-term satisfaction and comprehension data from a single non-representative cohort, not by evidence of awareness improvement. The body acknowledges this; the conclusion could be tightened to match.

A few smaller items. The text on page 4 cites the "Verizon 2023 Data Breach Investigations Report" while reference 5 lists the 2025 edition — the year should be reconciled. The plain language summary states the Rust rewrite is "more than five times faster" while the abstract and contributions feature the 326× figure for large batches; readers moving between sections may find this inconsistent. A short note in the plain language summary that the 5.6× figure refers to typical workloads while 326× applies at very large batch sizes would resolve this.

Strengths
The integration of detection and education is genuinely valuable and rare in the literature. The Rust rewrite is a real engineering contribution and the technical explanation added in v2 (no GIL, contiguous memory layout, deterministic deallocation) is well grounded. The open availability of code, model, and data on GitHub, HuggingFace, and Zenodo is exemplary for this kind of work and sets a useful reproducibility standard for the field. The GDPR-compliant ethics handling is appropriate. The honest discussion of the gap between immediate comprehension and durable behavioral change, together with a concrete plan for longitudinal follow-up, is exactly the right way to handle this limitation.

Items that should be addressed

1. Add at least one external benchmark comparison on CIC-IDS2017 or CIC-DDoS2019 to substantiate the claim of exceeding prior work, or soften the comparative framing in the Related Work and Comparison with Related Work sections.
2. Describe the train/test split protocol used for the Enhanced DNN evaluation, including whether the reported metrics are from a single split or cross-validation.
3. Add a caveat to the SQL injection detection result in Table 4 indicating that the 21-sample class size limits the inferential weight of percentage-point comparisons; consider reporting raw counts.
4. Reconcile the Verizon 2023 vs 2025 reference between the text and the reference list.
5. Bring the abstract claim about post-training comprehension into closer alignment with the unpaired anonymous nature of the survey data, as already discussed in the body.
6. Add a sentence in the conclusion or limitations clarifying that the 91.7% TLS comprehension figure is from a low-baseline pedagogy cohort and may not generalize to professional audiences without further validation.
7. Reconcile the "more than five times faster" framing in the plain language summary with the 326× headline figure elsewhere by noting which figure applies to which workload.

Overall
This is a substantial and useful piece of applied cybersecurity work with a real implementation, open artefacts, and honest reporting of limitations. The v2 revision has addressed the previous reviewers' main concerns thoroughly. The reservations above are about strengthening the empirical case and tightening some framing rather than questioning the substance of the work, and most are addressable in a focused minor revision. The paper deserves to be approved with these reservations addressed.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial Intelligence and Cybersecurity, with focus on deep learning for intrusion detection, federated learning, privacy-preserving analytics, and LLM applications in

security

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 17 June 2026

https://doi.org/10.21956/openreseurope.26095.r75844

© 2026 Zangana H. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Hewa Majeed Zangana
Duhok Technical College, Duhok Polytechnic University, Duhok, Iraq

The authors have successfully addressed all of my previous concerns and reservations. Specifically, the addition of the detailed participant cohort table (Table 6) provides the necessary demographic clarity, the exact parameters and quantization framework for Microsoft's Phi-4 model have been clearly specified, all technical acronyms have been systematically defined, and the claims regarding long-term human-factor behavioral change have been appropriately softened to match the current scope of the workshop data. Furthermore, the technical explanation provided for the $326\times$ Rust-based speedup at scale is highly robust and scientifically sound. The manuscript is now exceptionally well-detailed, technically sound, and fully reproducible. I have no further comments to make, and I am pleased to recommend the paper for immediate acceptance and indexing in its current form.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 15 June 2026

https://doi.org/10.21956/openreseurope.26095.r75845

© 2026 Dalal A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aryendra Dalal
Middle Georgia State University, Macon, Georgia, USA

All previous comments have been addressed as required, no further comments

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cybersecurity

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 27 March 2026

https://doi.org/10.21956/openreseurope.24826.r71580

© 2026 Dalal A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Aryendra Dalal
[1] Middle Georgia State University, Macon, Georgia, USA
[2] Middle Georgia State University, Macon, Georgia, USA

OVERALL RECOMMENDATION
This paper tackles an important problem—combining technical threat detection with human security awareness—and does so with a practical, well-executed system. The work is fundamentally sound and makes genuine contributions. However, there are some gaps in how the evaluation is presented and a few technical details that need clarification before I can fully endorse it for indexing.
Recommendation: Approved with reservations (Minor revisions required)
WHAT WORKS WELL
The core idea is smart. Most cybersecurity tools treat detection and education as completely separate problems. Your approach of using every security event as both a threat to neutralize and a teaching moment is genuinely novel and makes practical sense.
The technical execution is solid. The Rust rewrite delivering 5-6x speedup while maintaining 100% accuracy equivalence shows careful engineering. The 69.5% false positive reduction is meaningful—that's the difference between a tool analysts might actually use versus one they'll ignore due to alert fatigue.
The validation effort is substantial. You didn't just build a system and call it done. You actually put it in front of 76 real people across different sectors and measured their responses. The 86.1% satisfaction rate and 91.7% comprehension of your Traffic Light System are strong indicators the design works.
Reproducibility is excellent. Providing code on GitHub, the quantized model on HuggingFace, and aggregated data on Zenodo sets a good standard. Other researchers can actually build on this work.
ISSUES THAT MUST BE FIXED

1. **Workshop Participant Details Are Too Vague**
You mention "76 stakeholders" and "54 participants" at different points, with pre-workshop N=39 and post-workshop N=36. This is confusing.
What's actually unclear:
- How many people attended the professional workshop versus the educational workshop?
- What was the baseline cybersecurity expertise of each group?
- Why did participant numbers drop from 39 to 36 post-workshop?

Why this matters: Saying "91.7% found the Traffic Light System clear" means something very different if you're talking about security professionals versus college students studying pedagogy. The reader can't properly interpret your validation results without knowing who these people actually were.
Fix required: Add a clear table showing:
- Workshop 1 (Professional): N=X, sectors represented, prior security experience
- Workshop 2 (Educational): N=76 (39 pre-survey, 36 post-survey), student vs. faculty breakdown, why some didn't complete both surveys

2. **The LLM Implementation Needs Clearer Specification**
You say you used "Microsoft's Phi-4 model" with "Q4_K_M quantisation" but this is ambiguous.
What's missing:
- Which specific Phi-4 variant? (Phi-4-mini? Phi-4-base? Something else?)
- What was the exact model size before quantization (you say "28 GB to 2.5 GB" but Phi-4-mini is typically around 14GB)
- Which quantization framework did you use? (llama.cpp, GGUF, something else?)

Why this matters: LLM behavior varies significantly between versions. Someone trying to replicate your educational chatbot needs to know exactly which model to download.
Fix required: State explicitly in the text (not just the HuggingFace link): "We used Phi-4-mini-instruct (14B parameters, original size 28GB) quantized to Q4_K_M format using llama.cpp, resulting in a 2.5GB deployment model."
3. **"Educational Retention" Is Mentioned But Not Measured**
Your workshop shows people liked the chatbot and understood the traffic light system immediately after seeing it. But does this actually change their behavior over time?
You acknowledge this limitation briefly but it deserves more discussion because it's central to your claimed contribution.
Current state: You measured immediate comprehension and satisfaction. You didn't measure whether people actually started using 2FA, stopped reusing passwords, or became better at spotting phishing emails weeks or months later.
Why this matters: The cybersecurity education literature you cite (Bada et al., Aldawood & Skinner) specifically points out that awareness campaigns often fail to change actual behavior. Your system might be better, but you haven't shown that yet.
Fix required: Add a paragraph in the Discussion or Limitations section explicitly stating:
- This study measured immediate comprehension and satisfaction
- Actual behavior change (e.g., adoption of 2FA, reduced password reuse) was not measured
- Longitudinal studies tracking participants over 3-6 months are needed to validate educational effectiveness
- This is a priority for future work

4. **Some Acronyms Are Not Defined at First Use**

Examples I caught:

- "RNN" appears on page 5 without definition
- "CNN-LSTM" appears without spelling out what CNN and LSTM stand for
- "SOAR platforms" mentioned in Future Work without definition

Why this matters: Your paper bridges cybersecurity and education domains. Not all readers will know every machine learning acronym.
Fix required: Go through the entire manuscript and ensure every technical abbreviation is defined the first time it appears. This is a simple editing pass.
QUESTIONS AND CLARIFICATIONS
Q1: The 326x speedup for 100,000 events—is that real?
Your Table 2 shows processing 100,000 events in 5.39 seconds (Rust) versus 1,756.9 seconds (Python). That's a 326x improvement which seems almost too good to be true.
I'm not saying it's wrong—Rust absolutely can deliver these gains for I/O-heavy workloads. But it would be helpful to explain what specifically caused such a dramatic difference. Is this mostly eliminating Python's Global Interpreter Lock for parallel processing? Better memory layout? Something else?
A one-sentence technical explanation would make this more credible.
Q2: How did you handle the class imbalance in training?
You mention 168,186 benign samples versus only 21 SQL injection samples. That's an extreme imbalance. Did you:

- Use class weights during training?
- Apply SMOTE or other oversampling?
- Just accept the imbalance and focus on precision/recall for minority classes?

This matters for understanding how your model actually works.
MINOR WRITING ISSUES
The "Plain Language Summary" is excellent —genuinely accessible while still being accurate. Keep that.
Some sentences are unnecessarily complex. Example from page 4: *"The architecture operates as a strategic enhancement to the ResilMesh security framework, integrating as an advanced threat analytics layer while preserving the existing architecture's integrity."*
Simpler version: *"SentinelSphere extends the ResilMesh framework by adding advanced threat analytics without disrupting existing workflows."*
Not a big deal, but simpler prose helps international readers.
WHAT ABOUT THE STATISTICAL ANALYSIS?
The statistics you present are appropriate for this type of work:

- Standard classification metrics (F1, precision, recall, accuracy)
- Confusion matrices properly constructed
- Performance benchmarks with clear methodology

No issues here. You're comparing against a well-defined baseline (the original ResilMesh detector) and the improvements are substantial enough that statistical significance testing isn't really necessary—a 69.5% false positive reduction is clearly meaningful.
DATA AVAILABILITY
You've done everything right here:

- Public datasets clearly referenced (CIC-IDS2017, CIC-DDoS2019)
- Code on GitHub with DOIs

- Model on HuggingFace
- Aggregated workshop data on Zenodo

The explanation of why you can't share individual survey responses (GDPR compliance, anonymous collection) is reasonable and well-stated.
No changes needed.
DO THE CONCLUSIONS FOLLOW FROM THE RESULTS?
Mostly yes, with one caveat.
Claims that are well-supported:

- Enhanced DNN achieves better detection with fewer false positives ✓
- Rust implementation delivers major performance gains ✓
- System can scale to enterprise workloads ✓
- Users find the Traffic Light System intuitive ✓
- Workshop participants were satisfied with the platform ✓

Claim that needs softening:

- "The system addresses both technical and human-factor cybersecurity challenges"

You've shown it addresses the technical detection challenge effectively. For the human factor, you've shown people like the educational component and understand it immediately after seeing it. But you haven't shown it actually changes security behavior over time.
Fix: Change this to "The system demonstrates potential to address both technical detection and human security awareness challenges, though longitudinal studies are needed to validate sustained behavioral change."
SUMMARY OF REQUIRED CHANGES
Must fix for scientific soundness:

1. **Clarify workshop demographics** - Table showing participant counts, expertise levels, and sector breakdowns for both workshops
2. **Specify LLM model exactly** - State precise model variant, original size, and quantization method in the text
3. **Acknowledge educational effectiveness limitation** - Add explicit discussion that you measured immediate satisfaction but not long-term behavior change
4. **Define all acronyms** - Edit pass to ensure RNN, CNN-LSTM, SOAR, etc. are defined at first use

Should fix to strengthen the paper:

1. Explain what causes the dramatic 326x speedup for large batches
2. Clarify how you handled class imbalance during training
3. Soften the claim about addressing "human-factor challenges" to reflect that behavior change wasn't measured

FINAL ASSESSMENT
This is good work that makes genuine contributions to an important problem. The technical system is well-designed, the performance improvements are substantial, and the integration of education with detection is genuinely novel.
The issues I've raised are fixable with revisions—none of them undermine the core findings. Once you address the participant demographics,
LLM specification, and educational effectiveness limitations, this will be a solid contribution to the cybersecurity literature.

The work deserves indexing after these revisions.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cybersecurity

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 13 March 2026

https://doi.org/10.21956/openreseurope.24826.r70373

© 2026 Zangana H. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Hewa Majeed Zangana
[1] Duhok Technical College, Duhok Polytechnic University, Duhok, Iraq
[2] Duhok Technical College, Duhok Polytechnic University, Duhok, Iraq

The manuscript presents a timely integration of AI-driven threat detection with real-time security education. By utilizing an "Enhanced DNN" for network traffic analysis and an LLM for generating contextual educational content, the authors address both technical defense and the "human factor" vulnerability. The transition from Python to Rust for performance optimization is a practical contribution that enhances the paper's technical depth.
Detailed Critical Assessment

1. Presentation and Literature Review The work is clearly presented. The authors successfully bridge two often-disparate fields: machine learning for intrusion detection and pedagogical strategies for cybersecurity awareness. The citation of the 2024 NIST Cybersecurity Framework 2.0 demonstrates that the literature review is current.

- *Requirement for Soundness:* Ensure all acronyms (e.g., RNN, CNN-LSTM, LLM) are defined at first mention to maintain accessibility for interdisciplinary readers.

2. Study Design and Academic Merit The hybrid approach—treating a security alert as a "teachable moment"—is conceptually strong. The study design appropriately benchmarks the proposed DNN against the ResilMesh baseline.

- *Constructive Criticism:* While the study includes a "multi-sector stakeholder validation," the sample size and specific demographics of the workshop participants are slightly opaque. To strengthen the academic merit, the authors should explicitly state the number of participants and their prior expertise levels to better contextualize the "90% positive feedback."

3. Methodology and Replication The methodology is detailed, particularly regarding feature engineering (HTTP-specific indicators). Providing the source code via GitHub and the LLM via Hugging Face is a gold standard for reproducibility.

- *Requirement for Soundness:* The authors should specify the exact version/checkpoint of the LLM used (e.g., Llama-3-8B-Instruct) within the text, as model behavior can vary significantly across versions.

4. Statistical Analysis and Interpretation The metrics (F1, Precision, Recall) are standard and correctly applied. The 69.5% reduction in false positives is a significant finding that is well-supported by the dual-stage filtering logic described.

5. Data Availability The use of Zenodo for workshop data and the reliance on public datasets (CIC-IDS2017) fulfills open-science requirements. No further action is needed here.

6. Support for Conclusions and Limitations The conclusions follow logically from the results. The authors honestly address the "LLM hallucination" risk and the need for quantization.

- *Constructive Criticism:* The paper would benefit from a more rigorous discussion on the "educational retention" of users. While the platform *provides* education, the study does not yet prove long-term behavior change. Adding a "Future Work" note regarding longitudinal studies on user behavior after interacting with SentinelSphere would round out the discussion.

Summary of Required Revisions for Scientific Soundness:

1. **Quantify Stakeholders:** Clearly define the N-size and expertise of the workshop participants.
2. **Model Versioning:** Explicitly name the base LLM model and version used in the final implementation.

Clarity: Ensure all technical abbreviations are defined in the Introduction or Background sections.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and does the work have academic merit?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.