LLM-Based Framework for Next-Generation Cyber Threat Detection

Adamu Hussaini, Almustapha A. Wakili, Usman Shuaibu Musa, and Wei Yu
Dept. of Computer and Information Sciences, Towson University, USA 21252
{ahussa7, awakili1, umusa1}@students.towson.edu, wyu@towson.edu

Abstract

The rapid growth of Internet of Things (IoT) technologies leads to many devices deployed and interconnected over cyberspace, posing serious cybersecurity challenges. Traditional techniques, such as centralized Intrusion Detection Systems (IDS) and resource-intensive firewalls, present scalability concerns in large-scale IoT networks. These systems are inefficient when monitoring the diverse and distributed nature of networking systems and struggle with scalability and efficiency in these complex, distributed networks. As systems become more interconnected, they are increasingly vulnerable to sophisticated attacks. We propose a security framework that utilizes a Large Language Model (LLM) to enhance intrusion detection for cybersecurity. Our work involves developing several innovative LLM algorithms for dynamic cyber threat detection. Furthermore, we recommend future research to explore the integration of LLMs with safeguard mechanisms, edge computing, and blockchain technology, as well as extending our framework to cyber-physical system (CPS) security, ensuring safe and reliable operations in complex environments.

1 Introduction

Recent research by [Ali, 2024] highlights a sharp rise in cyberattacks on critical national infrastructure and enterprise systems. By the end of 2025, annual damages are expected to reach $10.5 trillion—up from $3 trillion in 2015 [Morgan, 2023]. To address these growing threats, the National Institute of Standards and Technology (NIST) introduced a cybersecurity framework in 2014, outlining policies for identifying, protecting, detecting, responding to, and recovering from cyber incidents [Cybersecurity C.I., 2018].

Advancements in machine learning (ML) and deep learning (DL) techniques have significantly improved a variety of areas, such as intrusion detection across various domains [G. Lira *et al.*, 2024; Hatcher and Yu, 2018; Tian *et al.*, 2025; He *et al.*, 2023; Song *et al.*, 2025; Kheddar *et al.*, 2024; Liang *et al.*, 2019; Mohammadi *et al.*, 2018; Liu *et al.*, 2021; Ge *et al.*, 2025; Alharbi *et al.*, 2024; Yu *et al.*, 2015; Dong *et al.*, 2021; Qian *et al.*, 2025]. Integrating AI and ML into Large Language Models (LLM) has further advanced this field. LLMs, trained on vast datasets, can analyze network logs, adapt to evolving threats, and distinguish between normal and malicious activities [Ridwan *et al.*, 2021; Huang *et al.*, 2024]. Their powerful Natural language processing (NLP) capabilities enable accurate detection and proactive mitigation of even unknown attacks [Ferrag *et al.*, 2024; Li *et al.*, 2024]. This research explores how LLMs enhance the security and adaptability of emerging technologies.

Cybercrime has recently surged in frequency and matured into a formidable industry capable of undermining even robust digital infrastructures so it is essential to design computing infrastructure and systematical defend to deal with threats (e.g., intrusion detection, forensics) [Roshanaei *et al.*, 2024; Mirkovic and Reiher, 2004; Chen *et al.*, 2016; Chen *et al.*, 2021; Yu *et al.*, 2013; Yu *et al.*, 2007; Yang *et al.*, 2015]. For instance, as reported in 2023 by The Hackers News [The Hacker News, 2023], Distributed Denial-of-Service (DDoS) attacks had shot down Microsoft services. Similarly, Google Cloud announced that it handled significant DDoS attack that consists of 398 million requests per second, in the same year of 2023 [TechTarget, 2023]. This underscores the growing scale and precision of modern cyber threats.

Traditional IDS methods struggle to detect novel attacks, driving demand for more robust solutions. While ML and DL have improved detection, recent advances using LLMs show greater promise by overcoming rule-based limitations and enhancing real-time accuracy and adaptability in domains like smart transportation, smart energy grid, and smart manufacturing. Cyber-attacks have become a major global concern, amplified by Industry 4.0 advancements. LLMs offer innovative solutions to cybersecurity challenges across sectors of the economy [Thalpage and Nisansala, 2023]. Leveraging LLMs in IDS enables real-time analysis of large cybersecurity datasets and proactive threat detection. If deployed at the network edge, these systems can offload complex computational tasks from IoT devices, allowing faster, context-aware responses without straining device performance.

Deep learning has advanced threat detection with high accuracy and low false positives, but its need for training from scratch limits scalability. In contrast, LLMs leverage pretrained knowledge, reducing the need for extensive labeled data and computational resources. This research proposes

LLM-based intrusion detection frameworks with feature selection to address complex cyber threats in real time and overcome the limitations of traditional defenses.

It is worth noting that conventional cybersecurity detection mechanisms, such as signature-based, rule-based, and anomaly-based systems, as well as classical ML models, often lack the adaptability to detect emerging, polymorphic, or zero-day attacks. These techniques primarily rely on static features, such as predefined signatures, which makes them insufficient for handling the scale and complexity of modern threat landscapes. In contrast, LLMs introduce a novel capability by leveraging contextual understanding, semantic reasoning, and few-shot generalization, which enables them to interpret diverse data sources such as system logs, threat intelligence reports, and adversarial content. LLMs offer a robust capability for real-time detection of advanced cyber threats. Accordingly, this study examines the incorporation of LLMs as a crucial step toward developing intelligent and adaptive threat detection frameworks.

Our key contributions to this paper are listed as follows: (i) We propose a security framework that leverages an LLM-driven intrusion detection for cyberspace security. (ii) We develop various novel LLM algorithms for dynamic cyber threat detection. (iii) We suggest future research leveraging LLMs with safeguard mechanisms, edge computing, and blockchain to develop robust and adaptive security frameworks for CPS, ensuring safe and reliable operation in complex environments.

The remainder of the paper is organized as follows. Section 2 introduces the background of Next-Generation Cyber-Threats (NGCTs), IDS, and LLM. Section 3 presents our approach in detail including the results and discussion. Section 4 presents the performance evaluation. Section 5 highlights the challenges and future directions. Finally, Section 6 concludes the paper.

2 Preliminaries

2.1 Next-Generation Cyber-Threats (NGCTs)

These advanced cybersecurity threats represent a paradigm shift in malicious cyber activities' nature, sophistication, and impact. It differs from the traditional threats relying on predefined malware signatures or simple phishing schemes; NGCTs leverage emerging technologies (AI, ML, edge intelligence, etc.) and sophisticated attack techniques and exploit vulnerabilities in modern digital ecosystems [Rao *et al.*, 2023]. As stated by [Chisty *et al.*, 2022], modern organizations can better safeguard their IT infrastructure against sophisticated cyber-attacks and guarantee the confidentiality, integrity, and availability of cyber-assets in a world that is becoming increasingly digital by implementing cybersecurity strategies of the next generation that focus on risk management, collaboration, and resilience.

2.2 Intrusion Detection System (IDS)

Recent advancements in AI and other emerging trends show an increase in the global interconnectivity of modern and complex networks. There is an urgent need for the highest security measures in human history. The reliability of interconnected systems has been increasing daily. On the other hand, this has brought a wide range of vulnerabilities, making computer networks vulnerable to zero-day cyber-attacks [Musa *et al.*, 2021]. To combat these emerging challenges, IDS has emerged as an essential component of cyber-security, aiming to address this challenge. These systems monitor network traffic and identify potential cyber threats. They can analyze a system's network and identify potential security threats like malware [Hussaini *et al.*, 2021] and distributed denial-of-service attacks [Paya *et al.*, 2024]. IDS can be classified based on the ways of signature-based, anomaly-based, or hybrid-based [Mishra and Mishra, 2024] or based on deployment strategies such as network-based, host-based, and cloud-based deployment [Liu *et al.*, 2018].

2.3 Large Language Model (LLM)

LLM represents sophisticated AI systems with expert knowledge across various domains, including specialized areas like network intrusion detection. These models are developed using a type of neural network such as autoregressive transformers that generate sequences by predicting the next element based on previous elements in a sequential manner, which are sequential data applications such as language models, time-series predictions, etc., and are initially trained on vast, self-supervised datasets, which enable them to understand and generate contextually relevant responses [Huang *et al.*, 2024; Wan *et al.*, 2024; Zhang *et al.*, 2024]. An alignment process with human preferences often follows this foundational training. It uses human-based reinforcement learning (RLHF) to refine its performance for domain-specific tasks such as detecting and analyzing potential intrusions in network traffic [Grumeza *et al.*, 2024].

3 Our Approach

3.1 Model Architecture

Our model features an LLM-based IDS using GPT [OpenAI, 2022] and BERT variants [Devlin *et al.*, 2019], optimized for real-time intrusion detection. Designed for future deployment, it supports low-resource IoT environments and leverages edge computing for efficient, local threat analysis.

Generative Pre-trained Transformer (GPT)

The GPT model architecture is chosen for binary classification tasks' capability and optimization power to structure input data rather than conventional text-based sequences, as shown in Fig. 1, which follows a pipeline of tokenization, embedding extraction, and classification. The input data supplied to the GPT is converted to an acceptable format compatible with the GPT tokenizer, enabling the model to extract deep contextual embeddings from the input data. The input data are transformed into a tokenized sequence by the model tokenizer from individual numerical values suitable for language-model-compatible format at the tokenization layer.

Furthermore, a classification head consisting of a fully connected dense layer is added to the GPT model. This layer processes the contextualized embeddings and outputs a probability distribution over two target classes using a softmax activation function. Optimization uses the Adam optimizer

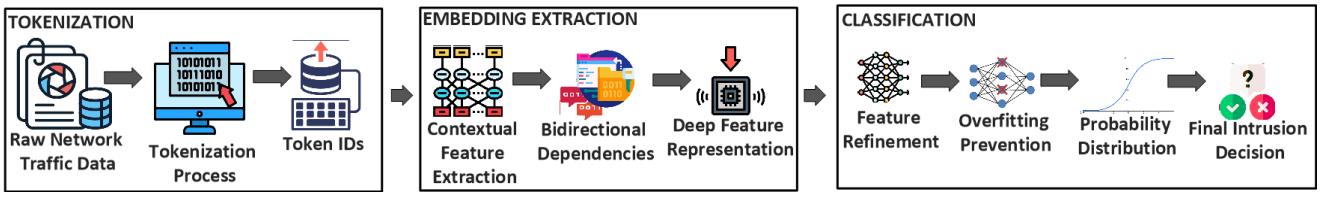

Figure 1: Our Model Pipeline

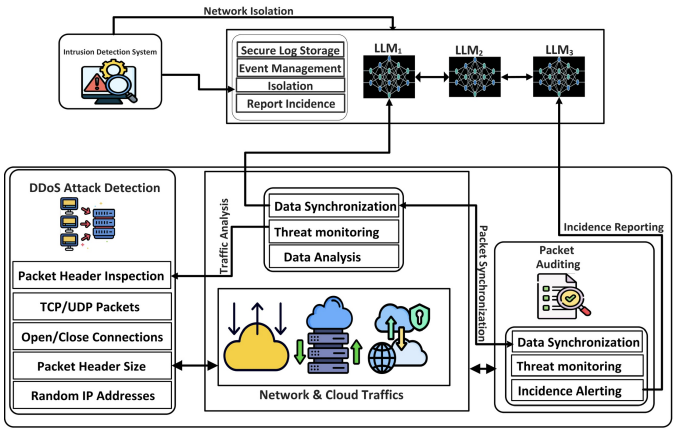

Figure 2: Proposed LLM-based IDS Framework

with a 5×10^{-5} learning rate and a Sparse Categorical Cross entropy loss function. This setup facilitates stable and efficient backpropagation, ensuring the model converges effectively during training. This approach is among the best solutions for real-time binary classification tasks because of its ability to process complex data patterns.

Bidirectional Encoder Representations from Transformers (BERT)

The second proposed model is based on two variants of BERT model architecture, which follows a standard BERT-based classification pipeline. Both BERT models use a pre-trained Bert-base-uncased encoder from the HuggingFace library. It is developed to process network traffic data represented as tokenized input sequences. It is divided into a BERT encoder, fully connected layers, and a softmax activation function. Specifically, the BERT encoder layer converts the input sequences into contextual embeddings by processing the tokenized sequences through multiple transformer layers. It also captures bidirectional dependencies between tokens, allowing the model to understand the complex relationships in network traffic data.

The input sequence is prepared in three steps. First, categorical features (e.g., protocol type, service, and flag) are label-encoded and converted into token sequences. Second, numerical features are normalized using a standard scaler and transformed into a textual format for tokenization. Lastly, the tokenized input is passed to the BERT tokenizer, which converts the sequences into token IDs and attention masks. The output from the first layer is passed through the second training layer (fully connected layers) to refine the extracted features and perform binary classification. The first stage is a dropout layer, which prevents overfitting by applying a 0.3 dropout rate. The dropout layer is followed by a dense classification layer with two output nodes corresponding to the two classification classes (regular and malicious). Like in the GPT model, a softmax activation function is also applied here to the output of the final dense layer to convert the logits into class probabilities. The softmax function is defined as: $P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$. Here z_i is the Logit for class i, k is the number of classes (2 in this case), and $P(y_i)$ is the probability that the input belongs to class i.

The softmax activation ensures that the sum of the predicted class probabilities equals 1, making it suitable for multi-class and binary classification problems. In the case of the loss function and optimization technique, the model is trained using cross-entropy loss. It aims to measure the difference between the predicted probability distribution and the true distribution. The cross-entropy loss function is defined as: $L = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$, where n is the number of samples, y_i refers to True label, $\hat{y}_i$ refers to the predicted probability for class i.

The Adam optimizer is used for model training with a learning rate of 5×10^{-5}. Adam leverages the benefits of momentum and adaptive learning rates, which makes it suitable for transformer-based models. The optimization strategy includes gradient clipping and early stopping to prevent overfitting.

3.2 Dataset and Preprocessing

The dataset used as the case study for this research is the NSL-KDD dataset [Tavallaee *et al.*, 2009], which was used for model training and evaluation. The NSL-KDD dataset addresses some inherent limitations in the original KDD Cup 99 dataset, such as the high redundancy of records and imbalance in class distribution, making it more suitable for intrusion detection research. The dataset contains 41 features categorized into continuous and discrete attributes and a label indicating whether the record corresponds to regular activity or an attack. The attacks are further classified into four categories: (i) *DoS (Denial of Service):* Flooding the network with traffic to exhaust resources and disrupt services. (ii) *R2L (Remote to Local):* Gaining unauthorized access to a machine over the network. (iii) *U2R (User to Root):* Gaining root-level access after initially accessing the system as a user. (iv) *Probe:* Scanning the network to gather information for future attacks.

3.3 Proposed IDS-Security Framework

Fig. 2 illustrates an LLM-enhanced IDS Security Framework designed to safeguard emerging technology environments

such as cloud computing, edge infrastructure, and distributed systems. This framework integrates several advanced cybersecurity modules, including the Intrusion Detection System (IDS), DDoS Attack Detection and Traffic Analysis, Packet Synchronization and Auditing, and a Large Language Model (LLM) Ensemble for deep threat analysis, supported by secure log management and network isolation.

At the core is the intrusion detection system, which continuously monitors network activity to identify suspicious behaviors. Upon detecting anomalies, the IDS logs the incidents and forwards the data to a secure log storage and event management system, initiating isolation protocols if necessary. A powerful feature of this architecture is the integration of large language models (LLM_1, LLM_2, and LLM_3). These models collaboratively handle tasks such as incident reporting, event isolation, and log interpretation. They provide contextual analysis and pattern recognition capabilities that outperform traditional rule-based detection, enabling the identification of complex or stealthy intrusions in real time.

The DDoS Attack Detection Module strengthens the system's resilience by performing deep traffic analysis. It examines packet headers, connection states (TCP/UDP), and IP address anomalies to recognize distributed denial-of-service attacks, leveraging this data to identify patterns associated with volumetric or stealth-based DDoS campaigns. Simultaneously, packet synchronization and auditing modules are responsible for continuous data synchronization, threat monitoring, and incident alerting. These ensure that any malicious packets are flagged, logged, and compared across the network, helping to localize and neutralize threats efficiently.

This model is designed for cloud-native and hybrid environments encompassing enterprise data centres, IoT-driven edge systems, and secure internet-connected devices. The shift towards network isolation, real-time analytics, and LLM-powered automation reflects the growing need for adaptive, intelligent, and scalable cybersecurity in modern digital ecosystems.

In summary, this security framework demonstrates how combining LLMs with modular IDS components can lead to robust, context-aware cybersecurity strategies for next-generation infrastructures. It enables proactive threat mitigation, autonomous incident response, and adaptive protection for complex, interconnected systems.

4 Performance Evaluation

We start by using a Kernel Density Estimate (KDE) plot and a simple pie chart to examine the relationship among the data values in the dataset. The KDE plot helps us understand the underlying distribution of the data, identify patterns, and compare different groups within the data, as illustrated in Fig. 3 and Fig. 4.

Fig. 3 shows a Kernel Density Estimate (KDE) plot representing the probability density of duration values for network protocol types: UDP, TCP, and ICMP. The x-axis represents the duration variable, measured in milliseconds. In contrast, the y axis represents the density, which shows how dense the duration values are for a given protocol type in the data. Different colors are used to describe the network protocol. For instance, light blue is used for UDP, medium blue represents TCP, and ICMP uses dark blue. Fig. 4 visualizes the distribution of a continuous variable (Duration) between different categories (Result). The x axis represents the values of the variable analyzed, likely normalized or scaled values denoted by $r_{\max}$ and other numerical points (0.8, 0.6, 0.4, 0.2). The y-axis represents the density estimate, which indicates how densely packed the data points are at different values of the x-axis. The values range from 0 to 40,000, suggesting the density scale.

Also, Fig. 5 consists of two pie charts that analyze the data set distribution according to the type of network protocol and the classification of attack outcomes. The left graph represents the proportion of network protocols observed in the dataset. Most network traffic uses the TCP protocol (82%), followed by UDP (12%) and ICMP (7%). This indicates that most network activity in the dataset is based on TCP, which aligns with common real-world network usage patterns. In contrast, the right chart shows the classification of network events into normal (53%) and attack (47%) categories. The nearly balanced distribution suggests a well-structured dataset, making it suitable for cybersecurity research, particularly in intrusion and anomaly detection studies.

4.1 GPT Results

The GPT-based IDS model was evaluated using key performance metrics to assess its effectiveness in detecting network intrusions. The results are summarized below in Table 1.

Table 1: Performance Metrics for GPT-Based IDS

Metric	Value
Test Accuracy	0.82
Precision (weighted avg)	0.85
Recall (weighted avg)	0.82
F1-Score (weighted avg)	0.82
AUC (Area Under the Curve)	0.87

Table 1 presents the performance metrics for a GPT-based IDS. The model achieves a test accuracy of 0.82, indicating that it correctly classifies network traffic 82% of the time. The weighted average precision is 0.85, meaning that the model is correct 85% of the time when it makes the decision on an attack or normal traffic. The recall, also 0.82, signifies that the model successfully identifies 82% of actual attacks and normal instances, reflecting a balanced detection capability. The F1 score, which balances precision and recall, is also 0.82, confirming consistent performance across attack and normal traffic classifications. The AUC (Area Under the Curve) of 0.87, as shown in Fig. 7 indicates a strong discriminatory power, meaning the model effectively differentiates between attack and normal traffic instances. These results suggest that the GPT-based IDS identifies intrusions well while maintaining a reasonable trade-off between false positives and false negatives, as shown in Fig. 6.

Fig. 6 demonstrates a confusion matrix that evaluates the performance of the GPT-based IDS. It shows the number of

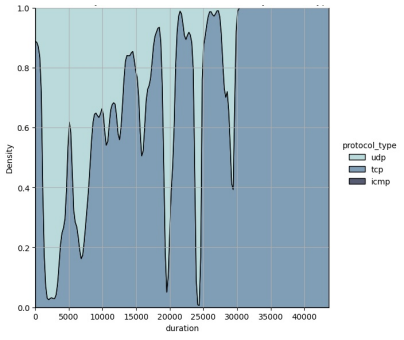

Figure 3: KDE of Network Traffic Duration by Protocol Type

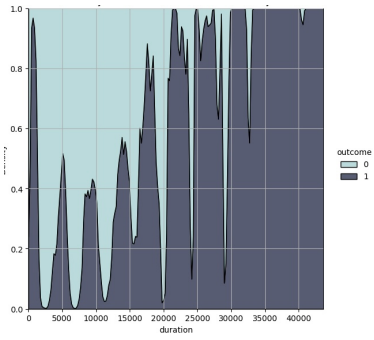

Figure 4: KDE Plot of Duration Distribution by Outcome

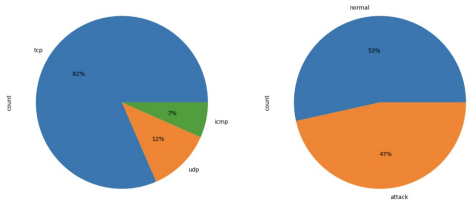

Figure 5: Protocol Distribution and Attack Classification

correct and incorrect predictions for two classes: *Attack* and *Normal*. We have the following results: (i) *True Positives*: The model correctly identified 9,311 attack instances. (ii) *False Negatives*: The model misclassified 3,522 attack instances as normal. (iii) *True Negatives*: The model correctly identified 9,188 normal instances. (iv) *False Positives*: The model mistakenly classified 523 normal instances as attacks. The confusion matrix shows that while the model performs well in detecting normal traffic, it struggles more with the recall for attacks, as evidenced by the relatively high number of false negatives. This suggests the model may need further optimization to reduce missed attacks while maintaining precision.

GPT Discussion

As shown in Fig. 7, the GPT-based IDS model offers competitive performance with an accuracy of 82% and an AUC of 0.87. Key observations include: (i) The model achieved high precision (0.95) for attack detection but had a lower recall (0.73), indicating some false negatives. (ii) Normal traffic detection showed high recall (0.95) but lower precision (0.72), suggesting some false positives. (iii) The overall F1-score of 0.82 reflects balanced performance, but there is room for improving recall in attack detection.

4.2 BERT Results

The BERT-based IDS model was evaluated using key performance metrics to assess its classification performance. Fig. 8 presents the precision, recall, and F1-score for attack and normal classes. The attack class exhibits high precision, suggesting that the model effectively minimizes false positives when identifying attacks. However, its recall is relatively lower, indicating that some attacks have not been detected. In contrast, the Normal class has a lower precision but higher recall, meaning that while more normal instances are correctly classified, there may be misclassifications. The F1 scores for both classes are reasonably balanced, highlighting a good trade-off between precision and recall. The results suggest that while the model performs well in distinguishing between attack and normal traffic, further tuning may be needed to improve recall for attack detection, ensuring fewer undetected threats.

Table 2 shows the results of the BERT-based IDS. The model achieves a high accuracy of 96%, demonstrating its overall effectiveness in correctly classifying network traffic.

Table 2: Results for BERT-Based IDS

Metric	Value
Accuracy	0.96
Precision (weighted avg)	0.95
Recall (weighted avg)	0.97
F1-Score (weighted avg)	0.96
PR-AUC (Precision-Recall Area Under Curve (AUC))	0.98
ROC-AUC (Receiver Operating Characteristic AUC)	0.97

The weighted average precision and recall scores are 0.95 and 0.97, respectively, indicating a strong balance between correctly identifying attacks (true positives) and minimizing false negatives. The F1 score that combines precision and recall is also high at 0.96, reinforcing the model's reliability in detecting intrusions.

Fig. 9 shows a confusion matrix that assess the performance of the BERT-based IDS. The matrix showcases the number of correct and incorrect predictions for two categories: 'Attack" and Normal." The model accurately classified 5,129 attack instances and 13,329 normal instances, indicating strong true positive and true negative rates. However, 6,597 attack instances were misclassified as normal cases (i.e., false negatives), and 140 normal instances were misclassified as attack cases (i.e., false positives). This suggests that the BERT-based model exhibits high accuracy and precision, and the confusion matrix confirms the model's robustness and reliability in distinguishing between malicious and benign network traffic.

Additionally, as reflected in the PR-AUC of 0.98 shown in Fig. 10 and ROC-AUC of 0.97 as shown in Fig. 11. The model distinguishes excellently between attack and normal traffic. The results suggest that the classifier maintains a high true positive rate while minimizing false positives across different threshold settings. In general, these results highlight the efficacy of BERT-based IDS in identifying malicious network activities with high confidence.

BERT Discussion

Overall, the BERT-based IDS model demonstrated high accuracy (0.96) and robust classification performance, with strong PR-AUC (0.98) and ROC-AUC (0.97) scores. Key observations include: (i) The model shows high sensitivity with a

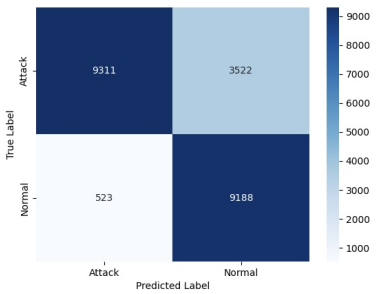

Figure 6: Confusion Matrix for GPT-Based IDS

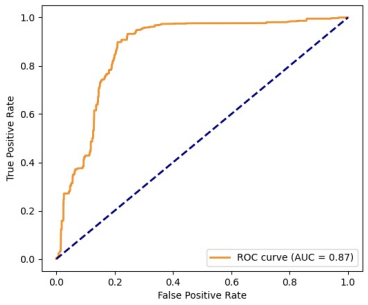

Figure 7: ROC Curve for GPT-Based Model

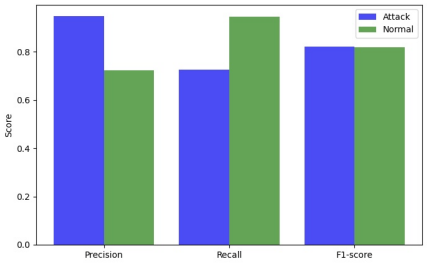

Figure 8: Performance for BERT-based IDS

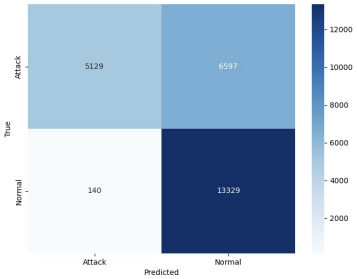

Figure 9: Confusion Matrix for BERT-based IDS

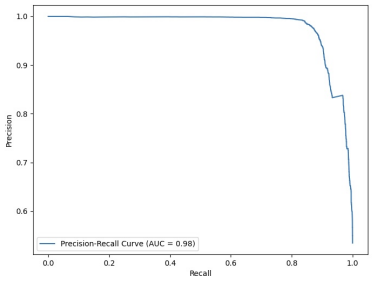

Figure 10: Precision-Recall Curve for BERT-based IDS

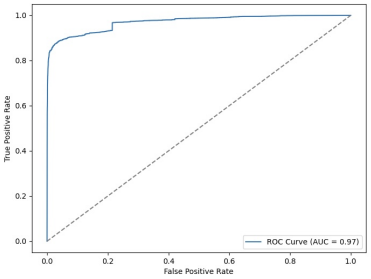

Figure 11: ROC Curve for BERT-based IDS

recall of 0.97, indicating a strong ability to identify most intrusions. (ii) A precision of 0.95 reflects the model's effectiveness in minimizing false positives. (iii) The F1 score of 0.96 suggests a balanced trade-off between precision and recall. It observed that the BERT-based model demonstrated strong performance in intrusion detection, with high accuracy and robust PR-AUC and ROC-AUC scores. Minor misclassifications were observed in attack instances, which could be improved by hyperparameter tuning and improved preprocessing.

4.3 DistilBERT Results

The DistilBERT-based IDS model is evaluated using key performance metrics to assess its effectiveness in detecting network intrusions. The results are summarized as follows. The DistilBERT-based IDS performance metrics are summarized in Table 3. The model achieves a test accuracy of 77%, with a weighted average precision of 0.81 and a recall of 0.78. The F1-score, which balances precision and recall, is 0.77. Additionally, the Precision-Recall AUC (PR-AUC) is 0.81, indicating a strong trade-off between precision and recall, while the ROC-AUC score of 0.78 suggests a moderate ability to differentiate between normal and attack traffic.

Figs. 12, 13, and 14 provide further insights into the classification performance of the model. In particular, the confusion matrix in Fig. 12 illustrates the distribution of true positives, true negatives, false positives, and false negatives, highlighting areas where the model misclassifies traffic. The Precision-Recall Curve in Fig. 13 demonstrates the model's effectiveness in handling imbalanced data, with an AUC of 0.81. Lastly, the ROC Curve in Fig. 14 shows the trade-off between the true positive rate and false positive rate, with an AUC of 0.78, reflecting a moderate classification capability.

Table 3: Performance Metrics for DistilBERT-Based IDS

Metric	Value
Test Accuracy	0.77
Precision (weighted avg)	0.81
Recall (weighted avg)	0.78
F1-Score (weighted avg)	0.77
PR-AUC (Precision-Recall AUC)	0.81
ROC-AUC (Receiver Operating Characteristic AUC)	0.78

4.4 Comparison Between BERT and DistilBERT

Table 4 shows the comparative analysis of the BERT-based and DistilBERT-based models. The table shows that the BERT model achieves higher accuracy and recall than DistilBERT. However, the DistilBERT model is significantly faster and requires fewer computational resources, making it a suitable option for timely intrusion detection in constrained environments.

BERT and DistilBERT Discussion

The BERT model demonstrates superior accuracy and recall, indicating a strong detection capability. However, the faster inference time and lower resource consumption of the DistilBERT model make it more suitable for real-time applications. Key observations include: (i) BERT outperforms DistilBERT in all key metrics, especially recall (0.97 vs. 0.78), reflecting better attack sensitivity. (ii) The higher precision

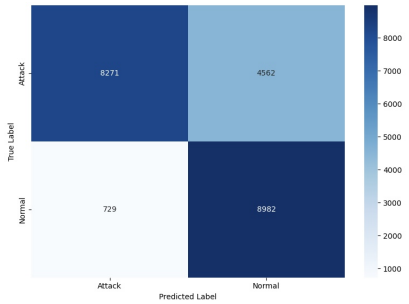

Figure 12: Confusion Matrix for DistilBERT-based IDS model

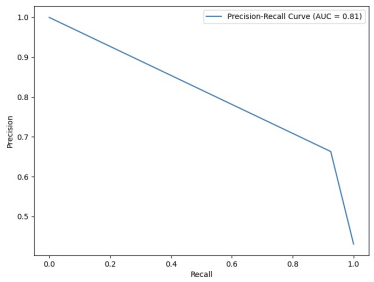

Figure 13: Precision-Recall Curve for DistilBERT-based IDS model

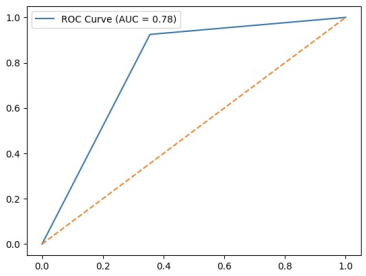

Figure 14: ROC Curve for DistilBERT-based IDS model

Table 4: Comparison of BERT and DistilBERT Performance Metrics

Metric	BERT Model	DistilBERT Model
Test Accuracy	0.96	0.77
Precision (weighted avg)	0.95	0.81
Recall (weighted avg)	0.97	0.78
F1-Score (weighted avg)	0.96	0.77
PR-AUC	0.98	0.81
ROC-AUC	0.97	0.78

(0.81) of DistilBERT compared to recall suggests that it generates fewer false positives but misses more attacks. (iii) The trade-off between accuracy and efficiency should be considered based on specific deployment requirements. In conclusion, although the BERT model provides better accuracy and recall, the DistilBERT model is a more lightweight alternative with a faster inference time.

5 Challenges and Future Directions

Even though leveraging LLMs in cybersecurity frameworks shows great promise for enhancing dynamic intrusion detection across emerging technologies, several challenges remain.

First, the development and deployment of such frameworks require accurate and high-quality data to train LLMs and build realistic, reliable security models effectively. Ensuring seamless integration with complex and evolving technological environments is essential for enabling real-time threat detection and adaptive responses. Moreover, it is critical to safeguard both the LLMs and the surrounding systems from data tampering, adversarial manipulation, and misuse, as such compromises could introduce new vulnerabilities rather than mitigate existing cyber risks.

Second, advancements in ML and LLM fine-tuning are expected to drive more sophisticated anomaly detection and threat prediction, making IDS increasingly adaptive and context-aware. Edge computing can help reduce latency and distribute computational tasks, facilitating real-time LLM-driven analysis even in environments with limited resources. Similarly, blockchain technology offers enhanced data integrity and security, ensuring tamper-proof records and secure communications across interconnected platforms. By integrating these innovations, LLM-based cybersecurity frameworks can evolve into robust, scalable, and adaptive solutions that address modern cyber threats facing emerging technologies.

Third, the proposed architecture can be extended with LLM-based intelligence tailored to cyber-physical system environments to enhance threat detection in CPS [Kim *et al.*, 2022; Liu *et al.*, 2019]. LLMs can analyze sensor data, control commands, and network traffic for subtle anomalies that traditional IDS may miss. Deploying lightweight LLMs on edge nodes will allow CPS components to gain localized decision-making and adaptive threat response, satisfying the timely performance requirement of CPS. LLM fine-tuning will enable the system to learn from CPS-specific operational contexts, improving situational awareness. Context-aware analysis from LLMs can help distinguish between legitimate anomalies and malicious behavior in CPS.

6 Final Remark

In this paper, we have proposed a security framework enhanced with LLMs to address cybersecurity challenges in emerging technologies. The research highlights the vulnerabilities of IoT and distributed systems, such as limited computational resources and exposure to sophisticated cyberattacks. Our hybrid approach combines traditional intrusion detection techniques with LLM-driven threat analysis. By leveraging models such as GPT, BERT, and DistilBERT, our framework processes the NSL-KDD dataset to detect threats, achieving high accuracy dynamically (BERT: 96% accuracy, 0.97 AUC) while maintaining a balance between precision and recall. The study also explores the trade-offs between performance and efficiency, with DistilBERT demonstrating faster inference speeds at the cost of slightly lower accuracy. Future research may engage safeguard mechanisms to enable the robustness and resilience of LLM, integrate edge computing for real-time analysis and blockchain technologies to ensure data integrity and extend our framework to a complex CPS environment. This will pave the way for scalable, adaptive security solutions in increasingly complex digital environments.

References

[Alharbi *et al.*, 2024] Saier Alharbi, Yifan Guo, and Wei Yu. Collusive backdoor attacks in federated learning frame-

works for IoT systems. *IEEE Internet of Things Journal*, 11(11):19694–19707, 2024.

[Ali, 2024] Tarek Ali. Next-generation intrusion detection systems with llms: real-time anomaly detection, explainable ai, and adaptive data generation. Master's thesis, T. Ali, 2024.

[Chen *et al.*, 2016] Zhijiang Chen, Guobin Xu, Vivek Mahalingam, Linqiang Ge, James Nguyen, Wei Yu, and Chao Lu. A cloud computing based network monitoring and threat detection system for critical infrastructures. *Big Data Research*, 3:10–23, 2016. Special Issue on Big Data from Networking Perspective.

[Chen *et al.*, 2021] Zheyi Chen, Pu Tian, Weixian Liao, and Wei Yu. Zero knowledge clustering based adversarial mitigation in heterogeneous federated learning. *IEEE Transactions on Network Science and Engineering*, 8(2):1070–1083, 2021.

[Chisty *et al.*, 2022] Nur Mohammad Ali Chisty, Parikshith Reddy Baddam, and Ruhul Amin. Strategic approaches to safeguarding the digital future: insights into next-generation cybersecurity. *Engineering International*, 10(2):69–84, 2022.

[Cybersecurity C.I., 2018] Cybersecurity C.I. Framework for improving critical infrastructure cybersecurity. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf, 2018. Accessed: 2025-04-28.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[Dong *et al.*, 2021] Shi Dong, Ping Wang, and Khushnood Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021.

[Ferrag *et al.*, 2024] Mohamed Amine Ferrag, Mthandazo Ndhlovu, Norbert Tihanyi, Lucas C Cordeiro, Merouane Debbah, Thierry Lestable, and Narinderjit Singh Thandi. Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices. *IEEE Access*, 2024.

[G. Lira *et al.*, 2024] Oscar G. Lira, Alberto Marroquin, and Marco Antonio To. Harnessing the advanced capabilities of llm for adaptive intrusion detection systems. In *International Conference on Advanced Information Networking and Applications*, pages 453–464. Springer, 2024.

[Ge *et al.*, 2025] Linqiang Ge, Jingyi Zheng, and Wei Yu. Chapter 4 - machine learning in cyber-physical systems. In Wei Yu, editor, *Edge Intelligence in Cyber-Physical Systems*, Intelligent Data-Centric Systems, pages 71–99. Academic Press, 2025.

[Grumeza *et al.*, 2024] Theodor-Radu Grumeza, Thomas-Andrei Lazăr, and Alexandra-Emilia Fortiş. Social robots and edge computing: Integrating cloud robotics in social interaction. In *International Conference on Advanced Information Networking and Applications*, pages 55–64. Springer, 2024.

[Hatcher and Yu, 2018] William Grant Hatcher and Wei Yu. A survey of deep learning: Platforms, applications and emerging research trends. *IEEE Access*, 6:24411–24432, 2018.

[He *et al.*, 2023] Ke He, Dan Dongseong Kim, and Muhammad Rizwan Asghar. Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys Tutorials*, 25(1):538–566, 2023.

[Huang *et al.*, 2024] Chao Huang, Xubin Ren, Jiabin Tang, Dawei Yin, and Nitesh Chawla. Large language models for graphs: Progresses and directions. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1284–1287, 2024.

[Hussaini *et al.*, 2021] Adamu Hussaini, Bassam Zahran, and Aisha Ali-Gombe. Object allocation pattern as an indicator for maliciousness-an exploratory analysis. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pages 313–315, 2021.

[Kheddar *et al.*, 2024] Hamza Kheddar, Diana W. Dawoud, Ali Ismail Awad, Yassine Himeur, and Muhammad Khurram Khan. Reinforcement-learning-based intrusion detection in communication networks: A review. *IEEE Communications Surveys Tutorials*, pages 1–1, 2024.

[Kim *et al.*, 2022] Sangjun Kim, Kyung-Joon Park, and Chenyang Lu. A survey on network security for cyber–physical systems: From threats to resilient design. *IEEE Communications Surveys Tutorials*, 24(3):1534–1573, 2022.

[Li *et al.*, 2024] Fang Li, Hang Shen, Jieai Mai, Tianjing Wang, Yuanfei Dai, and Xiaodong Miao. Pre-trained language model-enhanced conditional generative adversarial networks for intrusion detection. *Peer-to-Peer Networking and Applications*, 17(1):227–245, 2024.

[Liang *et al.*, 2019] Fan Liang, William Grant Hatcher, Weixian Liao, Weichao Gao, and Wei Yu. Machine learning for security and the internet of things: The good, the bad, and the ugly. *IEEE Access*, 7:158126–158147, 2019.

[Liu *et al.*, 2018] Ming Liu, Zhi Xue, Xianghua Xu, Changmin Zhong, and Jinjun Chen. Host-based intrusion detection system with system calls: Review and future trends. *ACM computing surveys (CSUR)*, 51(5):1–36, 2018.

[Liu *et al.*, 2019] Xing Liu, Cheng Qian, William Grant Hatcher, Hansong Xu, Weixian Liao, and Wei Yu. Secure Internet of things (IoT)-based smart-world critical infrastructures: Survey, case study and research opportunities. *IEEE Access*, 7:79523–79544, 2019.

[Liu *et al.*, 2021] Xing Liu, Wei Yu, Fan Liang, David Griffith, and Nada Golmie. On deep reinforcement learning security for industrial internet of things. *Computer Communications*, 168:20–32, 2021.

[Mirkovic and Reiher, 2004] Jelena Mirkovic and Peter Reiher. A taxonomy of DDoS attack and DDoS defense mechanisms. *SIGCOMM Comput. Commun. Rev.*, 34(2):39–53, April 2004.

[Mishra and Mishra, 2024] Nilamadhab Mishra and Sarojananda Mishra. A review of machine learning-based intrusion detection system. *EAI Endorsed Transactions on Internet of Things*, 10, 2024.

[Mohammadi *et al.*, 2018] Mehdi Mohammadi, Ala Al-Fuqaha, Sameh Sorour, and Mohsen Guizani. Deep learning for iot big data and streaming analytics: A survey. *IEEE Communications Surveys Tutorials*, 20(4):2923–2960, 2018.

[Morgan, 2023] Steve Morgan. 2023 cybersecurity almanac: 100 facts, figures, predictions, and statistics. https://cybersecurityventures.com/cybersecurity-almanac-2023/, 2023. Accessed: April 28, 2025.

[Musa *et al.*, 2021] Usman Shuaibu Musa, Sudeshna Chakraborty, Muhammad M Abdullahi, and Tarun Maini. A review on intrusion detection system using machine learning techniques. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 541–549. IEEE, 2021.

[OpenAI, 2022] OpenAI. Introducing chatgpt, 2022. (Accessed 16 March 2025).

[Paya *et al.*, 2024] Antonio Paya, Sergio Arroni, Vicente García-Díaz, and Alberto Gómez. Apollon: a robust defense system against adversarial machine learning attacks in intrusion detection systems. *Computers & Security*, 136:103546, 2024.

[Qian *et al.*, 2025] Cheng Qian, Yifan Guo, Hengshuo Liang, Jianchao Song, and Wei Yu. Chapter 15 - secured edge intelligence in smart manufacturing cps. In Wei Yu, editor, *Edge Intelligence in Cyber-Physical Systems*, Intelligent Data-Centric Systems, pages 377–401. Academic Press, 2025.

[Rao *et al.*, 2023] Pyla Srinivasa Rao, Tiruveedula Gopi Krishna, and Venkata Sai Srinivasa Rao Muramalla. Next-gen cybersecurity for securing towards navigating the future guardians of the digital realm. *International Journal of Progressive Research in Engineering Management and Science (IJPREMS) Vol*, 3:178–190, 2023.

[Ridwan *et al.*, 2021] Mohammad Azmi Ridwan, Nurul Asyikin Mohamed Radzi, Fairuz Abdullah, and YE Jalil. Applications of machine learning in networking: a survey of current issues and future challenges. *IEEE access*, 9:52523–52556, 2021.

[Roshanaei *et al.*, 2024] Maryam Roshanaei, Mahir R Khan, and Natalie N Sylvester. Enhancing cybersecurity through ai and ml: Strategies, challenges, and future directions. *Journal of Information Security*, 15(3):320–339, 2024.

[Song *et al.*, 2025] Jianchao Song, Yifan Guo, Cheng Qian, and Wei Yu. Chapter 16 - secured edge intelligence in smart home cps. In Wei Yu, editor, *Edge Intelligence in Cyber-Physical Systems*, Intelligent Data-Centric Systems, pages 403–429. Academic Press, 2025.

[Tavallaee *et al.*, 2009] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani. A detailed analysis of the kdd cup 99 dataset. Online, 2009. Available at https://www.unb.ca/cic/datasets/nsl.html.

[TechTarget, 2023] TechTarget. Rapid reset ddos attacks exploiting http/2 vulnerability, 2023. Accessed: 2025-05-02.

[Thalpage and Nisansala, 2023] Nipuna Sankalpa Thalpage and Thebona Arachchige Dushyanthi Nisansala. Exploring the opportunities of applying digital twins for intrusion detection in industrial control systems of production and manufacturing–a systematic review. *Data Protection in a Post-Pandemic Society*, pages 113–143, 2023.

[The Hacker News, 2023] The Hacker News. Microsoft blames massive ddos attack for outage of azure, outlook, and onedrive services, 2023. Accessed: 2025-05-02.

[Tian *et al.*, 2025] Pu Tian, Weixian Liao, Cheng Qian, Yifan Guo, and Wei Yu. Chapter 12 - foundation of secured edge intelligence in cps. In Wei Yu, editor, *Edge Intelligence in Cyber-Physical Systems*, Intelligent Data-Centric Systems, pages 297–323. Academic Press, 2025.

[Wan *et al.*, 2024] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. Efficient large language models: A survey. *Transactions on Machine Learning Research*, 2024. Survey Certification.

[Yang *et al.*, 2015] Xinyu Yang, Jie Lin, Wei Yu, Paul-Marie Moulema, Xinwen Fu, and Wei Zhao. A novel en-route filtering scheme against false data injection attacks in cyber-physical networked systems. *IEEE Transactions on Computers*, 64(1):4–18, 2015.

[Yu *et al.*, 2007] Wei Yu, Xinwen Fu, Steve Graham, Dong Xuan, and Wei Zhao. Dsss-based flow marking technique for invisible traceback. In *2007 IEEE Symposium on Security and Privacy (SP '07)*, pages 18–32, 2007.

[Yu *et al.*, 2013] Wei Yu, Guobin Xu, Zhijiang Chen, and Paul Moulema. A cloud computing based architecture for cyber security situation awareness. In *2013 IEEE Conference on Communications and Network Security (CNS)*, pages 488–492, 2013.

[Yu *et al.*, 2015] Wei Yu, David Griffith, Linqiang Ge, Sulabh Bhattarai, and Nada Golmie. An integrated detection system against false data injection attacks in the smart grid. *Security and Communication Networks*, 8(2):91–109, 2015.

[Zhang *et al.*, 2024] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K. Gupta, and Jingbo Shang. Large language models for time series: A survey. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8335–8343. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Survey Track.