Data Labelling Using Generative AI

Enhancing Cybersecurity in IoT Networks

Cooper Weeks, Kobi Chambers, Campbell Reid, Zakaria Wardere,
Edas Ramanauskas and Darcy McIntosh

Work Integrated Learning - Single Project (3821ICT)

Griffith University, Gold Coast, QLD

25 Oct 2024

1. Abstract

This report explores the application of generative AI for data labeling in intrusion detection systems (IDS) within Internet of Things (IoT) environments. We fine-tuned several large language models (LLMs), including **LLaMA-3.1-8b**, **Phi-3-mini-2b**, and **Gemma-2-27b**, using **Quantized Low-Rank Adaptation (QLoRA)** techniques. These models were tested on both binary and multi-class classification tasks to label network traffic from IoT devices. Our experiments demonstrate that generative AI holds promise for automating data labeling in cybersecurity contexts, particularly in binary classification tasks, where we achieved up to **99.8% accuracy** using **Phi-3-mini-2b** from only 12,800 training samples. However, further tuning and extended training are needed to enhance model performance for more complex multi-class scenarios where we were only able to achieve an accuracy level of 95.6% using **Gemma-2-27b**. The results suggest that with additional resources, these models could become integral in real-world IDS deployments, improving the efficiency of identifying and mitigating cyber threats in IoT networks.

2. Introduction

The rapid growth of digital infrastructure and networked devices has made cybersecurity an increasingly critical area of research. Machine learning (ML) applications, particularly in IDS, have gained significant attention due to their potential in identifying and mitigating cyber threats. IDS aims to monitor and analyze network traffic for signs of malicious activities, such as denial of service (DoS), man-in-the-middle (MitM) attacks, and malware infections. Supervised learning methods, a cornerstone of many IDS implementations, rely heavily on well-labeled data to train accurate models that can differentiate between normal and attack traffic.

However, one of the primary challenges in developing effective IDS solutions is the need for large, high-quality labeled datasets. Traditional approaches to data labeling often require cybersecurity experts to manually label network traffic data, a process that is not only labor-intensive but also prone to inconsistencies, especially with large, complex datasets. Given the increasing volume of network traffic in modern environments, manual labeling quickly becomes impractical. This creates a bottleneck in training ML models that can adapt to the evolving threat landscape in real time.

To address this issue, the focus has shifted towards automating the data labeling process. Generative AI and ML techniques offer promising solutions by automating the labeling of network traffic, allowing for faster, more scalable IDS implementations. Datasets like the TON_IoT dataset, which includes extensive telemetry and network data collected from IoT environments, have been integral in advancing research in this area. The TON_IoT dataset includes various types of attacks, such DoS and ransomware, along with normal network traffic, making it a valuable resource for training IDS models.

However, manually labeling such diverse and large datasets remains a challenge, and there is growing interest in using generative AI to automate the process. In this context, automated data labeling methods are becoming critical to enable faster model development and deployment. Automating the labeling process can significantly reduce the time and resources required for preparing datasets like TON_IoT,

which contains both benign and malicious traffic. By reducing reliance on manual labeling, these methods help facilitate the development of robust IDS systems that can detect complex, evolving cyber threats more effectively.

For automated labeling to be practical in cybersecurity, models must achieve a high level of accuracy to minimize the risk of misclassification, which could have serious consequences in a security context. Based on recent studies [1], achieving close to 100% accuracy on labeled datasets is ideal but challenging. For instance, the Inception Time model has demonstrated 100% accuracy on the Win10–Network subset of the ToN-IoT dataset in multiclass scenarios. However, real-world applications require high performance across various datasets, and achieving a similar level of accuracy across different environments remains a goal.

When tested on other datasets, such as Edge-IIoT, the Inception Time model achieved a maximum accuracy of 94.94% across 15 classes, while the accuracy for the UNSW-NB15 dataset reached 98.4% for 10 classes. These results improved to 98.6% using a sliding window approach, but still indicate that models often fall short of the ideal 100% accuracy threshold. Therefore, a more realistic benchmark for model viability would be achieving at least 95% accuracy on unseen test data, aligning with the highest performance observed in real-world scenarios.

If these accuracy metrics, ideally between 95% and 100%, can be met, it could indicate that automated data labeling systems might become feasible in the future for cybersecurity tasks. Such systems could significantly reduce the time and resources required, while enhancing the effectiveness of IDS in detecting increasingly complex and evolving threats. This suggests the potential for continual refinement in model training and adaptation, especially as new datasets and attack types emerge.

3. Related Work

The growing connectivity of the IoT and Industrial IIoT has led to increased attention toward their cybersecurity, as these systems are vulnerable to various types of cyber-attacks. Traditional methods of detecting cyber threats have proven inadequate in the face of evolving attack vectors [6][9][11]. Consequently, ML, deep learning (DL), and natural language processing (NLP) models are being explored as promising solutions [1][2][7][8]. This section reviews related work on the application of these techniques for cybersecurity in IoT networks, focusing on the use of DL architectures, LLMs, and specific IoT/IIoT datasets.

3.1. DL for Cybersecurity in IoT

DL methods have gained significant traction in cybersecurity, offering superior performance over traditional ML methods. Various DL models have been employed for detecting cyber-attacks, especially in IoT environments. For instance, DenseNet and Inception Time networks have been extensively used for multiclass classification of IoT cyber-attacks. These models have demonstrated remarkable success in detecting intrusions across different IoT datasets, such as ToN-IoT, Edge-IIoT, and UNSW-NB15. Tareq et al. (2022) trained DenseNet and Inception Time models and achieved near-perfect accuracy, with the highest result being 100% accuracy using Inception Time on the Windows 10 version of the

ToN-IoT dataset [1]. The study highlights the advantage of using deep architectures in identifying complex attack patterns in large and heterogeneous IoT datasets.

Previous studies have also utilized other DL models such as CNNs, LSTMs, and hybrid approaches. For example, the Edge-IIoT dataset has been widely employed in experiments related to DL-based IDS. Researchers such as Ferrag et al. (2023) have demonstrated the use of a GAN-Transformer architecture for cyber-attack detection, though it did not outperform simpler architectures like DenseNet in some cases [10]. These findings emphasize the ongoing exploration of model complexity and dataset characteristics in achieving the best performance.

3.2. LLMs for Cyber Threat Detection

As the cybersecurity landscape grows more complex, researchers are turning to NLP models, particularly pre-trained LLMs like BERT, for identifying and mitigating cyber threats. LLMs have traditionally been used for text-based tasks, but recent advancements have shown their potential in non-textual domains such as network traffic analysis.

Ferrag et al. (2023) introduced SecurityBERT, a BERT-based model designed for IoT/IIoT cyber threat detection. Unlike conventional DL models that rely on raw network traffic, SecurityBERT employs a privacy-preserving fixed-length encoding (PPFLE) to transform network traffic data into a format interpretable by BERT [2]. This novel encoding technique enables SecurityBERT to achieve high classification accuracy (98.2%) on the Edge-IIoT dataset while remaining lightweight and efficient enough to run on resource-constrained IoT devices.

3.3. Efficient Fine-tuning Techniques for Large Models

One of the primary challenges in using large models in cybersecurity is the high computational cost associated with training and fine-tuning. To address this, recent works have explored Parameter Efficient Fine-Tuning (PEFT) techniques for making these models more efficient. Dettmers et al. (2023) proposed Quantized Low-Rank Adaption (QLoRA), an innovative method for finetuning large quantized LLMs using 4-bit precision without compromising accuracy [5]. By introducing novel techniques such as 4-bit Normal Float quantization and double quantization, QLoRA significantly reduces memory requirements, enabling the finetuning of massive models like a 65B parameter LLaMA model on a single 48GB GPU.

QLoRA has proven particularly useful in reducing the barriers to entry for working with large models in domains like cybersecurity, where resource-constrained devices and datasets with high dimensionality are common. The method shows that efficient fine-tuning can achieve state-of-the-art performance with significantly lower computational costs, making it a viable approach for large-scale cybersecurity applications.

3.4. IoT/IIoT Cybersecurity Datasets

The choice of dataset is crucial in evaluating the performance of ML and DL models in cybersecurity. Several public datasets have become benchmarks in IoT/IIoT research, such as the ToN-IoT, Edge-IIoT, and UNSW-NB15 datasets. These datasets contain diverse types of attacks, ranging from Distributed Denial of Service (DDoS) to malware and reconnaissance attacks, and have been extensively used for training ML and DL models.

The ToN-IoT dataset has been designed to simulate real-world IoT environments, including telemetry data from connected devices and system logs from both Windows and Linux operating systems. Edge-IIoT, a more recent dataset, was developed to address the increasing integration of IoT devices in industrial settings. It provides a broad range of attack scenarios specific to IIoT protocols, making it a valuable resource for developing robust and scalable IDS [6].

4. Approach

In this section, we outline the methodology used to train LLMs using the QLoRA framework. Our focus was on efficient fine-tuning of models such as LLaMA3.1, Gemma, and Phi3Mini, allowing us to handle resource-constrained environments while achieving high performance for cyber threat detection in IoT/IIoT systems.

4.1. QLoRA for Efficient Fine-Tuning of Large Models

QLoRA is a novel technique that enables the efficient fine-tuning of LLMs by reducing their memory footprint through 4-bit quantization. This approach allowed us to fine-tune massive models such as LLaMA3.1, Gemma, and Phi3Mini on a single GPU while preserving performance comparable to 16-bit fine-tuning.

The fine-tuning process using QLoRA consisted of the following steps:

4.1.1. Model Selection:

LLaMA3, Gemma, and Phi3Mini were selected due to their state-of-the-art performance on various NLP and domain-specific tasks. These models were pretrained on vast datasets and had demonstrated high generalization capabilities.

Using QLoRA, we were able to adapt these pretrained models to the specific task of cyber threat detection, focusing on IoT/IIoT datasets like ToN-IoT.

4.1.2. Data Formatting:

To effectively fine-tune the models for cyber threat detection, the data required proper formatting for use with the QLoRA framework. We implemented a chat-based prompt formatting approach to align with the pretraining structure of the LLMs. Using the Unsloth library's chat templates, we applied structured formatting to the input examples, ensuring that conversations between user and assistant

were appropriately tokenized and aligned with model expectations. The data formatting process involved defining the conversation template and mapping it into the appropriate structure for the model's tokenizer. This is an example of an input prompt

4.1.3. Quantization and Adapter Setup:

We loaded the quantized pretrained weights of each model using QLoRA's **4-bit format**. This data type is specifically optimized for normal distributions and provides a significant reduction in memory usage without sacrificing accuracy.

Additionally, **Low-Rank Adapters** were inserted into the transformer layers of the models. These adapters allowed us to perform task-specific learning by modifying only a small number of parameters, further improving the memory efficiency of the fine-tuning process.

To facilitate this process, we used **Hugging Face** to load Unsloth's pre-quantized models within **Google Colab notebooks**. This integration made it straightforward to implement the QLoRA techniques, enabling efficient training and fine-tuning of the models.

4.1.4. Fine-Tuning Process:

Each model (LLaMA3, Gemma, and Phi3Mini) was fine-tuned on the TON-IoT network dataset, which is a comprehensive dataset for network security comprising a variety of attack scenarios. The dataset is comprised of 10 classes (backdoor, ddos, dos, injection, mitm, normal, password, ransomware, scanning and xss), for binary classification all attack types were relabeled as 'attack' and normal traffic was kept 'normal'.

The fine-tuning process was conducted using 4-bit precision, enabling us to run the models on single GPUs without out-of-memory errors. This setup allowed for highly efficient training, reducing the need for high-cost computational infrastructure.

For each model, QLoRA's Double Quantization technique was applied to reduce the memory footprint even further by quantizing both the model weights and quantization constants.

4.1.5. Evaluation and Inference:

After fine-tuning, the models were evaluated using common classification metrics such as accuracy, F1-score, and precision-recall, specifically targeting the detection of multiple cyber-attack categories.

4.2. Integration with IoT/IIoT Datasets

The IoT/IIoT environments present a unique challenge due to the large volume of data and the diverse nature of cyber-attacks. We employed the TON-IoT dataset for training, which includes telemetry from various IoT devices and logs of cyber-attacks. By using QLoRA to fine-tune LLaMA3, Gemma, and Phi3Mini, we were able to:

- Train models that can efficiently detect cyber-attacks.

- Ensure that even large LLMs could be deployed in practical scenarios without needing extensive hardware resources.

4.3. Advantages of the QLoRA Approach

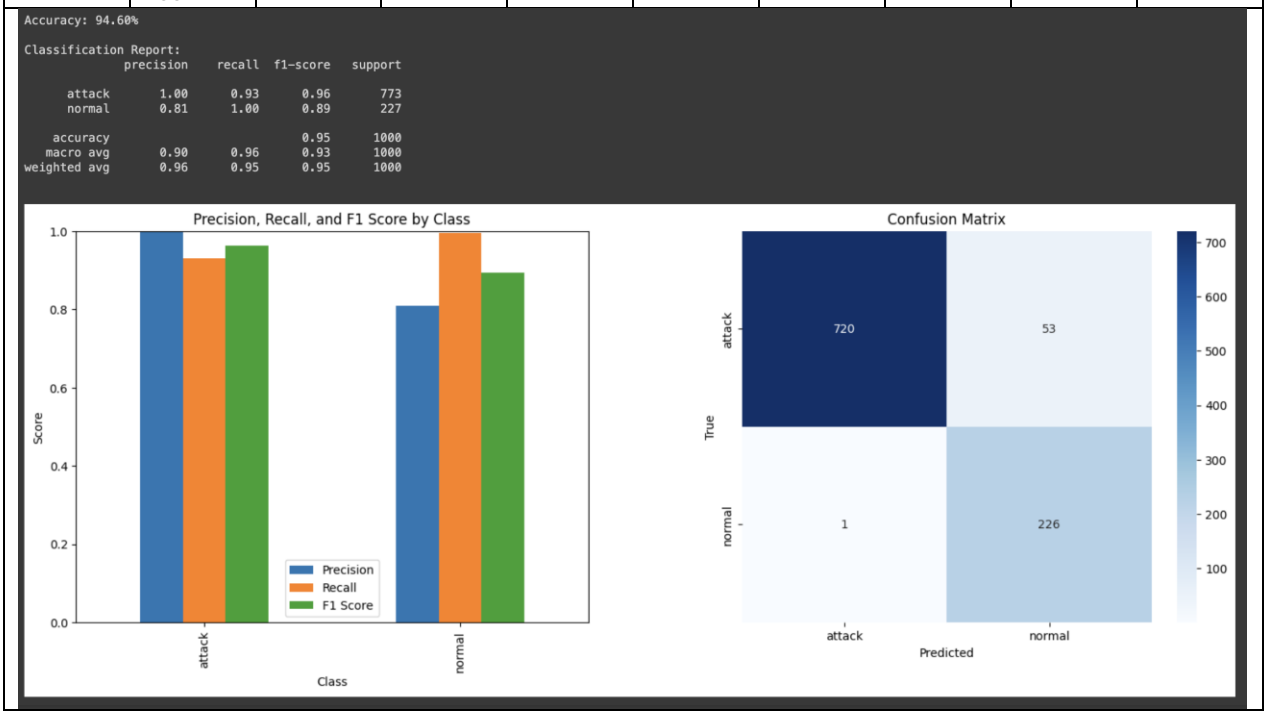
The use of QLoRA provided several key advantages in our approach:

- **Reduced Memory Usage:** By using 4-bit quantization and double quantization, we were able to significantly reduce the memory footprint of the models, allowing us to train on a single GPU.
- **Maintained Performance:** Despite the reduction in memory usage, QLoRA allowed us to fine-tune the models without any significant drop in performance, achieving near-state-of-the-art results for threat detection.
- **Scalability:** QLoRA's efficient fine-tuning made it feasible to apply large models in IoT/IIoT data labelling environments.

5. Results

Binary Classification

Model	Dataset	r	alpha	Batch Size	Learning Rate	Training Duration (Steps)	GPU	Accuracy (%)	Test Samples
Llama 3.1 8B	Network_Train_Test	128	8	128	2-e4	100	L4 – 22.5GB	94.6	1000



Phi 3.5 Mini 2.07B	Network_Train_Test	256	8	128	2-e4	100	T4 - 15GB	99.8	1000

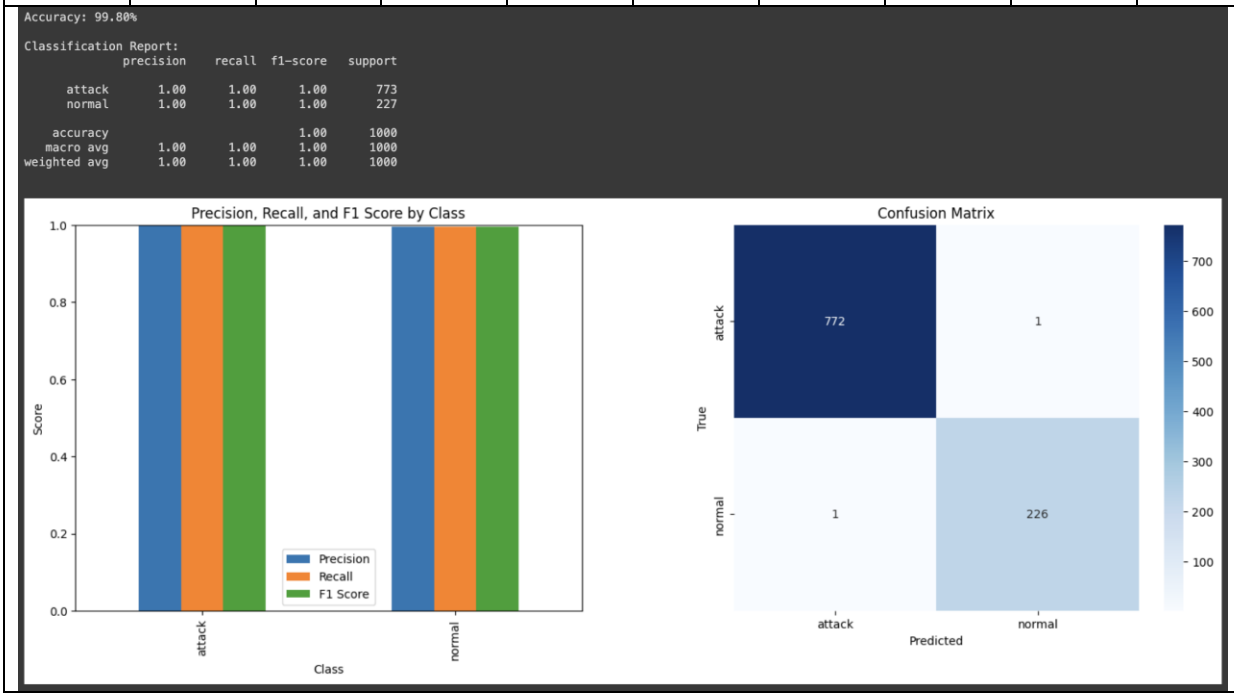


Table 1 Binary Classification Results.

Multi Classification

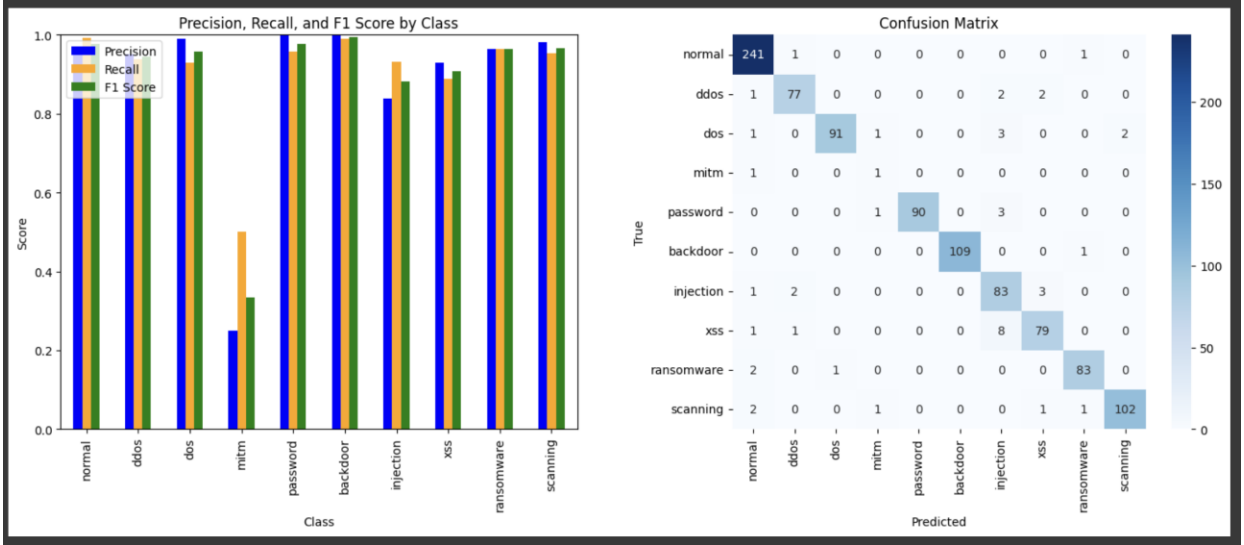
Model	Dataset	r	alpha	Batch Size	Learning Rate	Training Duration (Steps)	GPU	Accuracy%	Test Samples
Gemma 70B	Network_Train_Test	128	8	64	2-e4	100	A100 - 40GB	95.6	1000

```
Accuracy: 95.60%

Classification Report:
              precision    recall  f1-score   support

      normal       0.96      0.99      0.98       243
        ddos       0.95      0.94      0.94        82
         dos       0.99      0.93      0.96        98
        mitm       0.25      0.50      0.33         2
    password       1.00      0.96      0.98        94
    backdoor       1.00      0.99      1.00       110
   injection       0.84      0.93      0.88        89
         xss       0.93      0.89      0.91        89
  ransomware       0.97      0.97      0.97        86
    scanning       0.98      0.95      0.97       107

    accuracy                           0.96      1000
   macro avg       0.89      0.90      0.89      1000
weighted avg       0.96      0.96      0.96      1000
```



Phi 3 Mini 2.07B	Network_Train_Test	128	8	64	2-e4	100	T4-15GB	85.1	1000

```
Accuracy: 85.10%

Classification Report:
              precision    recall  f1-score   support

    scanning       1.00      1.00      1.00        94
  ransomware       0.00      0.00      0.00        95
         xss       0.42      0.95      0.58        76
   injection       0.90      0.67      0.76        90
      normal       0.60      0.50      0.55         6
    backdoor       1.00      1.00      1.00       232
    password       0.83      0.98      0.90        98
         dos       0.99      0.98      0.99       111
        mitm       0.95      0.94      0.95       101
        ddos       0.88      0.94      0.91        97

    accuracy                           0.85      1000
   macro avg       0.76      0.80      0.76      1000
weighted avg       0.81      0.85      0.82      1000
```

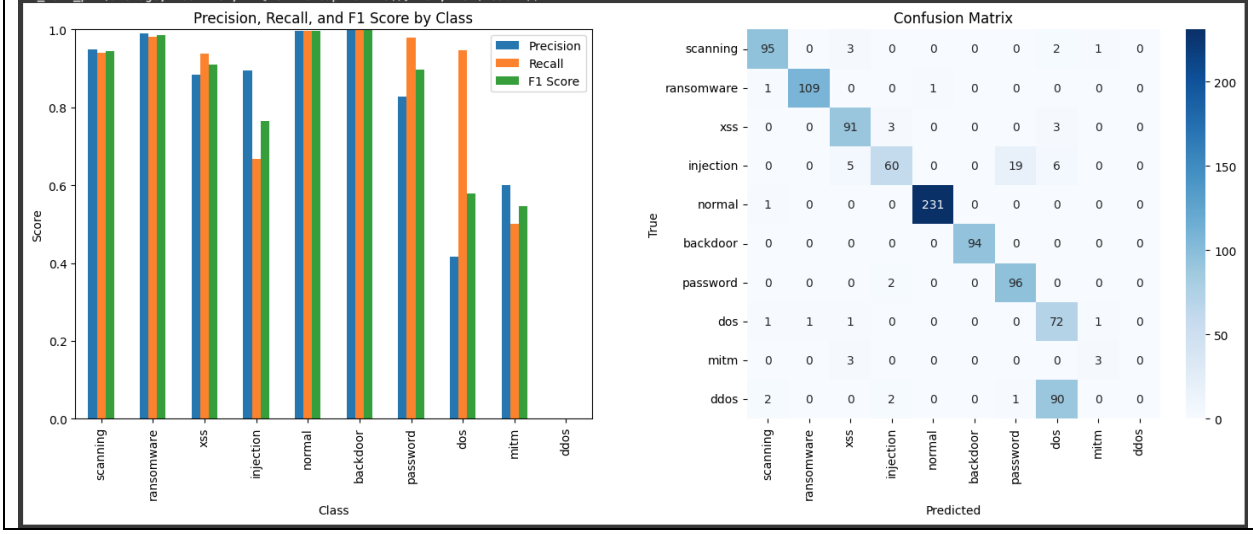


Table 2 Multi-Class Classification Results.

6. Discussion

Fine-tuning LLMs using LoRA offers a highly efficient approach to adapting pretrained models for specific tasks. Given that the task being fine-tuned is often a subset of the model's prior knowledge, models like LLaMA3, Gemma, and Phi3Mini can achieve exceptional performance with minimal fine-tuning. However, understanding the fine-tuning process and its limitations is key to optimizing performance.

6.1. Models Prior Learning

One of the primary reasons LoRA is efficient for fine-tuning is that any new task we fine-tune the model on typically falls within the domain of the model's existing knowledge. Pretrained LLMs, such as LLaMA3.1, Gemma, and Phi3Mini, are trained on extensive, diverse corpora, allowing them to acquire a vast amount of general knowledge. This broad pretraining enables the models to generalize well to new tasks with minimal fine-tuning [3][4], especially when these tasks are subsets or closely related to the concepts already encoded in the model.

For tasks such as classification, which require relatively simple decision-making processes, the model does not need to learn new representations from scratch. Instead, the model can rely on its existing representations and knowledge, requiring only minimal adjustments to the weights through fine-tuning. In our experiments, we found that the Gemma model performed exceptionally well on multi-class classification with as few as 100 training steps, achieving strong generalization to test data without the need for extensive fine-tuning.

Notably, Gemma, being the largest of the three models, achieved the best results from training. Interestingly, the performance ranking among the models did not scale linearly with parameter size. Phi-3, the smaller of the two binary models, outperformed LLaMA-3, defying the usual expectation that larger models perform better. This outcome likely reflects a more complex interplay of factors beyond just parameter count, although larger models generally possess greater capacity to learn and generalize from the training data.

6.2. Comparison with State-of-the-Art Performance

AI Type	Authors	Year	AI Model	Accuracy
Traditional ML	Ferrag et al.	2022	Decision Tree (DT)	67.11%
	Ferrag et al.	2022	Random Forest (RF)	80.83%
	Ferrag et al.	2022	Support Vector Machines (SVM)	77.61%
	Aouedi et al.	2023	DT + RF / FL	90.91%
	Zhang et al.	2023	K-Nearest Neighbor (KNN)	93.78%
	Ferrag et al.	2022	K-Nearest Neighbor (KNN)	79.18%
DL models	Friha et al.	2023	CNN / CL / No-DP	94.84%
	Friha et al.	2023	CNN / FL / No-DP	93.96%
	Aljuhani et al.	2023	CSAE + ABiLSTM	94.40%
	Friha et al.	2022	Recurrent Neural Network (RNN)	94%
	Ding et al.	2023	Long short-term memory (LSTM)	94.96%
	Ferrag et al.	2022	Deep Neural Network (DNN)	94.67%
	Friha et al.	2022	Deep Neural Network (DNN)	93%
	E. M.d. Elias et al.	2022	CNN-LSTM	97.14%
	Ferrag et al.	2023	Transformer model w/o Tokenization and Embedding	94.55%
Large language models	-	-	BERT without PPFLE	51.30%

	Ferrag et al.	2024	SecurityBERT	98.20%
This Work	Gemma Multi Classification	2024		95.6%
	Phi 3.5 mini, Binary Classification	2024		99.8%

Table 3 Comparison of Performance with State of The Art Models [2] pg.13

When comparing our results to state-of-the-art models like **SecurityBERT**, we must acknowledge the differences in data size, model architecture, and training durations. SecurityBERT is based on a fully fine-tuned **11,174,415-**parameter model that was trained on **1,765,482 samples** for **4 epochs**, achieving a multi-classification (15 classes) accuracy of **98.2%** on **441,371 test samples**. In contrast, the **Gemma model**, with **913,440,768 LoRA parameters**, was trained on **6,400 training samples** (10 classes), allowing us to achieve **95.6% accuracy** on **1000 test samples**. We observed that the accuracy varied by approximately 1-2% when compared to tests conducted with larger datasets of up to 10,000 samples. This suggests that while the model performs consistently, minor fluctuations in accuracy may occur as the number of test samples increases.

The significance of these results lies in several key aspects:

1. **Model Size and Complexity**: The **Gemma model** has a significantly larger number of parameters compared to SecurityBERT, which implies a greater capacity for learning complex patterns in the data. Despite being trained on a smaller dataset, Gemma's architecture allowed it to achieve competitive performance. This suggests that the model's design and training techniques, such as using **Low-Rank Adaptation (LoRA)**, can effectively leverage larger architectures even with limited training data.
2. **State Of the Art LLM's: Gemma** is a comparatively small model and ranks significantly below the top-performing models, falling short by nearly **20 points** on the quality index. This performance gap highlights the challenges faced by smaller models in competing with larger architectures that leverage extensive training data and vast parameter counts. Given these observations, we propose that proprietary LLMs are likely to generate much better performance in fine-tuning tasks. These proprietary models are typically developed with access to more extensive datasets and optimized training techniques, allowing them to achieve superior accuracy and generalization capabilities.
3. **Efficiency of Training**: Achieving **95.6% accuracy** with only **6,400 training samples** demonstrates the effectiveness of the training methodology employed. This is particularly noteworthy given that traditional approaches often require extensive datasets to reach high accuracy levels. It highlights the potential of fine-tuning

strategies like LoRA, which allow models to adapt quickly and effectively to new tasks with fewer examples.

4. **Comparative Performance**: While SecurityBERT achieved a higher accuracy, the proximity of the Gemma model's performance underscores its strength. The slight gap between the two models' accuracies suggests that with more extensive training data and additional epochs, Gemma could potentially surpass the performance of SecurityBERT.
5. **Scalability and Resource Utilization**: The results also reflect the scalability of using larger models with efficient fine-tuning methods. Given access to superior computational resources and longer training durations, we could further optimize the Gemma model, potentially achieving results comparable to or better than those of SecurityBERT.
6. **Implications for Future Research**: These findings open avenues for further exploration in the realm of cybersecurity and ML. They suggest that investing in more sophisticated model architectures combined with efficient training techniques like QLoRa can yield significant advancements in performance.

In summary, while our model currently achieves slightly lower accuracy than SecurityBERT, the implications of our findings point towards a promising direction for future developments in model training and architecture design. Given the right resources and methodologies, we are optimistic about the potential to enhance our model's performance.

6.3. Role of Dataset Diversity

When fine-tuning models on less diverse datasets, those with highly repetitive or similar examples, the model quickly learns the limited variability in the data. In contrast, more diverse datasets require a longer training period, as the model needs additional time to learn the varied patterns and relationships within the data. This dynamic can be defined by the **cosine similarity** between data points in the training dataset. If the dataset has high cosine similarity (i.e., the data points are very similar), the model can achieve high performance with fewer training steps.

This is why we can achieve impressive results with just **100 training steps** when working with binary classification tasks. In these cases, the limited variability allows the model to rapidly adjust its weights and generalize well to the test data. However, the same efficiency does not necessarily translate to multiclass classification tasks, where the increased complexity and diversity of the classes necessitate a longer training period. In such scenarios, the model struggles to learn the distinct features and relationships among the

various classes, resulting in a need for more extensive training to achieve optimal performance.

7. Key Takeaways

7.1. Generative AI for Data Labeling:

Generative AI shows significant potential for automating the data labeling process in IDS for IoT networks. Fine-tuning LLMs like LLaMA3, Gemma, and Phi3Mini with QLoRA techniques can efficiently adapt these models to specific tasks such as cyber threat detection.

7.2. QLoRA's Efficiency:

QLoRA enables the fine-tuning of large models using **4-bit quantization**, which dramatically reduces the memory footprint while maintaining strong performance. This method allows massive models to be trained on **resource-constrained devices**. By using quantization, QLoRA makes it feasible to leverage **large models** with greater understanding and capacity for complex tasks, such as language processing or cyber threat detection, even in environments with limited computational resources. This enables the use of advanced models that have been pretrained on vast datasets, ensuring they retain their comprehensive knowledge while being fine-tuned for specific tasks, such as labeling or classification, in a cost-efficient manner.

7.3. Performance in Binary vs. Multiclass Classification:

The models achieved 99.8% accuracy in binary classification tasks using the phi-3 mini model with just 12,800 training samples. However, performance in multiclass classification tasks was lower, with the best accuracy being 95.6%, highlighting the increased complexity and need for longer training in diverse datasets.

7.4. Role of Dataset Diversity:

Models trained on less diverse datasets, such as those used for binary classification, can achieve higher performance in the same number of steps due to the limited variability in the data. In contrast, more diverse datasets, like those used for multiclass classification, require longer training periods due to the higher complexity and diversity in the data.

7.5. Comparison with State-of-the-Art Models:

While the Gemma model performed well, it still lags behind models like SecurityBERT in terms of multi-classification accuracy. SecurityBERT achieved 98.2% accuracy with full parameter fine-tuning on a larger dataset.

7.6. Scalability and Resource Utilization:

The fine-tuning process with QLoRA proves to be highly scalable, allowing even large models like Gemma to be fine-tuned efficiently on a single GPU. This makes the approach practical for deployment in resource-constrained environments such as edge devices within IoT/IIoT networks.

7.7. Future Implications for Cybersecurity:

As the landscape of cyber threats continues to evolve, the results of this study suggest that future work should focus on utilizing PEFT methods, such as QLoRA, to enhance performance further. Achieving higher accuracy in multiclass scenarios is critical to improving real-world IDS effectiveness.

8. References

[1] Tareq, I., Elbagoury, B. M., El-Regaily, S., & El-Horbaty, E.-S. M. (2022). Analysis of ToN-IoT, UNW-NB15, and Edge-IIoT datasets using DL in cybersecurity for IoT. *Applied Sciences, 12*(19), 9572. https://doi.org/10.3390/app12199572

[2] Ferrag, M. A., Ndhlovu, M., Tihanyi, N., Cordeiro, L. C., Debbah, M., Lestable, T., & Thandi, N. S. (2023). *SecurityBERT: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Cyber Threat Detection*. https://arxiv.org/pdf/2306.14263

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language Models are Few-Shot Learners. *NeurIPS* arXiv:2005.14165

[4] Kaplan, J., McCandlish, S., Henighan, T., Brown, T., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv preprint* arXiv:2001.08361

[5] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. (2023). *QLORA: Efficient Finetuning of Quantized LLMs*. Preprint. https://arxiv.org/pdf/2305.14314

[6] Ferrag, M. A., et al. (2022). Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9751703

[7] Natasha Alkhatib, Maria Mushtaq, Hadi Ghauch, and Jean-Luc Danger. Can-bert do it? controller area network intrusion detection system based on bert language model. https://arxiv.org/pdf/2210.09439

[8] Othmane Friha, Mohamed Amine Ferrag, Mohamed Benbouzid, Tarek Berghout, Burak Kantarci, and Kim-Kwang Raymond Choo. 2dfids: Decentralized and differentially private federated learning-based intrusion detection system for industrial iot. Computers & Security, page 103097, 2023.

[9] Ons Aouedi and Kandaraj Piamrat. F-bids: Federated-blending based intrusion detection system. Pervasive and Mobile Computing, 89:101750, 2023

[10] Mohamed Amine Ferrag, Merouane Debbah, and Muna Al-Hawawreh. Generative ai for cyber threat-hunting in 6g-enabled iot networks. Pages 16–25. https://arxiv.org/pdf/2303.11751

[11] Xixi Zhang, Liang Hao, Guan Gui, Yu Wang, Bamidele Adebisi, and Hikmet Sari. An automatic and efficient malware traffic classification method for secure internet of things. IEEE Internet of Things Journal, 2023.