

# Integrated LLM-Based Intrusion Detection with Secure Slicing xApp for Securing O-RAN-Enabled Wireless Network Deployments

Joshua Moore, Aly Sabri Abdalla, Prabesh Khanal, and Vuk Marojevic  
Dept. of Electrical and Computer Engineering, Mississippi State University, USA  
Emails: {jjm702; asa298; pk571; vuk.marojevic}@msstate.edu

**Abstract**—The Open Radio Access Network (O-RAN) architecture is reshaping telecommunications by promoting openness, flexibility, and intelligent closed-loop optimization. By decoupling hardware and software and enabling multi-vendor deployments, O-RAN reduces costs, enhances performance, and allows rapid adaptation to new technologies. A key innovation is intelligent network slicing, which partitions networks into isolated slices tailored for specific use cases or quality of service requirements. The RAN Intelligent Controller further optimizes resource allocation, ensuring efficient utilization and improved service quality for user equipment (UEs). However, the modular and dynamic nature of O-RAN expands the threat surface, necessitating advanced security measures to maintain network integrity, confidentiality, and availability. Intrusion detection systems have become essential for identifying and mitigating attacks. This research explores using large language models (LLMs) to generate security recommendations based on the temporal traffic patterns of connected UEs. The paper introduces an LLM-driven intrusion detection framework and demonstrates its efficacy through experimental deployments, comparing non-fine-tuned and fine-tuned models for task-specific accuracy.

**Index Terms**—Intrusion detection, LLM, latency, Open Artificial Intelligence Cellular, O-RAN, security, slicing, xApp.

## I. INTRODUCTION

The Open Radio Access Network (O-RAN) is transforming the telecommunications landscape by promoting openness, flexibility, and intelligent closed-loop RAN optimization [1]. The main components of O-RAN include the Central Unit (CU), Distributed Unit (DU), and Radio Unit (RU), which together enable the disaggregation of traditional RAN functions for greater flexibility and multi-vendor interoperability. Additionally, the RAN Intelligent Controllers (RICs), the near-real time (RT) RIC and the non-RT RIC, enable artificial intelligence (AI)-driven optimization and management, enhancing network efficiency and adaptability. By decoupling hardware and software components and enabling multi-vendor deployments, O-RAN reduces capital expenditures while enhancing network performance as well as allowing operators to adapt rapidly to new technologies [2].

One such innovation is intelligent network slicing, a technique that allows operators to partition their networks into custom, isolated slices, each tailored to a specific application or quality of service (QoS) requirement. This capability is further enhanced by the ability of the RICs to optimize the resource allocation within network slices, enabling efficient

resource utilization, which improves service quality for connected user equipment (UEs).

However as illustrated in Fig. 1, while O-RAN establishes a more dynamic and modular architecture, it also increases the threat surface [3]. For instance, open interfaces in O-RAN enable seamless interoperability among equipment, but this openness also introduces significant security risks, as attackers may exploit these interfaces to inject malicious traffic or intercept sensitive communications. The reliance on over-the-air (OTA) transmissions between network elements exposes O-RAN to jamming, eavesdropping, and replay attacks, among others, which can compromise the integrity and availability of data. Furthermore, the intelligence of RICs presents additional vulnerabilities, as adversaries might manipulate AI models or introduce adversarial inputs, disrupting network optimization and decision-making processes. Advanced security measures must thus be considered to ensure the integrity, confidentiality, and availability of the network. Intrusion detection systems have become a critical component in identifying and mitigating malicious attacks to both traditional RAN and to future O-RAN deployments.

Recently, large language models (LLMs) have received significant attention from academia and industry for enabling intelligent decision-making across a wide array of fields, including education, finance, healthcare, and biology [4]. The capabilities of LLMs to process vast amounts of data and generate insights have positioned them as powerful assets for addressing complex challenges in these domains. In the field of wireless communications, LLMs have demonstrated remarkable potential in managing and optimizing network performance across diverse deployment scenarios [5]. Specifically, LLM-driven wireless network optimizations have been leveraged for edge intelligence [6], network intrusion detection [7], and reconfigurable intelligent surface deployments [8].

This research explores the unfolding possibilities of O-RAN in leveraging the capabilities of LLMs to generate security recommendations based on temporal traffic patterns of connected UEs. The goal is to evaluate the practical application of this emerging technology in enabling secure slicing, supporting heterogeneous traffic types on a shared O-RAN infrastructure. In this paper, we introduce a framework

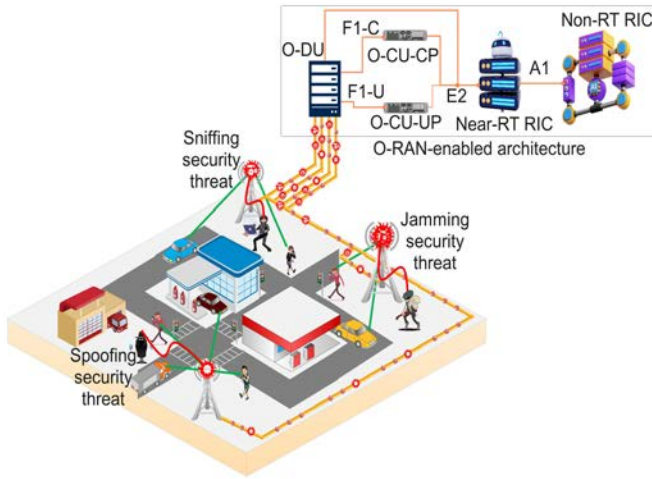


Fig. 1: The O-RAN architecture with wireless attacks.

for LLM-driven intrusion detection and, through experimental deployment, we show the efficacy of this approach for task-specific instructions comparing non-fine-tuned and fine-tuned models.

The main contributions of this paper are as follows:

- We propose a framework for LLM-based intrusion detection in the O-RAN context.
- We introduce an integrated framework composed of three xApps—KPIMON, LLM-based intrusion detection, and secure slicing—deployed in the near-RT RIC for collecting real-time data, detecting malicious activities, and isolating any detected intruder.
- Through an experimental deployment that leverages the Open Artificial Intelligence Cellular (OAIC) platform we demonstrate the effectiveness of this approach in safeguarding legitimate users in the presence of intruders.

The rest of this paper is organized as follows: Section II provides background related to intrusion detection in O-RAN as well as work leveraging LLMs for intrusion detection. In Section III, we introduce the experimental deployment and framework integration. Section IV details results obtained from testing the LLM as well as discusses the major findings and outcomes from testing various pre-trained models and one instruction tuned open source model. Section V presents the concluding remarks and future research directions.

## II. BACKGROUND AND PRIOR WORK

Intrusion detection systems in the O-RAN architecture are designed to process diverse data streams from multiple sources within the network, such as the RU, DU, and CU. These data streams provide a rich set of information, including network traffic patterns, signaling messages, and system logs, which are critical for comprehensive threat monitoring and analysis. Analysis engines employ a mix of techniques, including signature-based detection for known threats, anomaly-based models powered by machine learning to identify unusual behaviors, and heuristic or behavior-based methods to recognize deviations from expected patterns. The combination

of these engines ensures robust detection capabilities across a wide range of threat scenarios, from traditional attacks to sophisticated, zero-day exploits. Upon identifying potential intrusions, the intrusion detection systems can initiate various response mechanisms tailored to the severity and nature of the threat. Responses may include automated actions, such as isolating compromised components, triggering alarms for human intervention, or adapting network configurations to mitigate the impact [9].

Large language models offer potential to increase the capabilities of existing intrusion detection mechanisms within O-RAN. A proposed framework [10] uses a pre-trained LLM to automate intrusion detection in 5G networks by selecting relevant features, processing data, building prompts, and extracting decisions. In-context learning, a method where a pre-trained LLM is guided to improve performance on specific tasks by integrating examples directly into the prompts without altering the model’s parameters, is employed to improve detection accuracy using labeled examples and task-specific guidance. The work shows the capability of LLMs for feature extraction with a collected dataset but does not actualize the integration into a 5G network and uses pre-trained closed-source models for in-context learning results. Similarly, [11] shows the potential of LLMs in detecting anomalies in 5G network traffic. The work presents a centralized approach as well as a federated learning approach due to the edge device constraints. The results confirm that, while federated learning may have slightly lower accuracy than centralized approaches, it offers enhanced data privacy and scalability.

Llama 2 is used for detecting distributed denial of service (DDoS) attacks and improved using human feedback and compared against several established methods using the CIC-IDS2017 dataset [12]. The results demonstrate that Llama 2 achieves high accuracy and efficiency in detecting DDoS attacks with a performance comparable to long short-term memory, convolutional neural network, and deep neural network models, while also being suitable for real-time network traffic monitoring.

Unlike earlier studies that primarily explore anomaly detection or feature extraction from from previously collected network traffic datasets, this research proposes a comprehensive framework for LLM-driven intrusion detection specifically within the O-RAN context.

## III. O-RAN INTRUSION DETECTION FRAMEWORK

### A. LLM-based Intrusion Detection and Mitigation

To enhance security in O-RAN deployments, we propose a Large Language Model-based Intrusion Detection and Mitigation framework (LLM-ID), which leverages the reasoning capabilities of LLMs for real-time anomaly detection and response. This research introduces a comprehensive framework tailored specifically to the O-RAN context. By deploying this framework on the OAIC platform, we demonstrate its real-world applicability, going beyond theoretical models. Additionally, the research highlights how fine-tuning LLMs



We deploy a wireless communication system composed of a single base station with 3 UEs implemented over B210 Universal Software Radio Peripherals (USRP). These UEs are configured within a single slice as Enhanced Mobile Broadband (eMBB) users, which require continuous high-throughput requirements to support video streaming, immersive applications, and other data-intensive services. The bandwidth of the base station is set to 20 MHz with 100 RBs divided equally between the three users of the eMBB slice, where the eMBB user demands a 10 Mbps bit rate. The near-RT RIC is a modified version of the "E" release from the Open Source Community (OSC), enhanced with OSC's Key Performance Indicator Monitoring (KPIMON) xApp for performance monitoring and a custom secure slicing xApp (SSxApp) to enforce security measures and handle resource isolation upon detection of intrusion.

1) *Data Collection*: Data collection involves configuring the near-RT RIC to gather key performance indicators (KPIs), including packet transmission rates, resource allocation metrics, and UE connection status, from the RAN. We implement scripts for automated data logging and used Python tools for further analysis and visualization. We position the attacker at a fixed distance to ensure consistent signal propagation and data variability, monitoring performance metrics across various distances. We move the legitimate UEs with respect to the experimental base station in the laboratory to produce data sets with varying modulation and coding schemes. Collected data is preprocessed with Python scripts to remove obvious errors, and extract key parameters for fine-tuning the LLM.

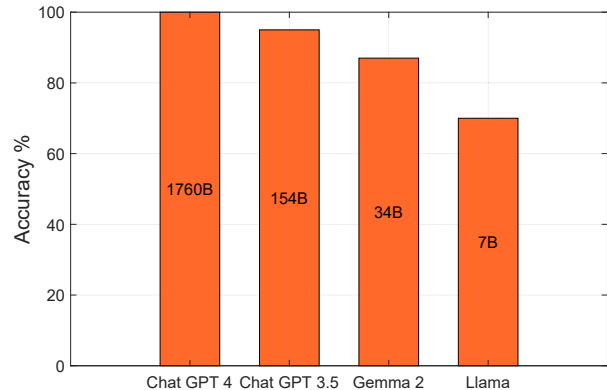
2) *Fine-Tuning the LLM*: Abiding by the O-RAN specifications [16], we fine-tune and deploy our LLM-based intrusion system offline on a GPU-accelerated server. The server is equipped with an Nvidia RTX 4090 GPU (24 GB VRAM), an AMD Ryzen 7 six-core CPU, and 128 GB of system RAM; it runs Ubuntu 22.04 LTS with a low-latency Linux kernel.

Gemma 2, an open-source LLM, is fine-tuned using the unsloth framework for a specialized instruction-based intrusion detection task in O-RAN. Fine-tuning Gemma 2 involves adapting the pre-trained model to recognize patterns indicative of security threats in the network traffic. This process trains the model on a data set containing labeled instances of both normal and malicious activities, with randomized variables such as the number of user devices and transmitted packets to expose the model to diverse traffic patterns, enhancing generalization and reducing overfitting. During this fine-tuning, model parameters are adjusted to optimize the ability to distinguish between legitimate and abnormal traffic [17].

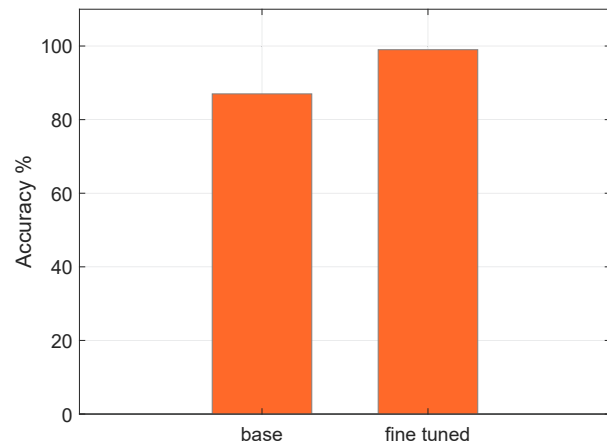
This fine-tuning leverages standardized O-RAN KPIs, including downlink and uplink bytes (DL BYTES, UL BYTES), physical resource blocks used (DL PRBS, UL PRBS), packet counts (TX PKTS, RX PKTS), transmission and reception errors (TX ERRORS, UL ERRORS), and active user counts (NUM UEs). These metrics enhance the model's understanding of network behavior. Aligning Gemma 2's learning objectives with KPIs, such as TX PKTS and NUM UEs, collected

from comprehensive OTA testing, optimizes the model's effectiveness in identifying intrusions and generating security recommendations. The use of standardized KPIs ensures that the fine-tuning aligns with the O-RAN philosophy and is relevant to real-world deployments.

#### IV. EXPERIMENTAL DEPLOYMENT AND RESULTS



(a) accuracy of base models with "few shot prompting"



(b) accuracy after fine tuning

Fig. 4: Comparison of model accuracy with few-shot prompting (a) and fine-tuning (b) for detecting anomalous user traffic flows.

We compare easily accessible closed and open source LLMs in two scenarios: (1) Non-fine-tuned models with few shot prompting techniques and (2) a fine-tuned open source model, Gemma 2, tuned for a task-specific prompt. We assess this fine-tuned model on the OAIC platform, where we deploy the LLM-ID and SSxApp and contrast it with a static threshold for intrusion detection and secure slicing.

The experimental setup involves a scenario where a malicious user device (UE) generates a flood of network requests to simulate a Denial of Service (DoS) attack. In this setup, the malicious UE overwhelms the resources allocated to specific network slices by sending a high volume of requests. This results in degraded service quality for legitimate users, demonstrating an effective strategy for disrupting service as

shown in prior research [18]. The attack exploits the absence of robust authentication mechanisms within the resource allocation policies of the slicing xApp. This vulnerability allows the attacker to disrupt the fair distribution of resources that the xApp is designed to ensure for connected user devices, ultimately causing potential service disruptions for legitimate users.

As shown in Fig. 4a, LLMs can perform well when instructed with a prompt that includes a clear example, such as defining the boundary of legitimate UEs in a network slice. This provides the model with context and helps it understand the task-specific constraints, leading to more accurate and relevant outputs. Specifically, we give the model an example of a legitimate UE based on its TxPkts and one malicious so that the model has the context of the boundary for TxPkts which is a common KPI in O-RAN.

The given prompt is:

PLEASE ONLY OUTPUT IN A WORD with TX Pack limits of 312 for 1 UE and 624 for 2 UEs, check if the following {NumUE} and {TXPackets} meet these bounds. If within bounds output *Legitimate* (input  $\leq$  bounds) or *Malicious* (input  $\geq$  bounds if exceeded).

This type of prompting is called “few-shot prompting” or “demonstration-based prompting.”

Larger models, such as Chat GPT 4 (1760B parameters) achieve higher accuracy (100%) in intrusion detection, demonstrating the advantage of complex architectures in capturing intricate patterns (Fig. 4a). Chat GPT 3.5 (154B) follows with 95% accuracy, showing strong performance but lagging behind Chat GPT 4. Gemma 2 (34B) achieves 87%, indicating that smaller models can still perform well, though less effectively than the larger ones. Llama (7B) performs the weakest at 70% accuracy, reflecting the limitations of smaller models in such complex tasks. Open-source models Gemma 2 and Llama provide flexibility and accessibility, while closed-source models Chat GPT 4 and Chat GPT 3.5 offer higher performance but are less customizable and accessible for research and development.

As illustrated in Fig. 4b, instruction tuning an open-source model can significantly enhance task-specific accuracy. Using demonstration-based prompting, the Gemma 2 model initially achieves an accuracy of 87% without any prior training. However, after fine-tuning, the accuracy for the same task increases to 99%, showcasing the effectiveness of the fine-tuning process.

Fig. 5 shows three UEs operating within the same slice with a maximum downlink throughput of 10 Mbps. Around 170 seconds into the experiment, one of the UEs becomes greedy and starts to consume more resources than prescribed, stealing resources from the other UEs operating within the same slice. Based on the traffic metrics collected from this UE, the LLM is able to identify the abnormality in the Tx Pkts as well as the throughput and classify the UE as an intruder. This is then used for the secure slicing xApp to mitigate further damage from occurring within the slice due to its ability to move the malicious UE to a slice with no resources. After

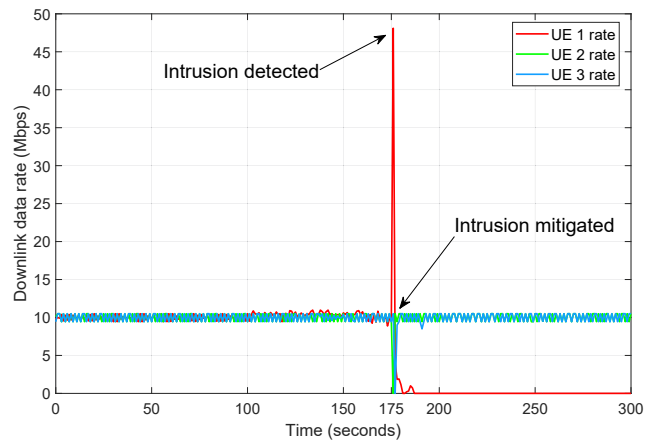


Fig. 5: Achieved data rate performance over time for legitimate UEs (UE2 and UE3) and malicious UE (UE1), illustrating the LLM-based ID and SSxApp mitigation events.

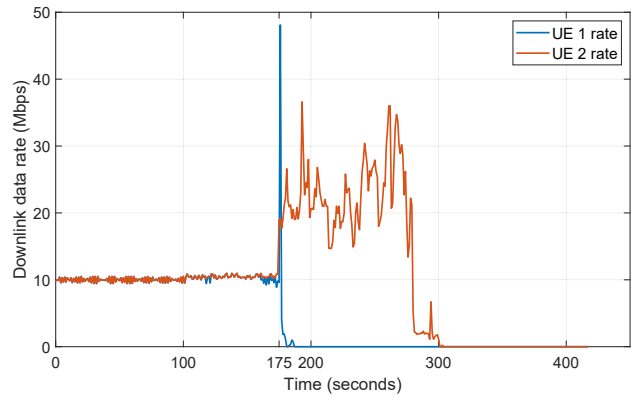


Fig. 6: Achieved datarate performance over experiment time for the LLM-based ID (UE 1) compared to a static ID method (UE 2).

that the malicious UE is eliminated, the legitimate UEs are able to quickly recover and access the previously captured resources by the malicious user. The average detection and response time is 239.66 ms, which reflects the time from the receipt of KPM reports to the mitigation of the intrusion. The KPImon xApp collects the KPM reports, which are generated at intervals between 1 ms and 1 second, and then the LLM-ID xApp analyzes these reports to detect potential intrusions. Upon detecting an anomaly, the SSxApp isolates the malicious UE by reallocating resources, effectively slicing it away from the network and minimizing disruption to legitimate users.

Fig. 6 compares the LLM-enabled intrusion detection with secure slicing (UE1) against intrusion detection based on a static threshold (UE2). This comparison of throughput over simulation time shows the difference in time to detect an intrusion and to move the UE to a slice with no resources. The LLM is the fine-tuned Gemma 2 model tuned for the task of identifying anomalous UE traffic flows. The 10 Mbps intended downlink throughput for the connected UE is initially

observed. Anomalous behavior starts around 175 s for each separate trial. UE1 is analyzed by the proposed LLM-ID system, which identifies it promptly as malicious to initiate the process or rebinding it to a zero-resource slice. UE2, which is processed by the intrusion detection based on the static threshold, is not marked malicious until the 280 s mark. This longer identification is due to false positives, which we have shown in our previous work [19] in which the static threshold was used. Multiple KPM reports can be collected to mitigate these false positives, leading to a slower detection. The LLM-enabled intrusion detection system, on the other hand, can accurately assess if an intrusion has occurred with only one KPM report. This capability is driven by the model's advanced contextual understanding, which enhances its ability to distinguish between normal network fluctuations and true security threats with greater precision, minimizing the reliance on multiple data points.

By leveraging the LLM's ability to analyze network traffic and identify malicious patterns in real-time, the system helps ensure compliance with service level agreements across network slices. The proactive security approach not only mitigates the risk of attacks but also optimizes detection time and network performance, ensuring integrity within network slices. This approach underscores the critical role of intelligent, adaptive network management for safeguarding O-RAN deployments.

## V. CONCLUSIONS

O-RAN represents a transformative advancement in telecommunications by promoting openness, flexibility, and intelligent resource management to support multi-vendor deployments, dynamic network slicing, and network intelligence. While these innovations significantly enhance network performance and adaptability, they also introduce new security challenges due to the system's modular and dynamic nature. This paper has presented a novel approach of using LLMs to strengthen O-RAN intrusion detection capabilities, effectively analyzing and generating security recommendations based on temporal traffic patterns of UEs. An intelligent framework composed of the proposed LLM-based intrusion detection xApp, the KPM report collection xApp, and the secure slicing xApp has been presented to effectively detect anomalies and malicious behavior and then isolate intruders to secure the network resources. Experimental results over the OAIC testbed have demonstrated the improved accuracy of an open-source model that is fine-tuned for the specific task of identifying anomalous traffic patterns, underscoring the potential of LLM-driven frameworks in enhancing the security and resilience of future O-RAN deployments.

## ACKNOWLEDGEMENT

This work was supported in part by NSF award 2120442 as well as NSF and Office of the Under Secretary of De-

fense (OUSD) – Research and Engineering, under Grant ITE2326898, as part of the NSF Convergence Accelerator Track G: Securely Operating Through 5G Infrastructure Program.

## REFERENCES

- [1] A. S. Abdalla, P. S. Upadhyaya, V. K. Shah, and V. Marojevic, "Toward next generation open radio access networks—what O-RAN can and cannot do!" *IEEE Network*, pp. 1–8, 2022.
- [2] M. Polese, L. Bonati, S. D'oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Communications Surveys & Tutorials*, 2023.
- [3] A. S. Abdalla and V. Marojevic, "End-to-end O-RAN security architecture, threat surface, coverage, and the case of the open fronthaul," *IEEE Communications Standards Magazine*, vol. 8, no. 1, pp. 36–43, 2024.
- [4] W. X. Zhao *et al.*, "A survey of large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2303.18223>
- [5] H. Zhou, C. Hu, D. Yuan, Y. Yuan, D. Wu, X. Liu, and C. Zhang, "Large language model (LLM)-enabled in-context learning for wireless network optimization: A case study of power control," 2024. [Online]. Available: <https://arxiv.org/abs/2408.00214>
- [6] Z. Lin, G. Qu, Q. Chen, X. Chen, Z. Chen, and K. Huang, "Pushing large language models to the 6G edge: Vision, challenges, and opportunities," 2024. [Online]. Available: <https://arxiv.org/abs/2309.16739>
- [7] M. Fu, P. Wang, M. Liu, Z. Zhang, and X. Zhou, "IoV-BERT-IDS: Hybrid network intrusion detection system in IoV using large language models," *IEEE Transactions on Vehicular Technology*, pp. 1–13, 2024.
- [8] Q. Liu, J. Mu, D. Chen, R. Zhang, Y. Liu, and T. Hong, "LLM enhanced reconfigurable intelligent surface for energy-efficient and reliable 6G IoV," *IEEE Transactions on Vehicular Technology*, pp. 1–9, 2024.
- [9] E. N. Amaghghi, M. Shojafar, C. H. Foh, and K. Moessner, "A survey for intrusion detection systems in open RAN," *IEEE Access*, vol. 12, pp. 88 146–88 173, 2024.
- [10] H. Zhang, A. B. Sediq, A. Afana, and M. Erol-Kantarci, "Large language models in wireless application design: In-context learning-enhanced automatic network intrusion detection," 2024. [Online]. Available: <https://arxiv.org/abs/2405.11002>
- [11] F. Adjewa, M. Esseghir, and L. Merghem-Bouhahia, "Efficient federated intrusion detection in 5G ecosystem using optimized BERT-based model," 2024. [Online]. Available: <https://arxiv.org/abs/2409.19390>
- [12] M. Mahmoodi and S. M. Jameii, "Utilizing large language models for ddos attack detection," in *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, 2024, pp. 1–6.
- [13] J. Moore, N. Adhikari, A. S. Abdalla, and V. Marojevic, "Toward secure and efficient O-RAN deployments: Secure slicing xApp use case," in *2023 IEEE Future Networks World Forum (FNWF)*, 2023, pp. 1–6.
- [14] OAIC, "Oaic," 2025, accessed: 2025-03-30. [Online]. Available: <https://www.openaicellular.org>
- [15] P. S. Upadhyaya, A. S. Abdalla, V. Marojevic, J. H. Reed, and V. K. Shah, "Prototyping next-generation O-RAN research testbeds with SDRs," *arXiv preprint arXiv:2205.13178*, 2022.
- [16] O-RAN Working Group 2, "O-RAN AI/ML workflow description and requirements," Open-RAN Alliance, Technical Specification, 10 2021, version 1.03.
- [17] G. Team *et al.*, "Gemma 2: Improving open language models at a practical size," 2024. [Online]. Available: <https://arxiv.org/abs/2408.00118>
- [18] Y. Shi and Y. E. Sagduyu, "Adversarial machine learning for flooding attacks on 5G radio access network slicing," 2021. [Online]. Available: <https://arxiv.org/abs/2101.08724>
- [19] A. S. Abdalla, J. Moore, N. Adhikari, and V. Marojevic, "ZTRAN: Prototyping zero trust security xApps for open radio access network deployments," *IEEE Wireless Communications*, vol. 31, no. 2, pp. 66–73, 2024.